Item analysis in DIGRAM 3.04

Part I: Guided tours

Svend Kreiner

Tine Nielsen

Table of contents

PART I: Guided tours

PART II: Detours and technicalities (tentative 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 Kreiner (2003).

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 a 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 loglinear modelling of ordinal categorical data.
5) Analysis of multidimensional Markov Chains.
6) Item analysis by graphical and loglinear Rasch models.

Item analysis in DIGRAM is described in two volumes of notes: Volume I containing guided tours and Volume II with detours and sundry technicalities.

1.2 Item analysis

The purpose of item analysis in DIGRAM is to

1) examine whether a summated index scale counting responses to a set of items provides a valid, objective and useful measure of a latent trait by analysis of the fit of item responses to a graphical Rasch 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. To obtain a copy of the program you have to send an email to Svend Kreiner (skm@biostat.ku.dk) , who will invite you to share a Dropbox folder with the program, user guides and data examples and where updates of the programs and the user guides will be available as they appear.

2) 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 loglinear Rasch model (GLLRM) where uniform DIF and uniform local dependence (LD) is permitted,

3) 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 during the item analysis, but these estimates are in most applications subordinate to the other purposes. Item parameter estimates are used during tests-of-fit of the model and during estimation of person parameters. They can, however, also be of interest in themselves in connection with special applications, for instance during development of computer adaptive tests or in studies of rater agreement where raters play the role of items.

1.3 Three examples

Three examples are used throughout these notes. 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 PF3 project

The second project originated in a Danish Health survey. We will here be concerned with the validity of the SF36 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 in the following way:

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

so that a low score indicates physical impairment. Gender and Age are also included in this project.

1.3.3 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 a total of 16 items covering three different domains as 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 Graphical Rasch models and graphical loglinear Rasch models

Rasch models are IRT models where items are locally independent and without DIF relative to all covariates and where the total score over items is statistically sufficient for the person parameters in the conditional distribution of item responses given the latent variable.

The models for item analysis implemented in DIGRAM differ from conventional IRT and Rasch models by assuming that the IRT models are imbedded in larger multivariate structural frameworks defined by chain graph models (Lauritzen, 1996) where covariates are expected to be associated with the latent variable and where analysis of DIF is restricted to the set of covariates in the complete framework. The network of covariates may on one hand be regarded as the frame of reference of the measurement model or as the nomological network described by Cronbach & Meehl in their definition of construct validity. In such models, the Rasch models are measurement components playing the same role as factor analysis models play in structural equation models and the person *parameters* of the Rasch models are regarded as outcomes on latent variables so that the

Rasch models describe the *conditional* distribution of the item responses given the latent variable. We refer to Rasch models embedded in chain graph structures as graphical Rasch models (GRMs).

Loglinear Rasch models (Kelderman, 1984) are Rasch models in the sense that the total score is statistically sufficient so that inference on item parameters can be separated from inference on person parameters, but loglinear Rasch models permit uniform DIF and uniform local dependence (LD). In such models, loglinear interaction parameters that do not depend on the person parameter represent DIF and LD. Loglinear Rasch models embedded in chain graph structures are called graphical loglinear Rasch models (GLLRMs). GLLRMs therefore are extensions of GRMs that may be helpful when item analyses have disclosed DIF and local dependence among items.

The item analysis in DIGRAM provides facilities for analysis by both GRMs and GLLRMs. Some of these 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 loglinear 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 analysis of the DHP-DE items shows that items fit a GLLRM with local dependence and DIF. We use this result to show what a GLLRM looks like and to explain why we think that GLLRMs provide measurements that are almost up to the standard of measurement from Rasch models.

Figure 1.1 shows what is known as the IRT graph in the GLLRM theory. The graph is the same type of Markov or independence graph that defines chain graph models where missing edges indicate that variables are conditionally independent. In this way, the IRT graph summarizes the results of the analysis of the DHP items by graphical Rasch and graphical loglinear Rasch models

where evidence of local dependence was found for two pairs of items (DHP34 & DHP38 and DHP38 & DHP39) and wher DIF relative to sex was found for one item (DHP36).

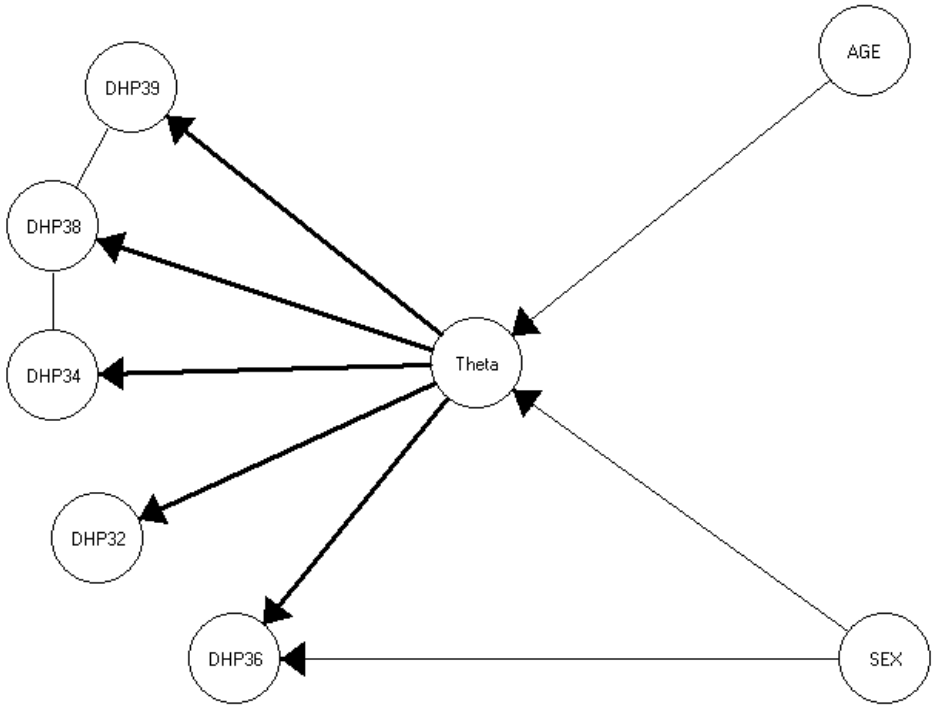


Figure 1.1. IRT graph of a GLLRM with two pairs of locally dependent items and one item with DIF relative to sex

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

$$\begin{aligned} &P(DHP_{32}=a, DHP_{34}=b, DHP_{36}=c, DHP_{38}=d, DHP_{39}=e \mid \theta, SEX=z, Age) \\ &= \\ &\frac{\exp\left(r\theta+\psi_{1a}+\psi_{2b}+\psi_{3c}+\psi_{4d}+\psi_{5e}+\lambda_{bd}^{(34,38)}+\lambda_{de}^{(38,39)}+\delta_{cz}^{(36,SEX)}\right)}{K} \end{aligned} \quad (1)$$

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

Models like (1) have many interesting features. The first is that the composite sum of items that are connected in Figure 1 behaves exactly like a partial credit item. If one instead of the separate item scores used the composite item scores, DHP32, DHP36 and DHP34+DHP38+DHP39 then one

would (except for the DIF of DHP36) have a perfect set of Rasch items. The second is that analyses of data for men and women separately would find perfect Rasch models in both groups. Except for the problem of comparing scores for men and women, we have no problems with the DHP-DE scores, and this problem can easily be overcome by adjusting scores form one group to be comparable with scores from the other group.

GLLRMs in this way may provide one solution to the 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 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 rest of this section illustrates that this is so.

The estimates of the item thresholds are as follows. Note that DHP has different thresholds for men and women because of the DIF and that the composite item counting scores on DHP34, DHP38 and DHP39 has nine thresholds because the range of this "item" consists of all integers from 0 to 9.

```
DHP32               -0.46   0.96  1.16
DHP36
   Sex =     Male   -0.52  -0.77  0.85
   Sex =   Female    0.60  -0.75  1.09

DHP34 + DHP38 + DHP39:
   -1.43  -0.42  -0.21  -0.52  -0.26   1.10  -0.12   1.27   0.86
```

The conditional likelihood ratio tests comparing item parameters in low and high score groups and item parameters in groups defined by sex and age comfortably accepts the GLLRM.

	CLR	df	p
Score groups	19.2	30	0.936
SEX	25.2	24	0.394
AGE	99.9	90	0.223

as does all the item fit statistics,

1.5 Publications on graphical (loglinear) Rasch models

The primary purpose of the guided tours is to show how to analyze item response data in DIGRAM using graphical Rasch models and graphical loglinear Rasch models. We have added a few appendices with technical details that are not covered elsewhere in Volume II, 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 appearing at the end of the following list of publications

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

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

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 loglinear 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 Loglinear 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.

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

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 Loglinear 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-restscore association. *Applied Psychological Measurement*, 35, 557-561

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

1.5 Item analysis commands

The commands for item analysis in DIGRAM are shown in Table 1.2 and illustrated in the guided tours in Chapter 2 and the detours in Chapter 3. 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.

Graphical Rasch models and graphical loglinear Rasch models are defined by so-called Markov graphs encapsulating the requirements of conditional independence made by the models. You can use DIGRAM's graph module to display the graphs and to redefine the models by adding or deleting edges and arrows between items and exogenous variables. These facilities are illustrated in Section 2.5 of the guided tour.

Table 1.2 Item analysis commands

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
CUT	Range and cutpoints	Redefines score groups	2.2.2
EXO	variables	Selects exogenous variables	2.1.2
Information on variables			
SHOW	**I**	Provides information on items	3.1
SHOW	**S**	Provides information on scores	3.2
Analysis of global Markov properties of Rasch models			
DIF	variables	Performs analyses of DIF	2.2.3
SCREEN	**I**	Performs item screening	2.4.1
SCREEN	**E**	Screening of the effect of exogenous variables on the score	2.4.1
CHECK	**I**	Check global Markov properties of the model	3.3
CHECK	**D**	Check the global Markov properties relating to DIF	3.3
MDIF	items	Tests the global Markov properties of multidimensional Rasch models	3.5
Parametric analyses			
GRM	Model generators	Analysis by graphical loglinear Rasch models	2.1.3
RASCH		Analysis by the Rasch model for dichotomous items	3.8
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	2.4.4
PROFILE	Item subsets	Analysis of profiles in different subpopulations	3.6
PRU	Item subsets	Assessment of the degree of practically unidimensionality	3.7
Score tables			
STABULATE	variables	Creates a contingency table containing the score groups and other variables.	3.4
Exploratory analyses			
DETECT	Number of dimensions	Identifies the optimal partitioning of items for a given number of dimensions	3.7.1
PURIFY		Provides help to select a pure subset of Rasch items	3.7.2

DIGRAM is constantly changed and (hopefully) improved. To keep track of what has happened, you can use the commands shown in Table 1.3.

Table 1.3 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**	Lists the nown (unsolved) problems
Information on commands		
HELP		Lists all available commands
command	**?**	Provides information on a specific command

2 Guided tours

This chapter describes four tours through DIGRAM where you will get a chance to take a look at 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 for unidimensionality and how to assess how well the items actually target the study population. During this tour we will also take a 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 loglinear Rasch models with uniform DIF and/or uniform local dependence and you will see that, apart from how to set up such models, there is no difference between inference in Rasch models and inference in graphical loglinear Rasch models because 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 loglinear Rasch modeling. It is during this tour that the differences between Rasch analysis and inference in graphical loglinear Rasch models become apparent.

Where to find it

Topic	***Sub topic***	***Sections***
Items	Selecting	2.1.1 2.2.1
	Flipping	2.2.1
	Info on	2.1.1 3.1
	Item parameter estimates	2.1.3.1 2.2.5.1 2.3.2
	Item difficulty, location and target	2.1.3.1
	Item characteristic curves	2.2.5.3 2.3.2
	Item information	
Exogenous variables	Selecting	2.1.2
	Disposing	2.1.2
	Info on	2.1.2
DIGRAM's GRM dialog		2.1.3
Scores	Info on	2.1.1 3.2
	Score tables	3.4
Score groups	Definition	2.1.1 2.2.2
	Info on	2.1.1 3.2
Rasch models		2.1 2.2
Graphical loglinear Rasch models		2.3 2.4
Global Markov properties		2.4.1 2.4.3 3.3
IRT and Rasch graphs		2.2.4 2.3.1 2.4.1
Overall tests of fit	Homogeneity	2.1.3.2
	No DIF	2.1.3.2
	Item correlation	2.1.3.4
Item fit statistics		2.1.3.3 2.3.2
Tests of no DIF		2.1.3.5 2.2.3 2.3.3 2.4.2
Tests of local independence		2.1.3.6 2.3.3 2.4.2
Item screening		2.4.1
Uni- and multidimensionality	Tests of unidimensionality	2.2.5.2
	DIF relative to other latent variables	3.5.1
	Practically unidimensional	3.5.2

Analysis of person fit		2.4.4
Person parameter estimates		2.1.3.7 2.3.4
Test difficulty, location and target		2.1.3.7 2.2.5.4
Reliability		2.1.3.7
Targeting		2.1.3.7 2.2.5.4 2.3.4
Item maps		2.2.5.4
Local homogeneity & DIF		2.4.5.1
Exploratory analyses	Multiple dimensions	3.6.1
	Item purification	3.6.2
Export of items to other programs	RUMM	3.8.1
		3.8.2

2.1 Rasch models. The very short tour

During the first tour where we use the Diabetes Health profile (DHP) project we will cover the basics of item analysis by Rasch models. 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.

During this tour you only need three DIGRAM commands: ITEMS, EXOGENOUS, and GRM. The GRM command invokes a graphical dialog where you have to select among a number of options ad where output will be displayed. The GRM 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 five items measuring disinhibited eating, 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. Note that two buttons ("IRT graph" and "Graphical Rasch models") have been enabled and that a list of items is shown in the panel below the main form.

Limitations:

The current version of DIGRAM requires that all items have the same number of response categories despite the fact that Rasch models do not share this limitation. This will be remedied in the future, but until that happens you have to cheat the program by insisting that all items have the same number of categories. Some of these will be dummy categories that never are used, but DIGRAM is able to handle such categories during both analysis of contingency tables and item analysis.

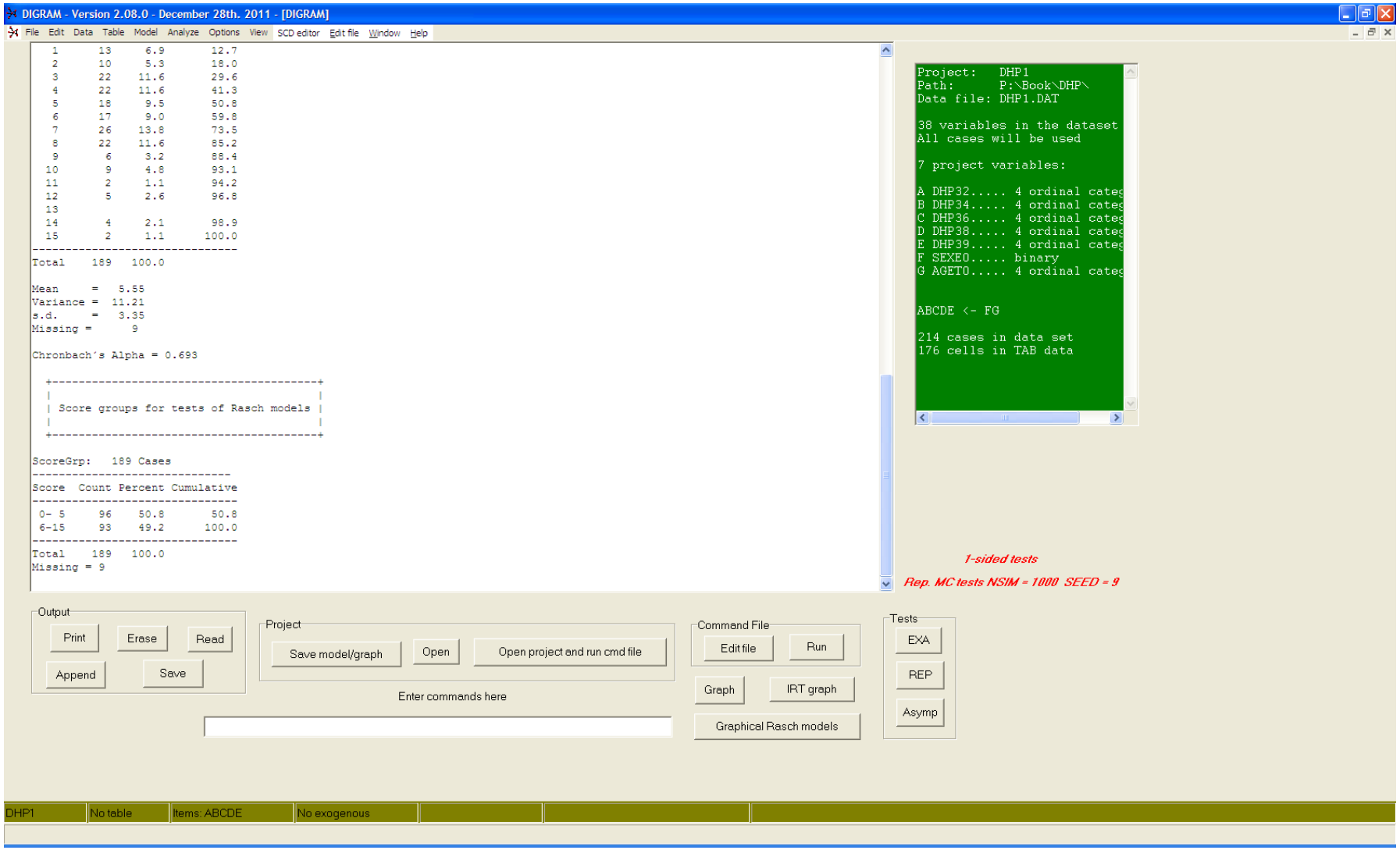


Figure 2.1.1 DIGRAM's main form after selection of items

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

Notes: Information on the average item scores is provided for all persons responding to an item and for the persons with complete responses on all items. Use this information to check that item responses seem to be missing at random. If item responses are lower for persons with complete responses it could be that missing responses should be coded as “errors” with item score equal to zero.

Much more information on items is available if you use the “SHOW I” command. This is described during the first detour in Chapter 3. We suggest that you wait with this detour until later.

```
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
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
Missing =  9
```

Figure 2.1.3 Information on the score and score groups

Notes: 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, of course, be of interest in themselves, but they do not provide information that can be used during the item analysis. The data used during the item analysis therefore consist of 176 persons.

A cut point equal to 5 defines two score groups so that 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 analyse 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 describes how you can use the “SHOW S” command to obtain 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 association with the latent variable. The DHP project has two exogenous variables, F = Sex and G = Age defined by four ordinal categories, 18-49, 50-59, 60-69, and 70-100. To select these two variables, we invoke the “**EXOGENOUS FG**” command. The result can be seen in Figure 2.1.4.

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

   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:    SEX       1       14.0        5.5     35.40 0.000
G:    AGE       1       14.0        5.5     35.40 0.000

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

              ABCDE# <- ¤ <- FG
```

Figure 2.1.4 Exogenous variables

Notes : Missing outcomes on exogenous variables reduce the number of cases that are covered by the graphical Rasch model. 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.

Finally, 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 have to be 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 pulled back to the same level as the items.

If you want to select other exogenous variables you just have to use the **EXOGENOUS** command again. When you do this, the old 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, the graphical Rasch model will be reinitialized and results obtained during the item analysis with previous set of items 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 form shown in Figure 2.1.5 turns up.

The GRM dialog box provides information on items and exogenous variables and offers a wide selection of options to choose from during the item analysis. On this tour we will only consider a few of those, but we will return for looks at the other options during the next tours.

Before we proceed you should familiarize yourself with the GRM dialog form.

Use the Start button when you have decided on the available options. 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 large field at the right side of the form. Below this field are a number of buttons that you can use to erase output and/or save the output on the output field of DIGRAM's main form. And, of course, an exit button that you have to use if you want to return to DIGRAM's main form.

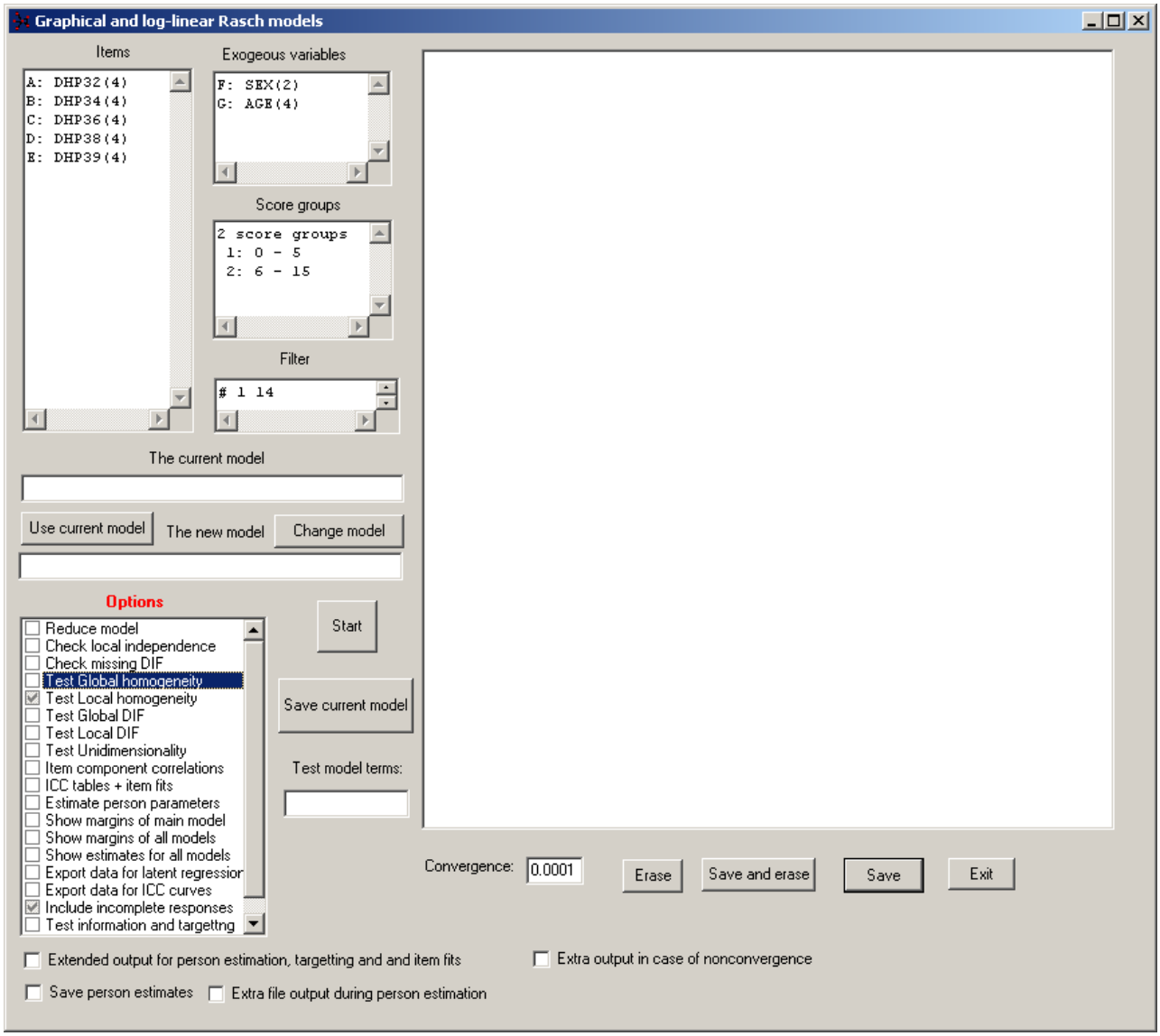


Figure 2.1.5 The GRM dialog form

The rest of the buttons and the fields with the current and new models are only of interest if you are working with graphical loglinear Rasch models so we return to these buttons during the tours through these models.

The information on items and exogenous variables cannot be edited 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[2] if you want to restrict the analysis to a specific subset of persons.

2.1.3.1 Estimating item parameters

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 parameter. To estimate the item parameters you just 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. The result is shown in Figures 2.1.6 and 2.1.7.

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


The model is estimated under the following assumptions:

   1 <= raw score <= 14

The assumed model:

All items are assumed to be locally independent
All items are assumed to be without DIF

176 valid cases included

Observed score range: 0 - 15

  +---------------------------+
  |                           |
  | Estimated item parameters |
  |                           |
  +---------------------------+

      item          0         1         2         3
-----------------------------------------------------
A:    DHP32       1.000     2.053     0.885     0.254
B:    DHP34       1.000     1.077     2.113     4.216
C:    DHP36       1.000     1.370     3.387     1.337
D:    DHP38       1.000     1.910     0.284     0.235
E:    DHP39       1.000     9.316     2.484     2.972
```

Figure 2.1.6 Multiplicative items parameters

Figure 2.1.6 shows the item parameters according to the formalization of the Rasch model as a multiplicative power series model[3]. These parameters are included for completeness because all computations during the item analysis are done on this version of the model. Most users of Rasch

[2] 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.

[3] The power series formalization is discussed in Volume II.

models are, however, more familiar with the partial credit model (PCM) formalization, and if you are one of them you will probably want to skip this part and instead look at the PCM thresholds that are shown in Figure 2.1.7 together with the "difficulty" of the items[4]. Disordered thresholds are a concern to many users of Rasch models. For this reason, a ">" is included between thresholds if thresholds are disordered. Four out of five DE items have disordered thresholds and DHP34 have completely disordered thresholds. DIGRAM offers no facilities for collapsing of categories during item analysis. If you want to analyse items with collapsed response categories you therefore have to define a new DIGRAM project with the collapsed versions of the items.

Finally, the list of items parameters include information on the conditional likelihood and on two information criteria (AIC and BIC).

```
  +-----------------------------+
  |                             |
  | Item main effect thresholds |
  |                             |
  +-----------------------------+

The item difficulty is equal to the person parameter where the expected
item score is equal to the highest obtained item score divided by 2.


A -    DHP32     -0.72      0.87      1.15   Difficulty =    0.53  (Exp. score = 1.5)

B -    DHP34     -0.07 >   -0.67 >   -0.68   Difficulty =   -0.49  (Exp. score = 1.5)

C -    DHP36     -0.31 >   -0.91      0.95   Difficulty =   -0.23  (Exp. score = 1.5)

D -    DHP38     -0.64      1.90 >    0.22   Difficulty =    0.70  (Exp. score = 1.5)

E -    DHP39     -2.23      1.32 >   -0.17   Difficulty =    0.04  (Exp. score = 1.5)


Nstep = 19   Delta =  0.0000045   14 estimated parameters


-ln(likelihood) =    577.748
            AIC =   1183.495
            BIC =   1227.882
              n =        176
```

Figure 2.1.7 Partial credit thresholds

2.1.3.2 Overall tests of fit

DIGRAM uses Andersen's (1973) conditional likelihood ratio (CLR) test for overall tests of homogeneity and no DIF. The test of homogeneity compares item parameter test in the two score

[4] The difficulty of an item is defined as the value of the person parameter with an expected score on the item equal to half the maximum score on the item. This is different from the "location" of the item defined by the average of the thresholds when the number of response categories is larger than three. More on this in Section 2.1.3.6 on item and test targeting.

groups defined during selection of items whereas the tests of no DIF compares item parameters in groups defined by the two exogenous variables. The results are collected in a small table presented at the end of the output produced during calculation of overall tests. This table is shown in Figure 2.1.8

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

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

Figure 2.1.8 Overall test of fit of the Rasch model

In most cases there will be no reason to look at anything but the final table containing the CLR tests. Situations may occur, however, where the rest of the output produced during the calculation of the tests might be of interest. The weak evidence against the hypothesis of homogeneity (p = .018) could motivate a closer look at what happened during the calculation of this test. This output is shown in Figure 2.1.9 comparing observed and expected average item scores together with standardized residuals in the two score groups. The results indicate 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.

Note: Taken by itself, the weak evidence against homogeneity and item DHP38 provided by one out of three overall tests is not enough to reject the Rasch model. It does, however, suggest that a more careful look at the kind of item fit statistics described in Section 2.1.3.3.

To us, the conditional likelihood ratio test is the fundamental overall fit statistic for the Rasch model. It address the same fit issue as other overall fit statistics like, for instance, the overall 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. Note, however, that Section 2.1.3.4 presents an overall fit statistic based on a completely different approach and with much more complicated asymptotics.

```
****  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 2.1.9 Analysis of homogeneity of item responses

2.1.3.3 Item fit statistics

Select item fit statistics to check whether responses for separate items appear to come from the Rasch model. DIGRAM calculates three item fit statistics. Outfits and Infits are well-known and much used item fit statistics going back to the early days of the theory of Rasch models. The Outfits and Infits calculated by DIGRAM compares observed item responses to the expected responses under the *conditional* distribution of responses given the total score to avoid bias and for realistic assessment of significance (see Kreiner & Christensen (2011b) for details). Both fit statistics have expected values equal to 1 under the Rasch model. Fit statistics above 1 indicate weaker item discrimination than expected under the Rasch model whereas fit statistics below 1 suggest that the item discrimination is too strong to be items from a Rasch model.

Outfits and Infits are sensitive to other types of departures from the Rasch model. For this reason, they are suspected to have less than optimal power against differential item discrimination. The item – rest score gamma coefficient is, on the other hand, targeted at the problem of item discrimination and therefore expected to have a little stronger power against such alternatives than the two other fit statistics. The item – restscore 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, this coefficient is also assessed under the conditional distribution of item responses given the total score on all items (Kreiner, 2011).

```
  +--------------------------------+
  |                                |
  | 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 restscore association  |
  |                             |
  +-----------------------------+

                  Item-restscore 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 2.1.10 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 % (***)

Comments on Figure 2.1.10: The item fits statistics only disclose misfit for one item (DHP38) that appears to have a much higher item discrimination than expected by the Rasch model. The Outfit and Infit statistics disagrees suggesting instead that the item discriminations of DHP32 and DHP34 are weaker than expected by the Rasch model, but the significance of these fit statistics were rejected after controlling for multiple testing.

Differential item discrimination indicate misfit to the Rasch model irrespective of whether item discrimination appears to be larger or smaller than expected by the Rasch model. The interpretation of the departures from the Rasch model is, however, somewhat different. Low discrimination is something that is to be expected for bad items that for some reason (e.g. bad item writing) do not relate exclusively to the latent variable for which reason such items are often eliminated. Evidence suggesting that the item discrimination is too strong does not support such interpretations. To some researchers, such evidence would suggest that the Rasch model is abandoned in favour of another type of IRT model. While this, at the end of the day, may be the only way to address the problem indicated by Outfits and/or Infits smaller than 1, we would 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 response dependency, multidimensionality and/or DIF because such problems require very different kinds of solutions.

2.1.3.4 Analysis of correlations among items

It follows from the properties of the Rasch models that items have to be positively correlated. Given the estimates of the item parameters and the distribution of the raw scores, DIGRAM can calculate the expected correlation and test whether the observed correlations among items are significantly different from the expected correlations.

To obtain these tests you should select “Item component correlations” from the list of options in Figure 2.1.5. The results are shown in Figure 2.1.11 where the observed and expected correlations are measured by Goodman and Kruskal’s γ. Test are provided both for the separate pairs of items and for the set of items as a whole.

```
  +-----------------------------------------------+
  |                                               |
  | Analysis of correlations among item components |
  |                                               |
  +-----------------------------------------------+

5 item components: 1: A 2: B 3: C 4: D 5: E

         A     B     C     D     E
 A obs       0.325 0.198 0.420 0.300
   exp       0.388 0.373 0.380 0.411
     p       0.441 0.036 0.670 0.227

 B obs 0.325       0.355 0.727 0.439
   exp 0.388       0.412 0.392 0.429
     p 0.441       0.441 0.000 0.901

 C obs 0.198 0.355       0.331 0.441
   exp 0.373 0.412       0.379 0.416
     p 0.036 0.441       0.579 0.773

 D obs 0.420 0.727 0.331       0.670
   exp 0.380 0.392 0.379       0.419
     p 0.670 0.000 0.579       0.009

 E obs 0.300 0.439 0.441 0.670
   exp 0.411 0.429 0.416 0.419
     p 0.227 0.901 0.773 0.009


Benjamini & Hochberg rejects at 0.01000 to control the FDR at 0.05
Benjamini & Hochberg rejects at 0.00100 to control the FDR at 0.01
Benjamini & Hochberg rejects at 0.00010 to control the FDR at 0.001

Overall chi square =    29.7  df = 10   p =  0.001

n = 176   RMSE =  0.130

Too strong association:

  B & D  0.727  0.392  0.00009
  D & E  0.670  0.419  0.00910

No evidence of associations that are weaker than expected
```

Figure 2.1.11 Comparison of observed and expected correlations among items. The evidence of differences is adjusted for multiple testing by the Benjamini-Hochberg procedure controlling the false discovery rate at 5 %.

The asymptotic p-values for the separate pairs of items in Figure 2.1.11 are sound, but the overall χ^2 defined as the sum of all these χ^2 statistics can not be assumed to have a χ^2 distribution with degrees of freedom equal to the number of pairs of items. There are two reasons for this: first, that the large

number of parameters that have been estimated has to be taken into account and second, that the separate χ^2 statistics cannot be assumed to be independent. For this reason DIGRAM suggests that you try to estimate the p-value by parametric bootstrapping, asking you (Figure 2.1.12) to tell us the bootstrap sample size, whether you want to sample in the marginal or the conditional inference frame and the maximum number of iterative steps you will let DIGRAM take during estimation of item parameters for each sample.

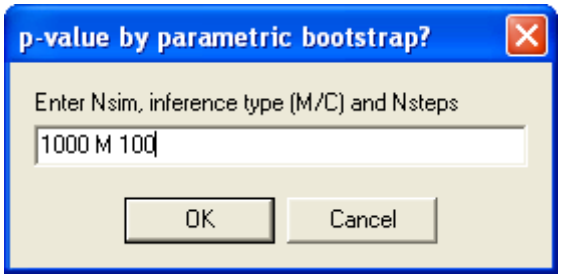


Figure 2.1.12 Information on parametric bootstrapping during analyses of marginal items correlations. (Nsim = sample size, Inference type = M (marginal) or C (conditional), Nsteps = maximum iterative steps permitted during item parameter estimation)

The result of the parametric bootstrap is shown in Figure 2.1.13 where the bootstrap estimate of the p-value agrees with the asymptotic p-value. In addition to what is shown in Figure 2.1.13, DIGRAM also reports on a number of technical issues that probably only are of interest to those that have developed the procedures so we do not show it and offer comments on these results for now.

A bit of advice concerning the use of parametric bootstrapping here is, perhaps, appropriate.

First, bootstrapping is time consuming. Since the asymptotic p-value is always larger than the bootstrap estimate of the true p-value, or has at least been larger in all the examples that we have examined. If the asymptotic p-value is equal to 0.000 there is no reason to waste time on bootstrapping.

Second, the reason why it is so time-consuming is that the item parameters are re-estimated for each bootstrap sample. To save time, you may set Nsteps = 0 in which case the expected correlation will be based on the estimates in your own data, that is equal to the expected values in Figure 2.1.11.

This will give you a too conservative assessment of the degree of misfit between observed and expected correlations, but again – if such a p-value is equal to 0.000 then the same has always also been true for the proper bootstrap estimate.

```
  +----------------------+
  |                      |
  | Parametric bootstrap |
  |                      |
  +----------------------+

Sample size    = 1000

Inference type = marginal

Maximum number of iterative steps during estimation = 100

Estimated p-value =  0.001
```

Figure 2.1.13 Information on parametric bootstrapping during analyses of marginal items correlations. (Nsim = sample size, Inference type = M (marginal) or C (conditional), Nsteps = maximum iterative steps permitted during item parameter estimation)

The analysis of marginal item correlations during Rasch analysis is perhaps unusual, so it may be in order to point out first, that this is similar one of the fundamental approaches to analysis of confirmatory factor analysis, and second, that in this example, the overall χ^2 is more confident in rejecting the Rasch model than the conditional likelihood ratio test in Section 2.1.3.2. Whether this is a general result (because the power of the χ^2 in general is stronger than the power of the CLR test) is not known.

Another important point is that the analysis of marginal item correlation clearly identifies two pairs of items B & D and D&E with stronger correlations than expected by the Rasch model. During confirmatory Rasch analysis, one would immediately think of a multidimensional alternative to the one-dimensional factor analysis. This is also an option here, but during Rasch analysis we would also be concerned about local response dependence. We return to this problem in Section 2.1.3.6.

2.1.3.5 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. The results are shown in Figure 2.1.14.

```
Check assumptions of no DIF


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

Benjamini & Hochberg rejects at 0.00500
```

Figure 2.1.14 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 %.

According to Figure 2.1.12 there is only evidence of DIF for item C (DHP36) relative to F (Sex). Adjustment for multiple testing suggests that the evidence should be discarded, confirming the results of the over-all tests of no DIF.

2.1.3.6 Tests of local independence

DIGRAM uses Kelderman's (1984) test of local independence to test that this assumption of the Rasch model is not violated. To obtain these estimates you must select "Check local independence" from the list of options. The results are shown in Figure 2.1.15.

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

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. We return in a later tour to show you what to do about that. For the moment we only need to point out that the results support the weak evidence against homogeneity in Figures 2.1.8 and 2.1.9 and that both cases of local dependence involves the item (DHP38) that the analysis of the item-rest score correlation in Figure 2.1.10 insisted had too strong item discrimination. This finding illustrates two points made above: first, that evidence of strong item discrimination may be caused by local dependence and second, that evidence of too strong marginal item correlation (Figure 2.1.12) often pinpoints the same pairs of items as the tests of local dependence. What the real culprit is can not be answered without a more careful analysis of data and item contents.

We return to these problems during a third tour and proceed for now as if nothing was wrong in order to show how to estimate person parameters and assess reliability and targeting.

2.1.3.7 Estimating person parameters

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 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 these estimates, DIGRAM 1) assumes that the item parameters are estimated without error, 2) uses these parameters to calculate the score parameters, and 3) calculate maximum 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.

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

Figure 2.1.14 shows the ML estimates together with the exact interval 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 reparameterized 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 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, $-4.061 \in]-\infty, 2.752]$ and $3.573 \in [2.122, +\infty[$.

```
  +---------------------------------+
  |                                 |
  | 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 2.1.16 Maximum likelihood (ML) and weighted maximum likelihood (WML) estimates and exact interval estimates of person parameters.

The properties of the ML estimate are summarized in the table shown in Figure 2.1.17.
For each value of the person parameter estimate, the table provides information on the expected (true) score, test information, asymptotic and exact standard error, root mean squared error (RMSE), expected estimate and bias, and skewness and kurtosis of the distribution of the estimate.

The most important information is the bias and the RMSE since this is these two statistics that 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.

```
+--------------+
|              |
| ML estimates |
|              |
+--------------+

                          Properties of ML estimates

     Theta      True       Test    asymptotic                          exact
  estimate      score      info         se    E(theta)      Bias         se        RMSE  Skew  Kurt
-------------------------------------------------------------------------------------------------
    -4.061       0.25     0.233      2.073     -3.685      0.376      0.700      0.794  1.42  0.32
    -2.493       1.00     0.843      1.089     -2.766     -0.274      0.987      1.024 -0.18 -1.24
    -1.672       2.00     1.713      0.764     -1.949     -0.277      0.922      0.962 -0.85  0.33
    -1.203       3.00     2.597      0.621     -1.402     -0.199      0.805      0.829 -1.10  1.79
    -0.865       4.00     3.318      0.549     -0.992     -0.127      0.704      0.715 -1.11  2.63
    -0.585       5.00     3.787      0.514     -0.656     -0.071      0.625      0.630 -0.92  2.67
    -0.329       6.00     4.008      0.499     -0.361     -0.032      0.571      0.572 -0.61  2.06
    -0.082       7.00     4.049      0.497     -0.089     -0.007      0.540      0.540 -0.26  1.38
     0.167       8.00     3.992      0.501      0.177      0.010      0.533      0.533  0.09  1.21
     0.420       9.00     3.885      0.507      0.447      0.026      0.551      0.552  0.46  1.75
     0.683      10.00     3.713      0.519      0.735      0.052      0.597      0.599  0.83  2.48
     0.963      11.00     3.410      0.542      1.059      0.096      0.674      0.681  1.07  2.39
     1.279      12.00     2.903      0.587      1.444      0.166      0.775      0.793  1.02  1.18
     1.675      13.00     2.137      0.684      1.932      0.257      0.869      0.907  0.65 -0.38
     2.304      14.00     1.117      0.946      2.589      0.286      0.866      0.912 -0.07 -1.37
     3.573      14.75     0.265      1.943      3.279     -0.294      0.573      0.644 -1.56  0.84
-------------------------------------------------------------------------------------------------
```

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

The weighted ML estimate reduces both the bias and the RMSE for a relatively wide range of person parameter values.

Results on WML estimates are shown in Figures 2.1.18 and 2.1.19. Figures 2.1.18 shows the WML and calculates bias and RMSE at these values. Figure 2.1.19 provides information on bias and RMSE at the values of the ML estimates to make it easier to compare the performance of estimates. The WML estimate controls bias in much larger ranges of values than the ML estimate.

Finally, DIGRAM prints information summarizing other properties of the score over all items. This is shown in Figure 2.1.21. The information includes:

1) The test *difficulty* defined by the person value with an expected (true) score equal to half the maximum score.
2) The test *target* equal to the person value where test information is maximized.
3) Estimates of the mean and standard deviation of the person parameter distribution under the assumption that the distribution is normal.
4) Estimates of the exact reliability calculated under the assumption that the person parameter has a normal distribution with the estimated mean and variance.

.

```
  +---------------+
  |               |
  | WML estimates |
  |               |
  +---------------+

                            Weighted ML estimates

     Theta     True       Test    asymptotic                          exact
  estimate     score      info        se    E(theta)      Bias          se        RMSE  Skew  Kurt
-------------------------------------------------------------------------------------------------
    -3.512      0.41     0.371     1.641     -2.963     0.549      0.754      0.933  0.74 -1.19
    -2.064      1.43     1.208     0.910     -2.012     0.052      0.837      0.839 -0.71 -0.48
    -1.437      2.45     2.119     0.687     -1.445    -0.007      0.721      0.721 -1.15  1.65
    -1.066      3.38     2.893     0.588     -1.078    -0.013      0.633      0.633 -1.15  2.78
    -0.787      4.27     3.468     0.537     -0.794    -0.008      0.573      0.573 -0.94  2.85
    -0.545      5.15     3.835     0.511     -0.546    -0.001      0.533      0.533 -0.64  2.24
    -0.316      6.05     4.014     0.499     -0.313     0.003      0.509      0.509 -0.35  1.43
    -0.084      6.99     4.049     0.497     -0.081     0.003      0.496      0.496 -0.11  0.79
     0.156      7.96     3.995     0.500      0.158     0.001      0.495      0.495  0.07  0.53
     0.404      8.94     3.893     0.507      0.402    -0.002      0.503      0.503  0.22  0.64
     0.655      9.89     3.736     0.517      0.652    -0.003      0.522      0.522  0.39  0.96
     0.911     10.82     3.477     0.536      0.909    -0.002      0.550      0.550  0.56  1.15
     1.185     11.72     3.069     0.571      1.184    -0.002      0.585      0.585  0.65  0.84
     1.503     12.60     2.473     0.636      1.492    -0.011      0.621      0.621  0.56  0.04
     1.933     13.49     1.665     0.775      1.864    -0.069      0.633      0.637  0.23 -0.90
     2.803     14.43     0.632     1.257      2.375    -0.427      0.529      0.680 -0.66 -1.01
-------------------------------------------------------------------------------------------------
```

Figure 2.1.18 Weighted maximum likelihood estimates of person parameters.

Assessment of bias of WML estimates at the values of the ML estimates

Theta estimate	True score	Test info	asymptotic se	E(theta)	Bias	exact se	RMSE	Skew	Kurt
-4.061	0.25	0.233	2.073	-3.168	0.894	0.638	1.098	1.38	0.10
-2.493	1.00	0.843	1.089	-2.353	0.139	0.864	0.875	-0.30	-1.30
-1.672	2.00	1.713	0.764	-1.665	0.007	0.773	0.773	-1.02	0.77
-1.203	3.00	2.597	0.621	-1.215	-0.013	0.665	0.666	-1.18	2.46
-0.865	4.00	3.318	0.549	-0.874	-0.009	0.589	0.589	-1.02	2.93
-0.585	5.00	3.787	0.514	-0.587	-0.002	0.539	0.539	-0.69	2.38
-0.329	6.00	4.008	0.499	-0.327	0.002	0.510	0.510	-0.36	1.47
-0.082	7.00	4.049	0.497	-0.078	0.003	0.496	0.496	-0.10	0.79
0.167	8.00	3.992	0.501	0.168	0.001	0.495	0.495	0.08	0.53
0.420	9.00	3.885	0.507	0.418	-0.002	0.504	0.504	0.23	0.66
0.683	10.00	3.713	0.519	0.680	-0.003	0.524	0.524	0.41	1.00
0.963	11.00	3.410	0.542	0.961	-0.002	0.556	0.556	0.59	1.13
1.279	12.00	2.903	0.587	1.276	-0.003	0.597	0.597	0.64	0.63
1.675	13.00	2.137	0.684	1.649	-0.026	0.632	0.633	0.45	-0.39
2.304	14.00	1.117	0.946	2.123	-0.181	0.604	0.631	-0.13	-1.24
3.573	14.75	0.265	1.943	2.601	-0.972	0.393	1.049	-1.57	0.89

Figure 2.1.19 Properties of the weighted maximum likelihood assessed at the values of the ML estimate.

```
Test difficulty =       0.042
Test target     =      -0.132   Test information =   4.051

  +--------------------------------------------------+
  |                                                  |
  | Monte Carlo estimate of normal trait distribution |
  |                                                  |
  +--------------------------------------------------+

Latent trait:                Mean = -0.519    s.d. =  0.855
Fitted score distribution:   Mean =   5.54    s.d. =   3.34

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

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

Normal trait distribution with mean = -0.52  s.d.(trait) =  0.86


Expected True Score =   5.65   Variance =     8.14
Observed mean Score =   5.62   Variance =    11.36

                                  Ratio =     0.72

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

Figure 2.1.20 Difficulty, target and reliability

The test target is equal to -0.132 which is a little above the estimated mean of the population distribution of the person parameters. Reliability is equal to 0.72 or 0.71 (depending on the definition of reliability). 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 separate persons whereas targeting and reliability attempt to assess measurement quality from a population point of view. We pursue this point of view during the next (somewhat longer) guided tour through DIGRAM.

2.2 Rasch models. A longer tour.

During this tour we trace the same path with the same DIGRAM project as in the previous short tour, but this time we take the time to show you some 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,
4) how to create IRT and Rasch graphs,
5) how to assess targeting
6) how to test for unidimensionality

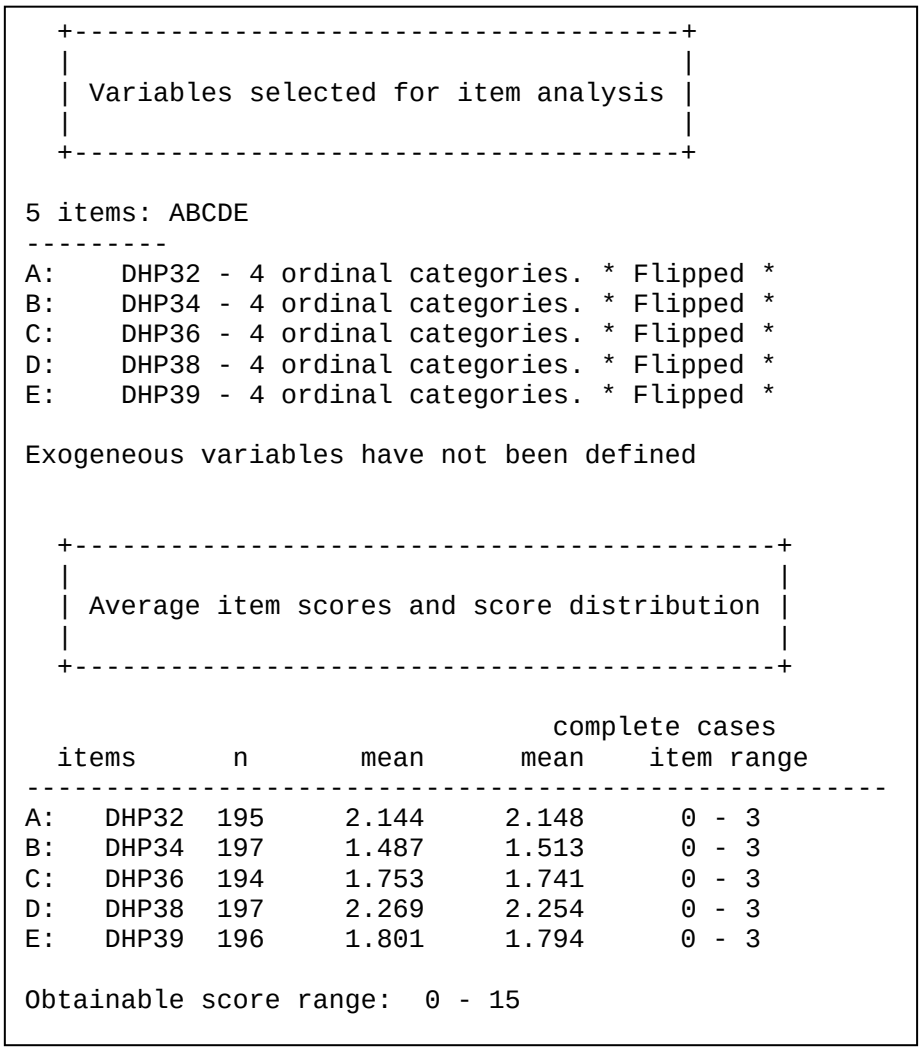
2.2.1 Changing the orientation of items

The DE 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 are free to change the orientation items. You can do this in two ways.

The first is to select items as before by followed by a "**FLIP**" command. The second is by adding a "-" to the items when you invoke the ITEMS command.

"**ITEMS ABCDE**" followed by "**FLIP**" is, in other words, the same as "**ITEMS -A-B-C-D-E**".

Figures 2.2.1 and 2.2.2 show the results of item flipping. Take a little time to compare the results with Figures 2.1.2 and 2.1.3. Note that Cronbach's Alpha and the variance of the total score is the same before and after items have been flipped.



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

Notes: The flipped item score is equal to 3 minus the original item score. The same is therefore true for the average items scores. From this it also follows that the correlation among items are the same as before flipping. It is this fact that implies that Cronbach's Alpha is unchanged after flipping (see Figure 2.2.2) because Alpha is a function of the correlations among 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.

```
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
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
Missing = 9
```

Figure 2.2.2 Information on the flipped score and score groups

Note: The flipped score is equal to 15 minus the original score. Since Var(15-X) = Var(X) it follows that the variance and standard deviation of the flipped score has to be equal to the variance and standard deviation of the original score.

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.

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[5], 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	$[s_0,s_1]$, $[s_1+1,s_2]$,.., $[s_{m-1}+1,s_m]$

Score groups are defined up to and including the cut points. "**CUT**" without parameters define 16 score groups, one for each separate score. "**CUT 9**" defines the two score groups shown in Figure 2.2.2. To define three score groups equal to 0-7, 8-10, and 11-15 you must invoke a "**CUT 0 7 10 15**" command. The result is shown in Figure 2.2.3.

Recall, that the most important role played by the score groups is during the conditional likelihood ratio test of the hypothesis that item parameters are the same for persons with high scores and persons with lower scores (Figure 2.1.9). Figure 2.2.9 below shows this test based on the flipped items and the redefined score groups.

[5] 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.

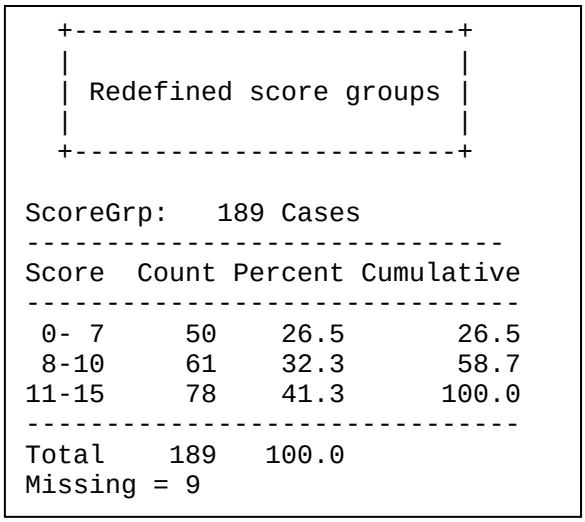


Figure 2.2.3 Score groups after "CUT 0 7 10 15"

2.2.3 Testing for DIF

Before we proceed to the analysis Rasch analysis after the GRM command, we want to show you another way to test for no DIF by tests for conditional independence in three way tables.

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, $Y_i \perp X_j | S$, is a hypothesis relating to a simple three way table where the association between the item and the exogenous variable is stratified according to 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 make it even simpler, 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[6]. 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

[6] "**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.

suggest that you use the facilities for analysis of score tables described during the detours in Section 3.4. For now, we only take a look at the summary results.

DIGRAM calculates χ^2 test 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:     SEX  26.6 26 0.429 0.667   0.07 0.595 0.524   21
G:     AGE  86.0 63 0.029 0.061   0.05 0.683 0.693 1000

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

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

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

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

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

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

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

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

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.

Comments on Figure 2.2.4: Evidence of DIF is disclosed in 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 tells the same story.

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


Analysis of DIF relative to F: 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


Analysis of DIF relative to G: 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
```

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

2.2.4 IRT and Rasch graphs

The tests for DIF described in the previous section are based on the fact that items and exogenous variables are conditionally independent given the total score. This result is well-known in the theory of Rasch models and can be proven in several ways. One easy way to prove it is to redefine the Rasch model as a chain graph model including items and exogenous variables together with the latent variable and the total score where it can be seen that the total score separates the items from all the other variables.

Ordinary graphical models of the kind that DIGRAM deals with only require one Markov graph to pinpoint the model's assumptions of conditional independency. Compared to this, graphical Rasch models are special because they require two Markov graphs that we refer to as IRT graph and Rasch graphs. As these become more and more important as we move along, it is convenient to take a first look at these graphs here.

Take a 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 2.2.6)[7].

The statistical model defined by this graph is an IRT model with local independence (because items are separated by the latent variable) and no DIF (items are separated from exogenous variables by the latent variable). It is for this reason that we refer to this graph as the IRT graph.

The Rasch model is an IRT model with a sufficient score. To show this property we add the total score to the graph in such a way that the total score separates items from the latent variable. Click the "Rasch" button to see this graph (Figure 2.2.7). In this graph, we have added undirected edges between the items because items are not conditionally independent given the total score. The total score on the other hand separates items from the latent variable and the exogenous variables from which it follows that the score is sufficient and that items are conditionally independent given the score.

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 next tour where they play a much more important role.

[7] 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.

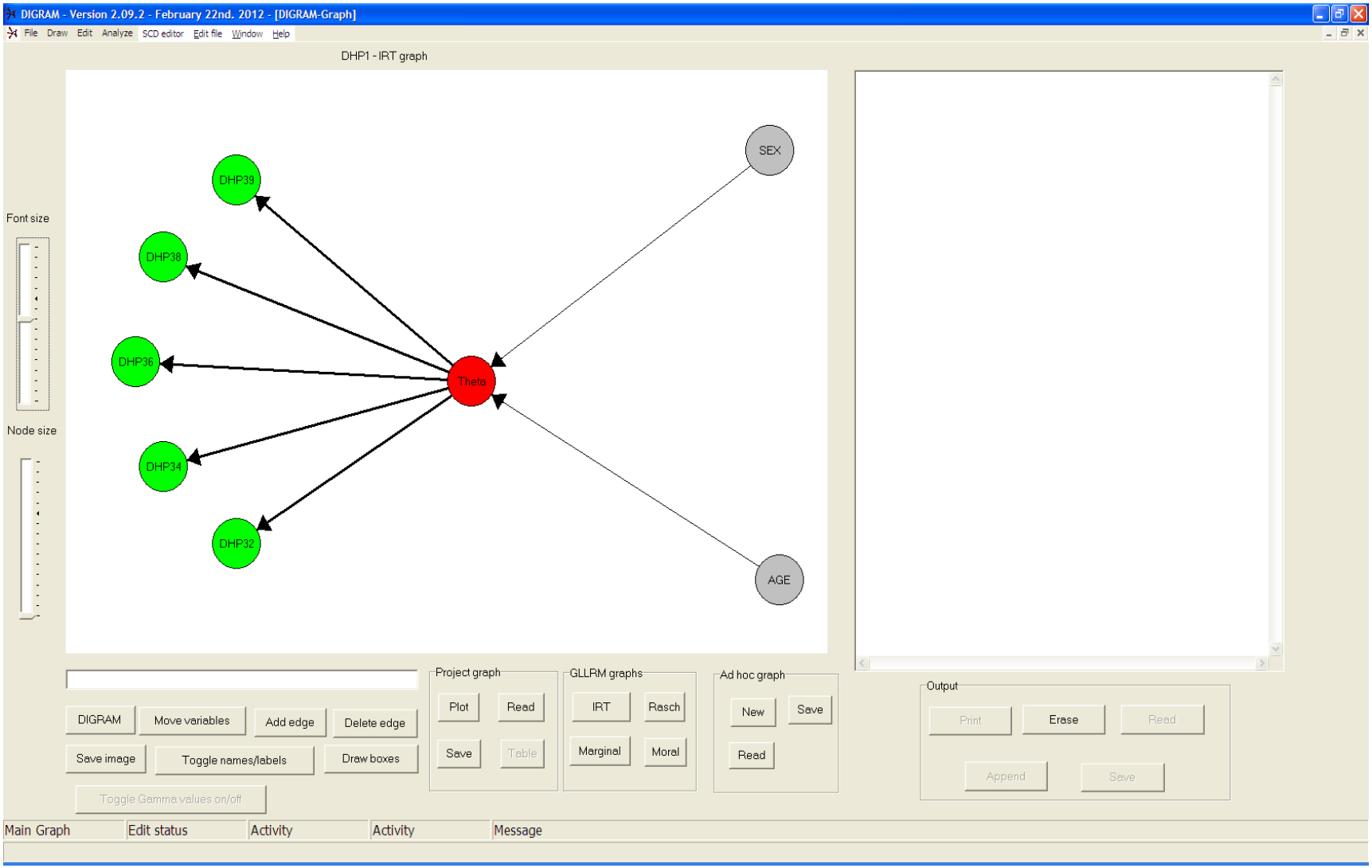


Figure 2.2.6. The IRT graph of the graphical Rasch model for the five DE items and with Age and Sex as exogenous variables. The graph has been edited

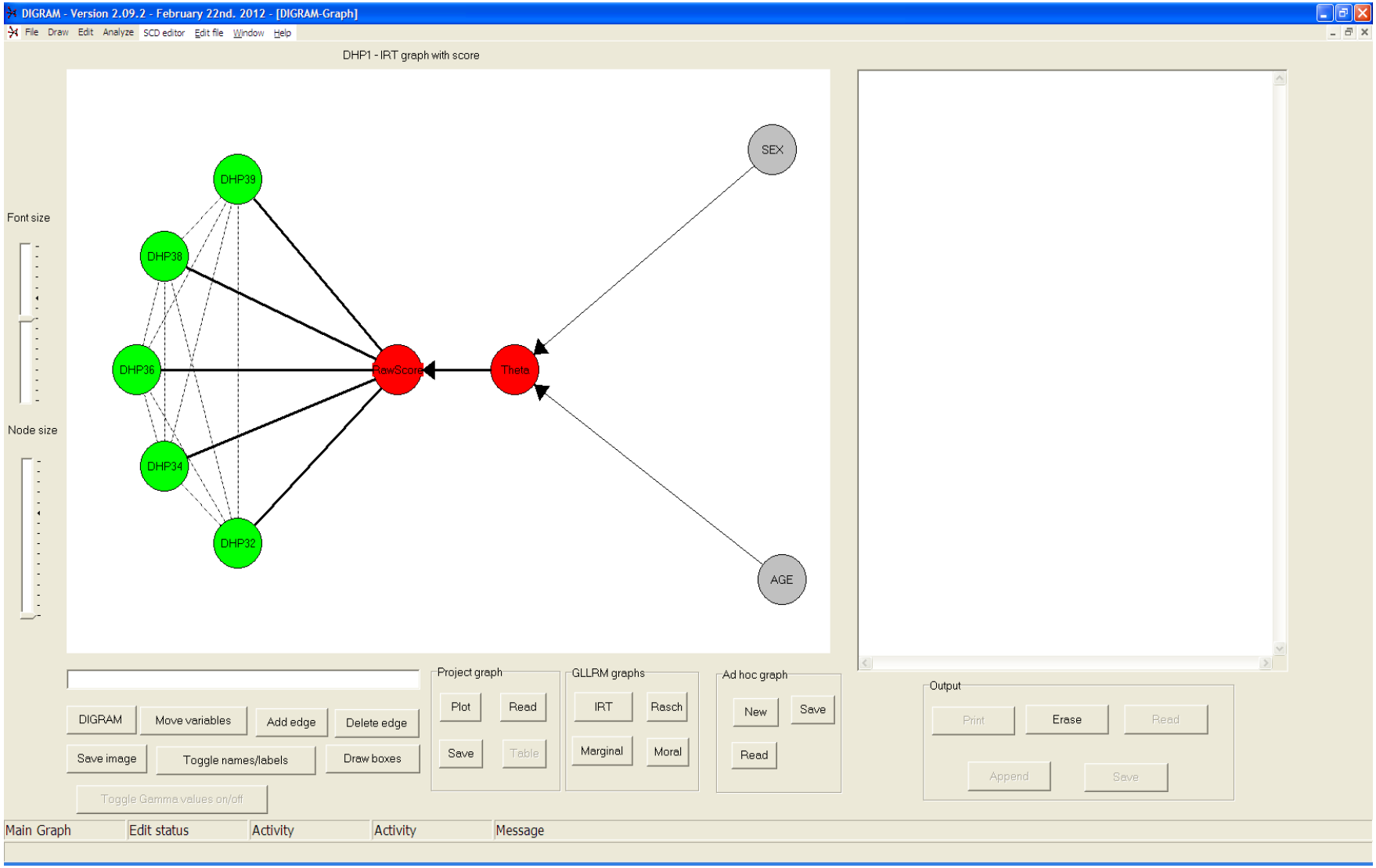


Figure 2.2.7. The Rasch graph of the graphical Rasch model for the five DE items with Age and Sex as exogenous variables. The graph has been edited

Press the “DIGRAM button to return to DIGRAM’s main form and press the “Graphical Rasch button” to kick off the parametric Rasch analysis.

2.2.5 Item analysis

2.2.5.1 Item parameter estimates, overall tests and person parameter estimates

Invoke the “**GRM”** command without parameters or click on the “Graphical Rasch model” button to initiate the item analysis. The GRM dialog box shown in Figure 2.1.5 turns up looking exactly as it did during the short tour except that there now are three rather than two score groups. Select “item fit statistics” together with “global tests of homogeneity and DIF” and compare the results with what we found during the short tour. Figure 2.2.8 shows the estimates of the item parameters of the flipped items and Figure 2.2.9 shows the summary of the overall tests of homogeneity and no DIF.

```
   +-----------------------------+
   |                             |
   | Item main effect thresholds |
   |                             |
   +-----------------------------+

The item difficulty is equal to the person parameter where the expected
item score is equal to the highest obtained item score divided by 2.


A -    DHP32   -1.15   -0.87    0.72   Difficulty = -0.53  (Exp. score = 1.5)
B -    DHP34    0.68 >  0.67 >  0.07   Difficulty =  0.49  (Exp. score = 1.5)
C -    DHP36   -0.95    0.91 >  0.31   Difficulty =  0.23  (Exp. score = 1.5)
D -    DHP38   -0.22 > -1.90    0.64   Difficulty = -0.70  (Exp. score = 1.5)
E -    DHP39    0.17 > -1.32    2.23   Difficulty = -0.04  (Exp. score = 1.5)

Nstep = 16   Delta =  0.0000517   14 estimated parameters

-ln(likelihood) =    577.748
            AIC =   1183.495
            BIC =   1227.882
                n =         176
```

Figure 2.2.8 Partial credit thresholds of flipped items

Compare the estimates of the flipped item parameters in Figure 2.2.8 with the estimates of item parameters in Figure 2.1.7. It is the same thresholds except that the sign of the thresholds have been changed. Notice also that the likelihood and the information criteria are the same indicating that it is the same model for the original items and the flipped items. These results extent to everything we saw before. Another illustration of this can be seen in Figure 2.2.9 where the overall tests of no DIF are the same as in Figure 2.1.8 for the original items. The test of homogeneity across score groups is

on the other hand different for the simple reason that Figure 2.1.8 compares item parameters in two score groups whereas Figure 2.2.9 compares item parameters in three score groups. The conclusion (weak evidence against homogeneity) is the same.

```
   +-------------------------------------------------------+
   |                                                       |
   | Overview of observed and expected average item scores |
   |                                                       |
   +-------------------------------------------------------+

   +----------------------+
   |                      |
   | Expected item scores |
   |                      |
   +----------------------+

C -    DHP36          SEX
    score    n      Male   Female
      0-7    47     0.90     0.86
     8-10    61     1.56     1.52
    11-15    67     2.40     2.40

   +------------------------------+
   |                              |
   | Observed average item scores |
   |                              |
   +------------------------------+

C -    DHP36          SEX
    score    n      Male   Female
      0-7    47     0.89     1.03
     8-10    61     1.45     1.77
    11-15    67     2.10     2.56


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

              CLR   df   p       delta
--------------------------------------
Score groups  48.2  28 0.010
F:      SEX   23.7  14 0.050
G:      AGE   42.4  42 0.454
```

Figure 2.2.9 Overall test of fit of the Rasch model. Flipped items and three score groups

When the number of score groups are larger than 2, DIGRA;M prints tables with the expected and observed item scores in groups defined by the score groups and exogenous variables. Figure 2.2.9 shows these tables for item C.

We leave the comparison of item fit statistics and person parameter estimates to yourself, but you will find that item fit statistics are the same and that person parameters are as before except that the signs of the estimates have changed.

2.2.5.2 Tests of unidimensionality

Recall the analysis of the original items disclosed evidence of local dependence between DHP34 (B) and DHP 38 (D) and between DHP38 (D) and DHP39 (E) (Figure 2.1.11). Evidence of local dependence is often caused by multidimensionality where more than one latent variable lies behind the responses to items. The natural next step in the analysis would be to test whether these three items measure a different latent variable than the two other items.

To check this we need a statistical test of the hypotheses that the two subsets of items measure the same latent variable. DIGRAM calculates two different tests[8] of this hypothesis if you select "Test unidimensionality" among the options in the GRM dialog box (Figure 2.1.5). Having done that, you have to tell DIGRAM that you want to test unidimensionality relative to items BDE on one hand and AC on the other. This is done in the little dialog box (Figure 2.2.10) that pops up where you write the subsets[9] of items separated by a '+'

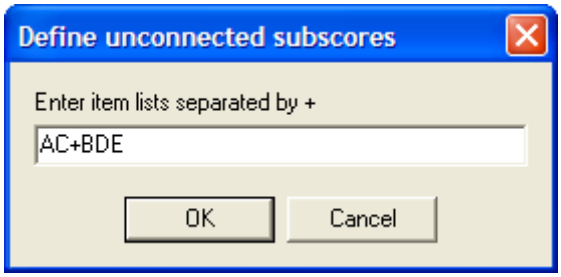


Figure 2.2.10 Definitions of subsets of items for tests of unidimensionality.

The first test of unidimensionality compares the observed and expected correlation of the subscores and concludes that the structure is multidimensional if the observed correlation is significantly lower than the expected correlation as evidence of multidimensionality. The second is a modified likelihood ratio test of the hypotheses that items come from a unidimensional Rasch model to the

[8] Information on these tests and additional references can be found in Chapter 9 of Christensen, Kreiner & Mesbah (2013)
[9] You can define more than two subsets of items.

alternative that responses come from two Rasch models depending on two different (but correlated) latent variables. Figures 2.2.11 – 2.2.17 show the output produced for these two tests.

Figure 2.2.11 shows the joint distribution of the two subscores while Figure 2.2.12 shows the expected distributions of the subscores under the Rasch model

Observed counts

Subscore 1: AC
Subscore 2: BDE

	Subscore 1						
Subscore 2	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 2.2.11 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 1						
Subscore 2	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

Figure 2.2.12 Expected joint distribution of subscores A+C and B+D+E.

Both tests of unidimensionality can be said to evaluate the differences between the observed distribution in Figure 2.2.11 and the expected distribution in Figure 2.2.12, expecting that multidimensionality will result in an abundance of persons with a relatively high score on one subscale and a relatively low score on the other. To give a first indication that this could actually be

the case, DIGRAM produces a table with standardized residuals comparing the observed and expected counts in the two tables. These residuals are shown in Figure 2.2.13.

Residuals (residuals between -0.01 and 0.01 are not printed)

Subscore 1: AC
Subscore 2: BDE

Subscore 2	Subscore 1: 0	1	2	3	4	5	6
0		-0.08		1.49	-0.29	-0.28	-0.10
1	0.08		-0.27	-0.58	**2.49**	-0.28	-0.25
2		-1.17	0.09	0.65	0.14	0.80	**2.59**
3	0.45	-0.85	-0.79	0.88	0.04	-0.84	-0.74
4	**2.11**	-0.79	-0.22	1.30	1.11	0.40	-1.33
5	-0.58	-0.90	-1.31	-0.67	-0.69	-0.04	0.28
6	-0.26	-1.06	-0.92	0.09	0.12	-0.75	-0.43
7	-0.33	-0.86	1.04	0.79	0.96	-1.92	-0.73
8	**4.56**	-0.46	-1.03	-1.03	1.87	-0.01	-0.79
9	-0.05	**5.83**	1.85	1.41	1.32	0.79	

Figure 2.2.13 Standardized residuals comparing observe and expected scores to the subscores. Significant residuals have been written in bold

Count outside 95% confidence region

Subscore 1: AC
Subscore 2: BDE

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

Count outside confidence regions

AC too low:

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

AC too high:

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

Figure 2.2.14 Observed distribution of subscores A+C and B+D+E outside the 95 % confidence region defined by the conditional distribution of the subscore given the total score

Finally, to further illustrate the extreme cases with discordant subscores, Figure 2.2.14 shows the table with observed counts outside the 95 % confidence region according to the conditional distribution of the subscores given the total score on all items. DIGRAM also counts the observed and expected numbers of cases outside the both the 95 % and 75 % confidence regions. Notice that the observed counts outside the 75 % and 95 % region for the subscores is a little higher than expected by the Rasch model.

If the structure underlying the item responses is two-dimensional we expect the number of discordant subscores to be too high and the correlation between the two subscores to be too low compared to what we would expect under the Rasch model. For this reason, DIGRAM calculates the observed and expected correlations based on Figures 2.2.11 and 2.2.12 together with the standard error of the correlation according to the expected distribution in Figure 2.2.12 and uses these results for a test of the difference between the observed and expected correlation. The asymptotic p-value of this test is based on the standard error, but DIGRAM suggest that you should use a Monte Carlo estimate of the p-value instead (Figure 2.2.15). It takes a little time, but we suggest that you always follow this advice.

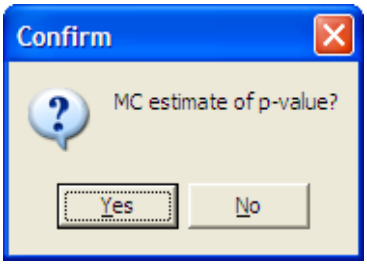


Figure 2.2.15 Monte Carlo estimate of p-value comparing observed and expected correlations between subscores.

The result is shown in Figure 2.2.16. The importance of using Monte Carlo estimates of p-values is obvious. In this case, there is a significant difference between the observed and expected correlation between the subscores.

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

Random gamma values

Estimate of p-value based on 1000 random tables p(below) =  0.042
```

Figure 2.2.16 Comparison of observed and expected correlations between subscores.

Despite the fact that the likelihood ratio test for unidimensionality in Rasch models dates back to at least Martin-Löf (1970) we are not completely happy with the standard implementation in DIGRAM because the extension of the test to polytomous items and the general class of loglinear Rasch model that you will visit during the next tour have created problems and because assessment of significance is a challenge. We have implemented parametric bootstrapping to provide better estimates of the true p-values. We will not illustrate these methods during this guided tour because they are very time-consuming, but they will be described in Part II of the introduction to item analysis in DIGRAM.

Figure 2.2.17 shows Per Martin- Löf's test of unidimensionality with the asymptotic p-value that we are concerned about. Unidimensionality is accepted[10].

```
PML test of unidimensionality:   z =    65.0  df = 51   p =  0.0898
```

Figure 2.2.17 Per Martin-Löf's test of unidimensionality.

[10] The bootstrap estimate of the p-value (based on a bootstrap sample of 400) is equal to 0.01, so it is a good idea to consider this possibility.

2.2.5.3 Item and test characteristic curves

Item characteristic curves (ICC) with expected item scores plotted against the values of the person parameter are very convenient if you want to illustrate how the Rasch model works and what may lie behind the numerical evidence provided by the item fit statistics. DIGRAM will not produce these curves for you, but will print information on a text file that you may put into a standard statistical program where the curves can be drawn.

If you select the “Export data for ICC curves” option in the GRM dialog box DIGRAM creates the text file and shows you where you can find it (Figure 2.2.18).

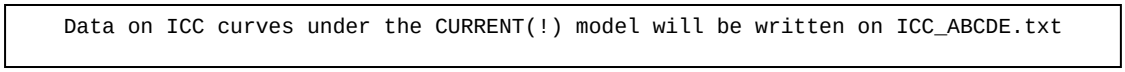


Figure 2.2.18 The file with data for IC curves.

Figure 2.2.19 shows the content of ICC_ABCDE.txt. Data is free formatted with variable names in the first row. The variables are,

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

```
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
-4,980 0,058 0,022 0,003 0,018 0,009 0,006 0 0
                    .........

-0,050 7,468 1,868 0,815 1,249 2,042 1,494 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
-0,020 7,589 1,889 0,847 1,275 2,062 1,517 0 0
-0,010 7,629 1,896 0,858 1,283 2,068 1,525 0 0
0,000 7,670 1,903 0,868 1,292 2,075 1,532 0 0
0,010 7,710 1,910 0,879 1,300 2,081 1,540 0 0
0,020 7,751 1,917 0,890 1,309 2,087 1,547 0 0
0,030 7,791 1,924 0,902 1,318 2,093 1,554 0 0
0,040 7,831 1,930 0,913 1,327 2,100 1,562 0 0
0,050 7,872 1,937 0,924 1,335 2,106 1,569 0 0
                    .........

4,980 14,896 2,986 2,992 2,990 2,987 2,940 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
-0,963 4 1,000 0,000 0,000 1,500 1,500 1 2
-0,683 5 2,000 0,333 0,778 1,111 0,778 1 9
-0,420 6 2,000 0,000 0,833 1,833 1,333 1 6
-0,167 7 1,818 0,500 1,409 1,818 1,455 1 22
0,082 8 1,962 0,577 1,577 2,000 1,885 1 26
0,329 9 2,235 1,412 1,353 2,412 1,588 1 17
0,585 10 1,889 1,889 1,889 2,389 1,944 1 18
0,865 11 2,409 2,000 2,000 2,636 1,955 1 22
1,203 12 2,409 2,636 2,091 2,818 2,045 1 22
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 2.2.19 The contents of file with data for ICC curves.

Figure 2.2.20 shows the ICC curves for the item where the item fit statistics (Figure 2.1.10) suggested that the item discrimination was stronger than expected by the Rasch model. This is easily recognized in the plot.

Figures 2.2.21 and 2.2.22 show the relationship between the person parameters and the expected scores (the true scores) on all items. The relationship between the observed scores and the maximum likelihood estimates of person parameters is also shown in Figure 2.2.23. Notice the almost linear relationship between the total score and the person parameters estimates for total scores equal between 3 and 12.

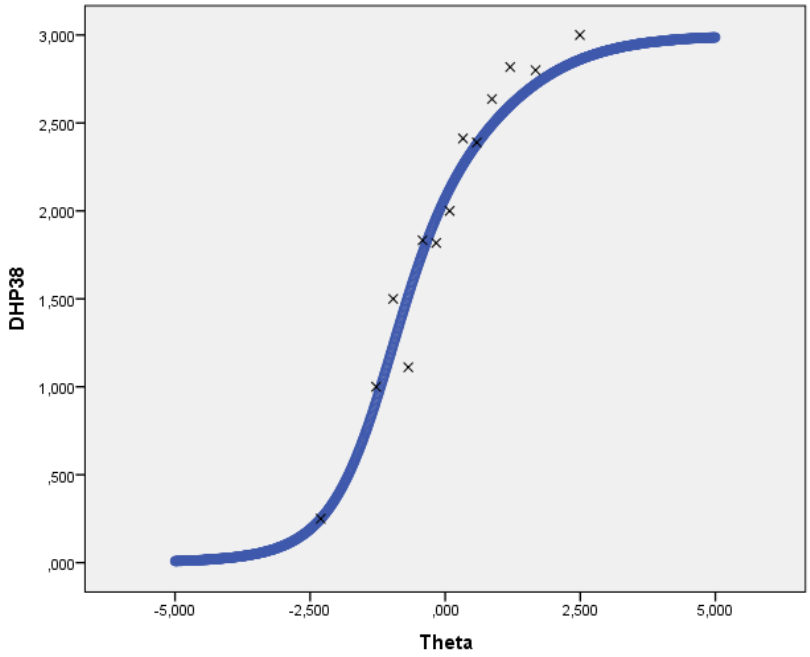


Figure 2.2.20 ICC curves for item D (DHP38). The x's show the average observed item scores for ML estimates of person parameters corresponding to the total score on all items

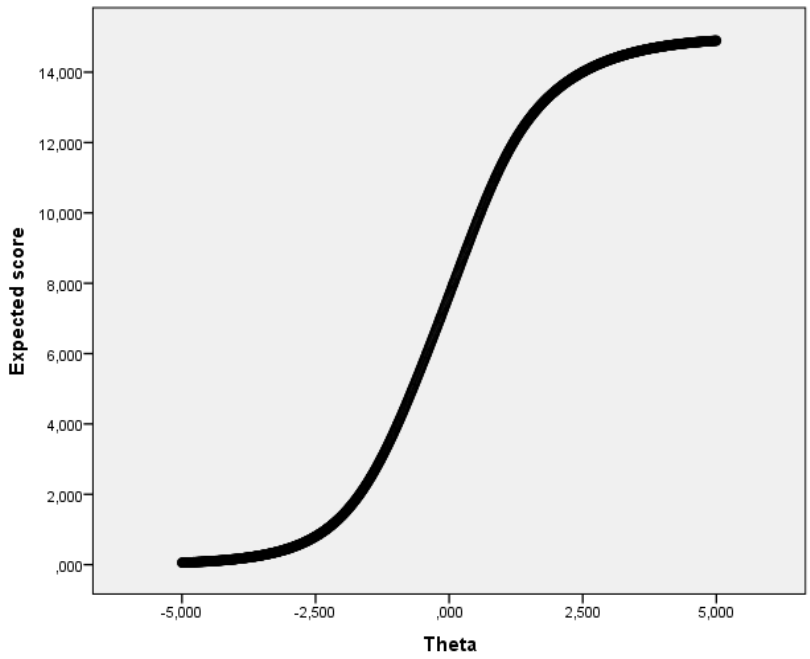


Figure 2.2.21 The test characteristic curve: the expected (true) score on all items plotted against the values of the person parameters.

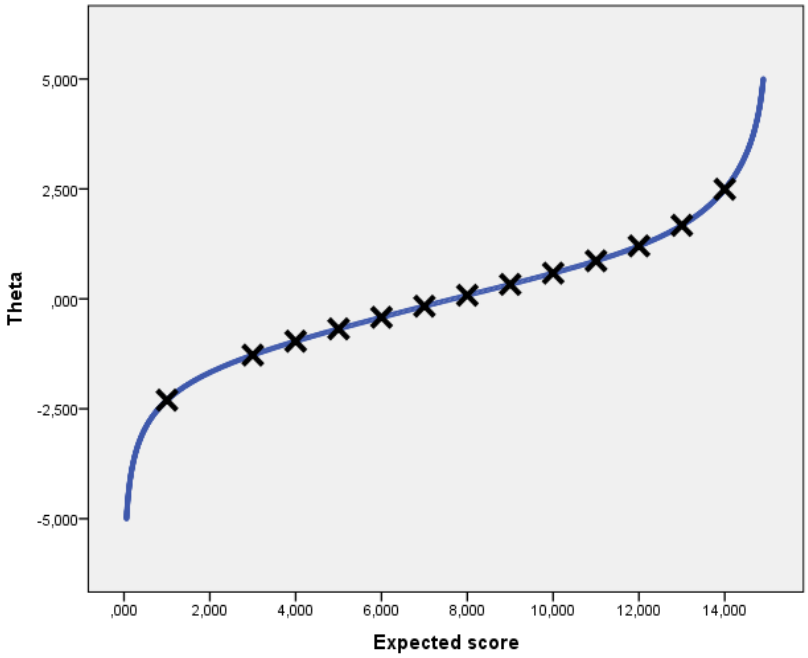


Figure 2.2.22 Person parameter values plotted against the expected (true) total score on all items. The X's shows the maximum likelihood estimates of person parameters corresponding to observed scores on all items.

2.2.5.4 Analysis of test information and targeting

The assessment of standard errors of measurement (Figures 2.1.17 – 2.1.19) describes the performance of the measurement function for specific person and shows that there can be large differences in measurement quality for different persons. Having recognized that, it is easy to imagine that measurement performance can be very different in different populations. To address this issue, DIGRAM examines the measurement performance in subpopulations defined by the outcomes on the exogenous variables providing a text file with input for so-called item maps and information on a number of test information statistics that are useful for assessment of targeting.

The file with input for item maps is called IMAP.txt. Figure 2.2.23 shows some of the contents on this file. Read this file into your favorite statistical program to draw maps like those shown in Figures 2.2.24 and 2.2.25. The item map plots 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.

```
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 2.2.23 The start of the IMAP.txt file. F and G are the exogenous variables, theta is the person parameter. Type indicates the ML estimates (0), the WML estimates (1), the estimated population distribution (2), the item parameter thresholds (3) and the score thresholds (4). The weight indicates the size of the bars of the item maps.

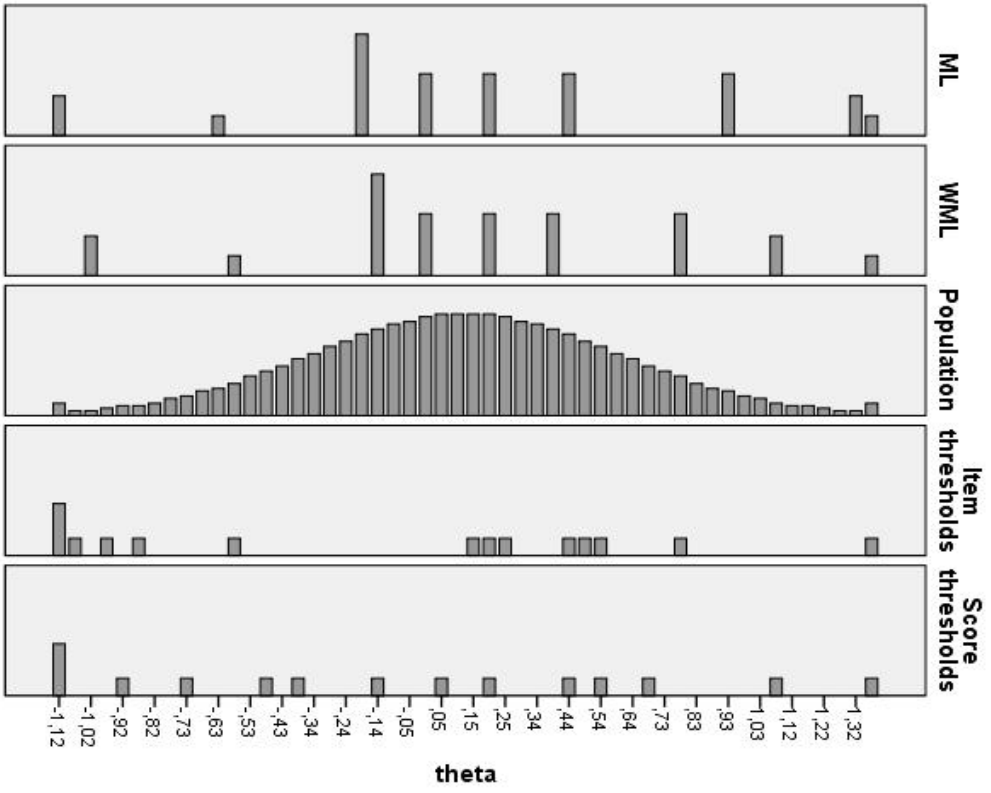


Figure 2.2.24 Item map for women, age 50-59. Person parameter mean = 0.59, sd = 0.59.

The range of the person parameter values is defined by the 99 % confidence region according the estimated person parameter distribution. The large columns to the left of the threshold distributions mean that many thresholds are smaller than the large majority of persons. The items appear to be somewhat out of target compared to this subpopulation.

In Figure 2.2.25, the range of person parameters cover the majority of the thresholds indicating that targeting is better relative to a population of 18-49 year old men. The implication of this becomes much more transparent when we look at the population characteristics of a number of statistics assessing different aspects of measurement quality[11]. These are collected in Figures 2.2.26 and 2.2.27 and summarized for all sub populations in Figure 2.2.28.

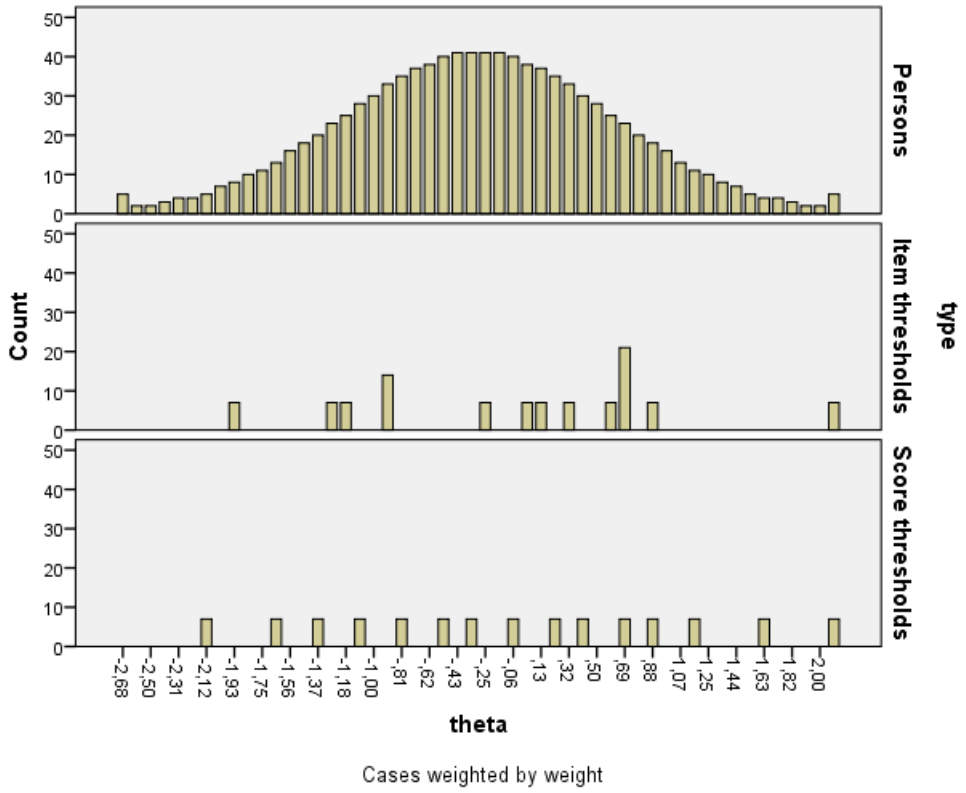


Figure 2.2.25 Item map for men, age 18-59. Person parameter mean = -0.29, sd = 0.90.

In each subpopulation DIGRAM calculates,

1. the mean and standard deviation of the total score,

2. estimates of the mean and standard deviation of the person parameter under the assumption that the distribution of the person parameter is normal,

[11] See Chapter 4 of Christensen, Kreiner & Mesbah (2013) for information on analyses of reliability and targetting in rasch models

3. the test *location* defined by the average of the thresholds of the partial credit distribution of the total score on all items, together with the test information and the expected standard error of measurement at the test location,

4. the test *difficulty* equal to the person parameter for which the expected total score is equal to half the maximum score on all items,

5. the test *target* equal to the person parameter for which the test information is maximized,

6. the expected bias, RMSE and standard error of measurement (SEM) of the ML estimate,

7. the expected bias, RMSE and standard error of measurement (SEM) of the WML estimate,

8. the mean and the standard error of the test information in the population together with a target index defined by the mean test information divided by the maximum obtainable test information,

9. the mean and the standard error of the SEM test information in the population together with a target index defined by the minimum SEM divided by mean of the SEM in the population,

10. parametric bootstrap estimates[12] of the reliability defined by the ratio between the variance of the true score and the variance of the score in the population,

11. parametric bootstrap estimates of the reliability defined by the test-retest correlation in the population,

12. parametric bootstrap estimates of the reliability measured by the correlation between the score and the true score in the population,

13. parametric bootstrap estimates of the probability that two random persons drawn from the distribution are correctly ordered by the total score on all items,

14. parametric bootstrap estimates of the probability that two random persons drawn from the population have the same scores on all items,

[12] based on bootstrap samples of 10,000 persons from the normal distribution with mean and variance equal to the estimates in point 2

15. the average bias of so-called interval estimates of the difference between two independent persons drawn randomly from the study population.

Note that the location, difficulty and target are person parameters values. These values together with the obtainable test information and SEM are the same in all groups if there is no DIF. All other results reported during the analysis of targeting depends on the mean and standard deviation of the person parameter distribution in the different subpopulations.

The two subpopulations described in Figures 2.2.26 and 2.2.27 are the extreme subpopulations in terms of reliability. Reliability is weak among females at age 50-59 and better among men at age 18-49 where the standard deviation of the person parameter is much larger. The example illustrates the fact that reliability depends as much on the distribution of persons as on the qualities of measurement. In this case, where we assume that the same Rasch model fits men and women at all ages, and where targeting according to the target indices is the same in both groups[13], the considerable difference in reliability is due to that fact that the standard deviation of the person parameter among women at age 50-59 is relatively small (sd = 0.59) while the than the standard deviation among men at age 18-49 is much larger (sd = 0.896).

Concerning the estimates of the person parameters is clearly shown that the WML estimate is superior to the ML in these populations and that targeting measured by the targeting indices is generally good. The differences in targeting appear to have little impact on the precision of the measurement as measured by the RMSE which is best among men at age 18-59.

Notice finally, that targeting was assessed in each subpopulation defined by Sex and Age because DIGRAM assumes the distribution of the person parameter depends on the exogenous variables. If you want an overall assessment you have to tell DIGRAM that the person parameter does not depend on these variables. How to do that will be illustrated in the long tour through the graphical loglinear models.

[13] Notice also, that the mean of the person parameter among women at age 50-59 lays 0.455 above test target whereas the mean of person parameter lays 0.420 below test target among men at age 18-49.

```
Group: SEX = Female   AGE = 50-59

  33 persons.   Mean score =   9.85  sd =  2.74   Mean Theta =  0.587   sd =  0.585

Location        =      -0.000   Test information     =   4.037   SEM =   0.498
Test difficulty =      -0.042   Test information     =   4.028   SEM =   0.498
Test target     =       0.132   Max test information =   4.051   SEM =   0.497

 **** ML estimates  ****
Mean bias =      0.096  sd =   0.082
Mean RMSE =      0.671  sd =   0.129
Mean SEM  =      0.662  sd =   0.117

 **** WML estimates ****
Mean bias =      0.006  sd =   0.009
Mean RMSE =      0.571  sd =   0.082
Mean SEM  =      0.571  sd =   0.082

Mean test information =   3.439  sd =   0.707   Target index =   0.849

Mean SEM              =   0.552  sd =   0.081   Target index =   0.901

var(true score)/var(score)     =  0.543
test-retest correlation        =  0.556
test-true score correlation    =  0.720

Probability of correct person separation =  0.705
Probability of no person separation      =  0.106

Bias of interval estimates
          ML estimate   : 0.0704    WML estimate  : 0.0850
```

Figure 2.2.26 Test information and targeting for women, age 50-59

```
Group: SEX = Male   AGE = 18-49

  17 persons.   Mean score =   6.65  sd =  3.62    Mean Theta = -0.288   sd =  0.896

Location        =      -0.000   Test information     =   4.037   SEM =   0.498
Test difficulty =      -0.042   Test information     =   4.028   SEM =   0.498
Test target     =       0.132   Max test information =   4.051   SEM =   0.497

 **** ML estimates  ****
Mean bias =      0.090  sd =   0.086
Mean RMSE =      0.660  sd =   0.131
Mean SEM  =      0.651  sd =   0.118

 **** WML estimates ****
Mean bias =      0.012  sd =   0.045
Mean RMSE =      0.553  sd =   0.066
Mean SEM  =      0.551  sd =   0.064

Mean test information =   3.419  sd =   0.757   Target index =   0.844

Mean SEM              =   0.558  sd =   0.108   Target index =   0.890

var(true score)/var(score)     =  0.730
test-retest correlation        =  0.734
test-true score correlation    =  0.841

Probability of correct person separation =  0.791
Probability of no person separation      =  0.075

Bias of interval estimates
          ML estimate   : 0.1153    WML estimate  : 0.0555
```

Figure 2.2.27 test information and targeting for men, age 18-49

+-----------------------------+
| Summary of test information |
+-----------------------------+

				theta		test information		target	RMSE(WML)		target		separation	
SEX	AGE	Target	n	Mean	sd	Mean	max	index	Mean	min	index	reliability	prob
Female	18-49	0.13	14	0.31	0.83	3.420	4.051	0.844	0.568	0.497	0.875	0.710	0.775
Male	18-49	0.13	17	-0.29	0.90	3.419	4.051	0.844	0.553	0.497	0.899	0.734	0.791
Female	50-59	0.13	33	0.59	0.59	3.439	4.051	0.849	0.571	0.497	0.870	0.556	0.705
Male	50-59	0.13	23	0.33	0.67	3.575	4.051	0.882	0.552	0.497	0.899	0.615	0.745
Female	60-69	0.13	47	0.80	0.84	3.026	4.051	0.747	0.618	0.497	0.804	0.682	0.760
Male	60-69	0.13	42	0.51	0.82	3.294	4.051	0.813	0.585	0.497	0.850	0.702	0.767

Figure 2.2.28 Targeting summary

2.3 Graphical loglinear Rasch models. The Short tour.

On this tour we return to the unflipped version of the DE subscale. The initial analysis provided evidence suggesting 1) DIF of item C (DHP36) relative to F (Sex), 2) local dependence (LD) between items B (DHP34) and D (DHP38) and items D (DHP38) and E (DHP39), and 3) that the item discrimination of item D was stronger than expected by the Rasch model[14].

2.3.1 Definition of graphical loglinear Rasch models

The evidence against the Rasch model is so comprehensive that the scale won't survive attempt to purify it by elimination of items. A better option is to attempt to fit a graphical loglinear Rasch model (GLLRM) where uniform DIF and uniform local dependence[15] is accepted, since such a model possess all the fundamental properties of Rasch models derived from the sufficiency of the total score.

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

There are three ways for you to tell DIGRAM that this is the model you want to use: 1) you can invoke the GRM command with the three pairs of variables as parameters, 2) you can invoke the GRM command without parameters or click on the GRM button and then add the three pairs of variables to the "New model" field in the GRM dialog box or 3) you can 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.

[14] Figures 2.1.8 – 2.1.10.

[15] 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.

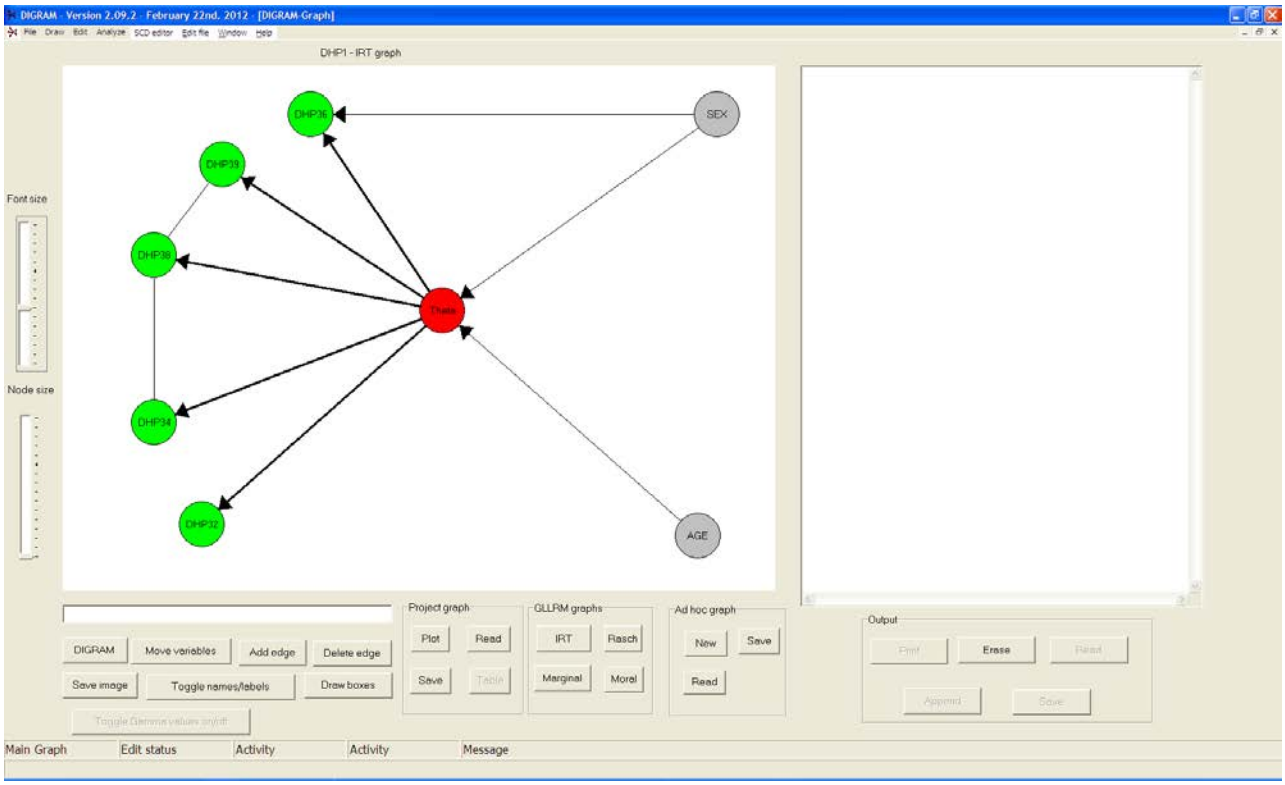


Figure 2.3.1 IRT graph of a GLLRM with DIF and local dependence. The graph has been edited to make it easier to read.

Before we go into the details of item analysis by GLLRMs it is useful to take a closer look at the GRM dialog box (Figure 2.3.2) that turns op 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 either by the edited IRT graph or by the parameters added to the GRM command. This is the model that DIGRAM has saved and DIGRAM assumes that this is your preferred model until you specifically tell the program that you want to revise the model.

The field with the current model cannot be edited. However, the model analyzed in the GRM dialog is the model defined in the “New model” field. To begin with DIGRAM copies the current model to the new model assuming that you want to examine the current model, but if you for some reason want to examine a different model you have to define this model in the “new model” field by adding and/or deleting model terms.

If you, having examined the new model, want to save this model as the current model, you must click on “Change model”. 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.

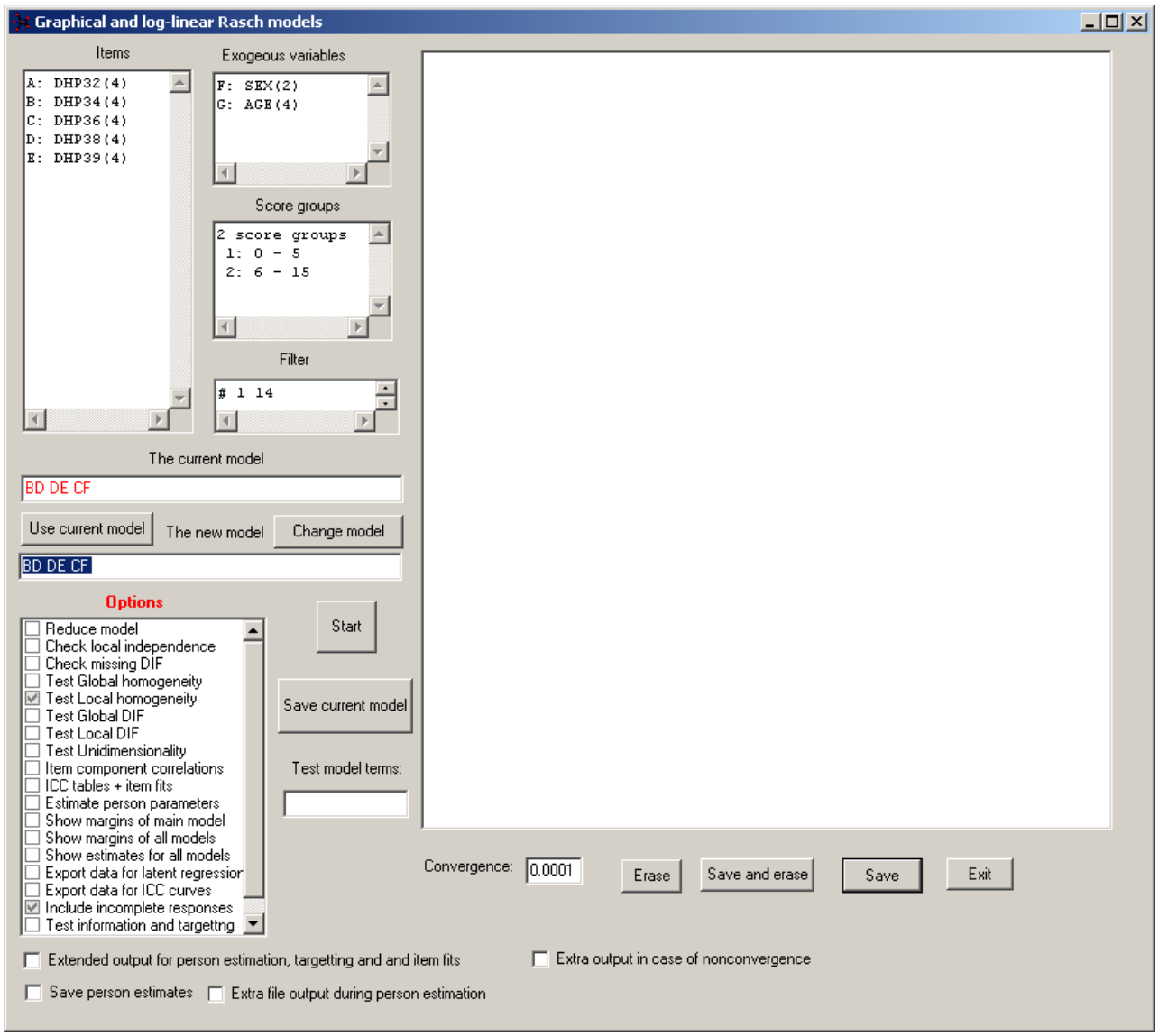


Figure 2.3.2 The GRM dialog form following either editing of the IRT graph or a “GRM BD DE CF” command.

Since we have just defined the current model we continue to use this model and do not attempt to define another new model here.

2.3.2 Item analysis

As long as you do not consider other models, the analysis proceeds in exactly the same way as for the GLLRM as for the ordinary Rasch model. Click start to estimate the parameters, select global

tests of homogeneity and DIF to obtain overall tests of fit, select item fits to obtain the same fit statistics as for the Rasch model, select check local independence and missing DIF to make sure that there is no evidence of DIF and LD above and beyond the DIF and LD in the model, select person parameter estimates to obtain person parameter estimates and select test information and targeting to assess the appropriateness of the items for the current study population. Everything works exactly as before except that item parameter estimates have to take the DIF and local dependence into account and that DIF has an effect on the targets of the items.

Some of the output is shown and commented upon in Figures 2.3.3 and 2.3.4, but we suggest that you to start try all these procedures and compare the results to the results obtained during the first short tour through the Rasch model analysis.

The multiplicative Item parameter estimates are shown in Figure 2.3.3.

```
       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

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

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

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

Figure 2.3.3 Estimates of 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 1. The interaction parameters are fixed in such a way that the interaction parameters corresponding to reference categories for both variables in an interaction terms is equal to 1 so that the parameters can be interpreted as odds-ratio coefficients.

One of the advantages of the multiplicative parameterization is that response categories that are not used do not create numerical problems because the maximum likelihood estimates of the parameters are equal to zero. DIGRAM prefer 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. This is the case for both sets of interaction parameters relating to local dependence, where it has to be admitted that the interpretation of the parameters is less than transparent.

One useful property of GLLRMs is that the main effect parameters and the interaction parameters under the multiplicative model can be easily reparameterized in terms of partial credit models.

First, it follows from the GLLRM that a DIF item has a partial credit distribution in each subgroup defined by outcomes on the sources of DIF. According to the model that we are using, Item C (DHP36) functions differently for men and women which means that this item is a partial credit item for both men and women except but that the thresholds are different.

Second, again under a GLLRM, it follows that the total score over all items belonging to the same item component[16] has a partial credit distribution. Items B, D and E (DHP34+DHP38+DHP39) is one such component for which reason, the subscore B+D+E is a composite partial credit item with thresholds depending on the multiplicative parameters in Figure 2.3.3.

Figure 2.3.4 shows the thresholds of the items under the GLLRM defined above. Note that there is one ordinary partial credit item A (DHP32), one item (DHP36) with somewhat different thresholds for men and women (the second and third thresholds are similar, but there is a big difference

[16] An item component is a subset of directly or indirectly connected items.

between the first threshold for men and women), and one composite item B+C+D (DHP34+DHP38+DHP39)

```
A -    DHP32          -0.56      0.86      1.06   Difficulty =    0.53

C -    DHP36
     F =     Male      -0.62 >   -0.87      0.75   Difficulty =   -0.36
     F =   Female       0.50 >   -0.85      0.99   Difficulty =    0.08
  +---------------------------------------------------------------------+
  |                                                                     |
  | Thresholds of item components defined by local response dependence  |
  |                                                                     |
  +---------------------------------------------------------------------+

B-DHP34 & D-DHP38 & E-DHP39

   Component scores from 0 to 9.  (Difficulty at expected score = 4.5)
   Thressholds:  -1.53 -0.52 -0.31 > -0.62 -0.36  1.01 > -0.22  1.17 >  0.76
   Difficulty =   -0.05
```

Figure 2.3.4 Partial credit thresholds (output has been edited to make it fit the page)

As for Rasch models, you can select “Export data for ICC curves” if you want to plot the ICCs, but the text file containing information on ICC curves is a little more complicated for a GLLRM than for the ordinary Rasch model, providing information on both item components with locally dependent items and on DIF.

The text file defined by the BD,DE,CF model provides 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 2.3.5 shows the ICC curves for F = 1. Note that the ICC curves are steeper for the locally dependent items than for the pure Rasch items.

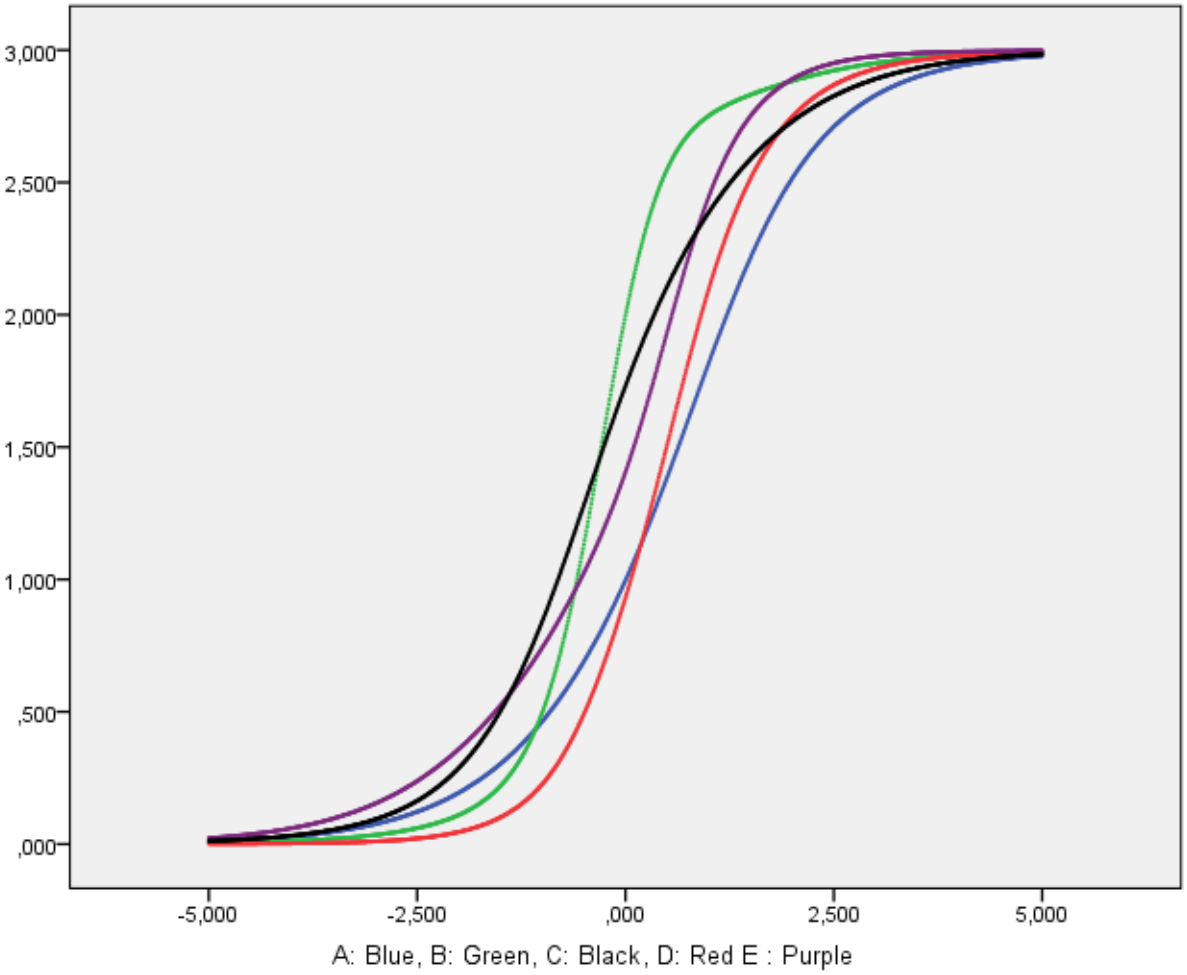


Figure 2.3.5 ICC curves for F = 1

Figure 2.3.6 shows the over-all fit statistics. 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:      SEX   25.2  24 0.394
G:      AGE   99.9  90 0.223     0.001
```

Figure 2.3.6 Overall fit statistics of the BD,DE,CF model

Figure 2.3.7 shows the item fit statistics. Compare these results item fit statistics derived under the Rasch model. Infits and outfits under the GLLRM are closer to 1 and the expected item restscore

correlation is closer to the observed, and all fit statistics are clearly insignificant. There is no evidence of a stronger discrimination of item D under this model.

```
  +--------------------------------+
  |                                |
  | 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 restscore association |
  |                            |
  +----------------------------+

                  Item-restscore 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
-------------------------------------------------
```

Figure 2.3.7 Item fit statistics under the BD,DE,CF model

2.3.3 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 using the model without the term as the null-hypothesis and the "new" model as the alternative. Select "Reduce model" to get these tests. The results are shown in Figure 2.3.8. All hypotheses are rejected, but it has to be admitted that the evidence of DIF is not strong (p = 0.024)

```
Test local dependence

B & D:   lr =   40.59  df =   9  p = 0.0000
D & E:   lr =   37.76  df =   9  p = 0.0000

               ---

Test of no DIF

C & F:   lr =    9.44  df =   3  p = 0.0240
```

Figure 2.3.8 Confirmatory tests of DIF and local dependence

2.3.4 Person estimation and targeting in GLLRMs

Figure 2.3.9 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 in this figure you therefore have to change the sign of the target value and the signs of the means of the population parameters in this figure. Relative to the original orientation of the items, the target (the person parameter where the test information is maximized) under the Rasch model is equal to -0.13 and the means of the person parameter are negative in all subpopulations defined by the Sex and Age except for females with age from 18 to 49 where the person parameter mean is equal to 0.29. All other parameters included in the assessment of test information and targeting do not depend on the orientation of the items so that this parameters are directly comparable between Figure 2.2.28 and Figure 2.3.9.

Due to DIF, person parameter estimates will be different for men and women. We do not show the estimates here, but after the tables with the person parameters estimates, DIGRAM prints a table (Figure 2.3.10) with equated scores where scores for groups are adjusted to scores for the first group. In the score range from 1 to 10, the scores for men (F=2) should be increased with 0.20 – 0.50 points to be comparable to the scores for women (F=1).

Summary of test information

SEX	AGE	Target	n	theta Mean	theta sd	test information Mean	test information max	target index	RMSE(WML) Mean	RMSE(WML) min	target index	separation reliability	separation prob
Male	18-49	-0.06	14	-0.19	0.65	4.215	5.052	0.834	0.514	0.445	0.866	0.642	0.760
Female	18-49	0.01	17	0.38	0.69	4.293	5.307	0.809	0.534	0.434	0.812	0.678	0.769
Male	50-59	-0.06	33	-0.43	0.42	4.296	5.052	0.850	0.513	0.445	0.867	0.436	0.676
Female	50-59	0.01	23	-0.11	0.46	4.733	5.307	0.892	0.526	0.434	0.825	0.502	0.701
Male	60-69	-0.06	47	-0.59	0.62	3.761	5.052	0.744	0.550	0.445	0.809	0.599	0.731
Female	60-69	0.01	42	-0.25	0.57	4.371	5.307	0.824	0.553	0.434	0.785	0.590	0.736

Figure 2.3.9 Targeting and test information under the BD, DE, CF model

DIF equated scores

DIF sources: F - SEX

score	F 1	F 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

Figure 2.3.10 DIF equated scores and assessment of test bias

2.3.5 Saving the model

Item analysis by GLLRMs can be quite time-consuming. In order 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: either 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,CF) 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
GRM BD DE CF
```

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

2.4 Graphical loglinear Rasch models. The longer tour.

On this tour, we abandon the DHP1 project and instead turn to the PF3 project with data on the physical functioning subscale of the SF36 inventory where we, to simplify things, exclude item A from the analysis and therefore only consider 9 out of 10 PF3 items. For numerical reasons, described at the end of Section 2.4.2 we also flip items, so that a high score implies physical impairment.

The items were described in Section 1.3.2. Before we start we suggest that you take a careful look at them again. We are sure that you will agree that items are phrased in such a way that local *response* dependence has to be expected for some pairs of items (e.g. PF4 & PF5, and PF7 & PF8 & PF9). For this reason, we could argue that it makes no sense to hope for anything looking like a Rasch model for these items, but we will nevertheless perform the analysis as if we had failed to think about the local dependence issue to show you that the methods implemented in DIGRAM will discover what we have overlooked.

Assume therefore that you expect items to fit a Rasch model, but that you prefer to attempt to replace it with a GLLRM to preserve sufficiency and essential validity and objectivity rather than eliminate all the items that do not fit the Rasch model. Unfortunately, it is often a challenge to identify an adequate GLLRM because of the very large number of potential GLLRM. To be able to solve this problem in finite time we therefore need a systematic approach to analysis by GLLRMs.

Such an approach has been implemented in DIGRAM. It consists of two parts. The first is a strategy for initial item screening that test for DIF and local dependence and eliminates what appears to be spurious evidence and proposes a starting point for the second part. The second consists of stepwise model search for an adequate and parsimonious model where interaction terms are added to and/or deleted from the model suggested by item screening and where the GLLRM defined by the local dependence and DIF found during item screening.

2.4.1 Item screening

Assume that we have defined the items and the two exogenous variables (K = sex and L = Age). Item screening is invoked by a "**SCREEN I**" command. The initial item screening is in itself a

stepwise procedure and the output from this analysis is extensive and provides information on every step taken during the screening. In most cases there will, however, be no reason to look at anything but the final part where the results are summarized and a GLLRM is proposed. So let's take a look at the summary of the screening of the PF3 items first, followed by a look at the model defined by the screening and some of the results leading to this model.

Figure 2.4.1 summarizes the result of the item screening and Figure 2.4.2 shows the IRT graph of the GLLRM proposed by the result.

```
Item screening has defined the following GLLRM: BC DE GH HI IJ CK

The score is associated with the following exogenous variables : K L

Local dependent items. 4 item components: BC DE F GHIJ
```

Figure 2.4.1 Summary of results of item screening of PF3 items.

So what does the summary of the item screening tell us?

First, the interaction terms "`BC DE GH HI IJ CK`" imply that item screening has found 5 pairs of locally dependent items (including the three pairs of locally dependent items that we expected to find) and that item C functions differentially relative to K,.

Second, item screening concludes that the total score appears to be associated with K (sex) and L (age).

The final part tells us that the local dependence defines four item components consisting of items that are connected in the IRT graph in Figure 2.4.2. The first consists of items B and D, The second of items D and E, the third of item F, and the fourth of items G, H, I, and J.

We refer to Kreiner & Christensen (2011) for details on item screening and another example with PF items. The rest of this section provides some details, but if you can do without these details for now we suggest that you skip these details and proceed to Section 2.4.2 to see what happens after screening.

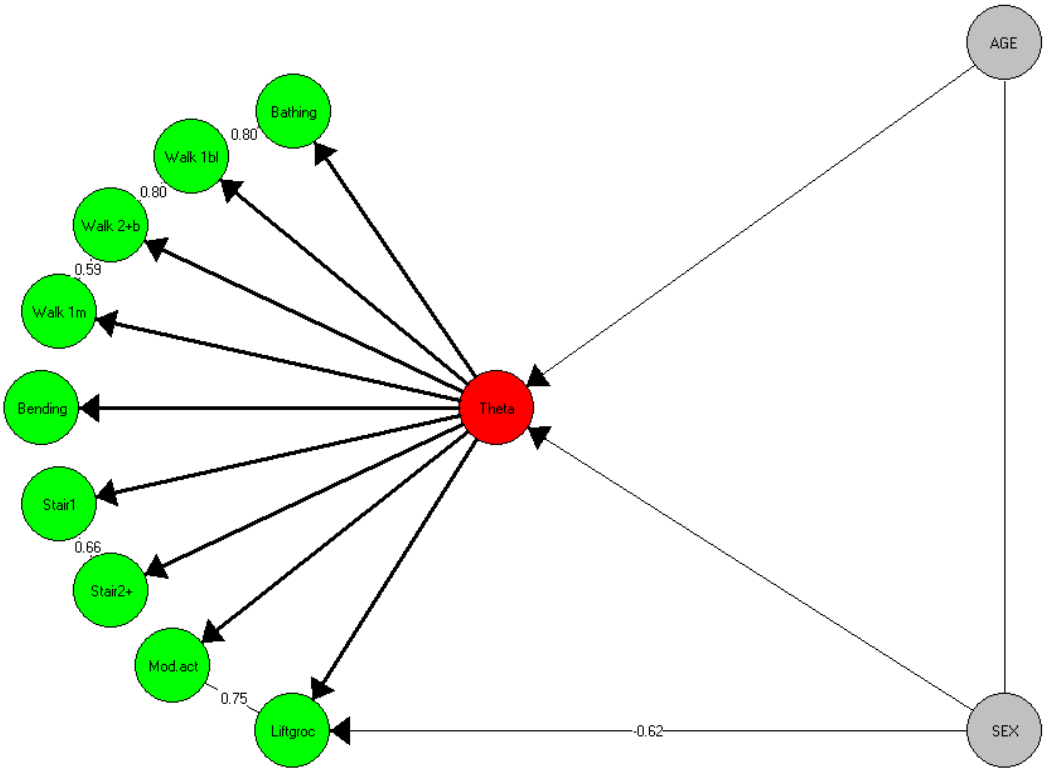


Figure 2.4.2 The IRT graph of the model defined by screening of nine PF3 items. The numbers on the edges of the graph are partial Gamma coefficients measuring the strength of the association among items and DIF sources. Associations are very strong.

The first part of the screening consists of an analysis of marginal association between the items, the total scores and the exogenous variables. Figures 2.4.3 – 2.4.7 give the results:

Figure 2.4.3 examines the marginal associations among items and exogenous variables.

Figure 2.4.4 examines the partial associations among items and exogenous variables given scores and restscores for tests of local independence and no DIF.

Figure 2.4.5 eliminates spurious evidence of local dependence

Figure 2.4.6 eliminates spurious evidence of DIF

Figure 2.4.7 examines the association between the total score and the exogenous variables.

```
  +----------------------------------------+
  |                                        |
  | Screening of marginal item relationships |
  |                                        |
  +----------------------------------------+


p-values are two-sided and exact(Nsim = 1000)

                                                                                  rest
                     B      C      D      E      F      G      H      I      J     score
-------------------------------------------------------------------------------------------
B  Mod.act Gamma           0.973  0.913  0.949  0.868  0.920  0.967  0.958  0.914  0.932
              p            0.000  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  0.888  0.951  0.957  0.936  0.916
              p     0.000         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  0.858  0.921  0.919  0.912  0.860  0.885
              p     0.000  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         0.923  0.927  0.974  0.983  0.970  0.969
              p     0.000  0.000  0.000         0.000  0.000  0.000  0.000  0.000  0.000

F  Bending Gamma    0.868  0.848  0.858  0.923         0.881  0.909  0.922  0.901  0.855
              p     0.000  0.000  0.000  0.000         0.000  0.000  0.000  0.000  0.000

G  Walk 1m Gamma    0.920  0.888  0.921  0.927  0.881         0.965  0.936  0.865  0.900
              p     0.000  0.000  0.000  0.000  0.000         0.000  0.000  0.000  0.000

H Walk 2+b Gamma    0.967  0.951  0.919  0.974  0.909  0.965         0.992  0.967  0.974
              p     0.000  0.000  0.000  0.000  0.000  0.000         0.000  0.000  0.000

I Walk 1bl Gamma    0.958  0.957  0.912  0.983  0.922  0.936  0.992         0.984  0.970
              p     0.000  0.000  0.000  0.000  0.000  0.000  0.000         0.000  0.000

J  Bathing Gamma    0.914  0.936  0.860  0.970  0.901  0.865  0.967  0.984         0.910
              p     0.000  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      J     score
-------------------------------------------------------------------------------------------
K      SEX Gamma   -0.366 -0.440 -0.155 -0.264 -0.131 -0.170 -0.185 -0.192 -0.033 -0.217
              p     0.000  0.000  0.014  0.004  0.037  0.021  0.043  0.059  0.385  0.000

L      AGE Gamma   -0.608 -0.577 -0.545 -0.584 -0.546 -0.557 -0.594 -0.661 -0.588 -0.468
              p     0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000  0.000
```

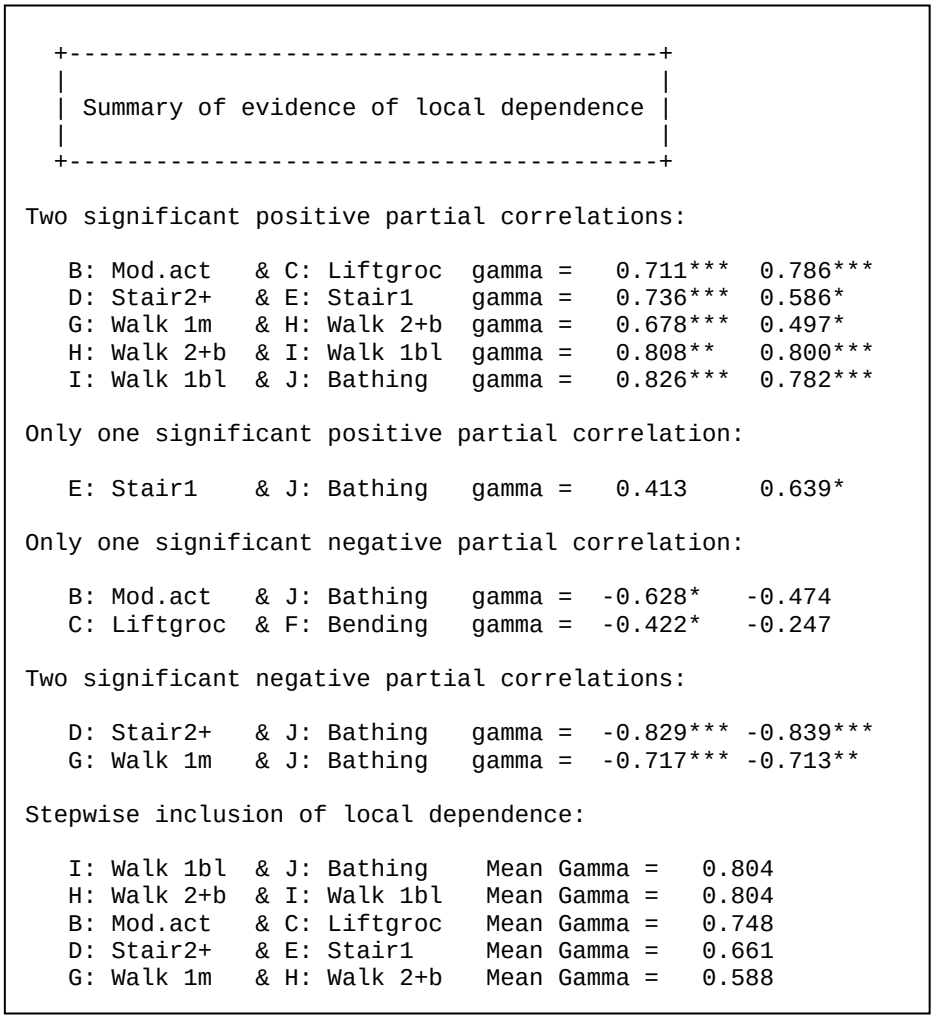
Figure 2.4.3 Gamma coefficients measuring marginal association among items, scores and exogenous variables. Restscores are scores without the item defining the row

There are few surprises in Figure 2.4.3. We expect positive association between items and between items and restscores and we expect the same type of association between an exogenous variable on one hand and the items and the total score on the other. The only surprise is the weak association between Item J and Sex (K).

```
  +----------------------------------------+
  |                                        |
  | 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      J
-------------------------------------------------------------------------------
B  Mod.act Gamma           0.711 -0.166 -0.216 -0.284 -0.194 -0.024 -0.392 -0.628
              p            0.000  0.357  0.350  0.047  0.242  0.932  0.227  0.010

C Liftgroc Gamma    0.786        -0.200 -0.136 -0.422 -0.371 -0.281 -0.519  0.203
              p     0.000         0.163  0.619  0.005  0.014  0.267  0.030  0.429

D  Stair2+ Gamma    0.086 -0.014        0.736 -0.097  0.252 -0.315 -0.455 -0.829
              p     0.286  0.938        0.000  0.381  0.102  0.224  0.136  0.000

E   Stair1 Gamma   -0.511 -0.361  0.586        0.102 -0.440 -0.093  0.531  0.413
              p     0.025  0.073  0.006        0.267  0.018  0.725  0.095  0.130

F  Bending Gamma   -0.015 -0.247 -0.095  0.460        0.239  0.098 -0.205  0.029
              p     0.905  0.095  0.500  0.023        0.115  0.654  0.417  0.952

G  Walk 1m Gamma    0.007 -0.344  0.177 -0.270  0.070        0.678  0.056 -0.717
              p     0.905  0.019  0.240  0.235  0.638        0.000  0.828  0.000

H Walk 2+b Gamma   -0.157 -0.418 -0.519 -0.138 -0.259  0.497        0.808  0.333
              p     0.619  0.077  0.024  0.556  0.223  0.006        0.001  0.215

I Walk 1bl Gamma   -0.509 -0.550 -0.582  0.560 -0.469 -0.061  0.800        0.826
              p     0.077  0.055  0.032  0.088  0.070  0.805  0.000        0.000

J  Bathing Gamma   -0.474  0.375 -0.839  0.639  0.106 -0.713  0.497  0.782
              p     0.086  0.137  0.000  0.008  0.636  0.001  0.050  0.000


        Test for DIF

                     B      C      D      E      F      G      H      I      J
-------------------------------------------------------------------------------
K       SEX Gamma   -0.503 -0.622  0.186  0.036  0.221  0.246 -0.030  0.393  0.295
              p     0.003  0.000  0.269  0.905  0.102  0.118  0.962  0.158  0.250

L       AGE Gamma   -0.113 -0.006  0.030  0.371 -0.078 -0.043  0.331  0.067  0.215
              p     0.292  0.974  0.786  0.036  0.524  0.759  0.103  0.746  0.278
```

Figure 2.4.4 Partial Gamma coefficient measuring conditional association among items given restscores without one of the items and between items and exogenous variables given the total score over all items. Restscores are scores without the item defining the row

The serious part of the item screening starts in Figure 2.4.4 where partial gamma coefficients are used to measure the conditional association among pairs of items given restscores without one of the items and the conditional association between items and exogenous variables given the total score on all items. The Rasch model expects all the partial gamma coefficients to be equal to zero. DIGRAM adjusts assessment of significance due to multiple testing but still concludes that there is strong evidence of both local dependence and DIF.



```
  +----------------------------------------+
  |                                        |
  | Summary of evidence of local dependence |
  |                                        |
  +----------------------------------------+

Two significant positive partial correlations:

   B: Mod.act   & C: Liftgroc  gamma =   0.711***  0.786***
   D: Stair2+   & E: Stair1    gamma =   0.736***  0.586*
   G: Walk 1m   & H: Walk 2+b  gamma =   0.678***  0.497*
   H: Walk 2+b  & I: Walk 1bl  gamma =   0.808**   0.800***
   I: Walk 1bl  & J: Bathing   gamma =   0.826***  0.782***

Only one significant positive partial correlation:

   E: Stair1    & J: Bathing   gamma =   0.413     0.639*

Only one significant negative partial correlation:

   B: Mod.act   & J: Bathing   gamma =  -0.628*   -0.474
   C: Liftgroc  & F: Bending   gamma =  -0.422*   -0.247

Two significant negative partial correlations:

   D: Stair2+   & J: Bathing   gamma =  -0.829*** -0.839***
   G: Walk 1m   & J: Bathing   gamma =  -0.717*** -0.713**

Stepwise inclusion of local dependence:

   I: Walk 1bl  & J: Bathing    Mean Gamma =   0.804
   H: Walk 2+b  & I: Walk 1bl   Mean Gamma =   0.804
   B: Mod.act   & C: Liftgroc   Mean Gamma =   0.748
   D: Stair2+   & E: Stair1     Mean Gamma =   0.661
   G: Walk 1m   & H: Walk 2+b   Mean Gamma =   0.588
```

Figure 2.4.5 Analysis of evidence of local dependence

The problem with the evidence collected in Figure 2.4.4 is that it probably draws a too dark picture of the situation because a lot of the evidence is spurious in the sense that it is caused by departures from the model that are different from those that the evidence suggests. For this reason, DIGRAM attempt to distinguish between genuine and spurious evidence of local dependence (Figure 2.4.5) and DIF (Figures 2.46 & 2.4.7).

The end result on local dependence is written at the bottom of Figure 2.4.5. Out of 10 pairs of items with evidence of local dependence, DIGRAM selects five pairs that look genuine and dismisses five other pairs where the evidence looks spurious. The five pairs are selected in a stepwise manner. In each step DIGRAM selects the pair with the strongest positive association as genuine and dismisses all significant test results that could be caused if the selected pair really was locally dependent.

```
Evidence of several biased items relative to SEX(K)

-------------------------------------------------------------------------------------
                         p-values                p-values (2-sided)
Hypothesis       X²  df asymp exact              Gamma asymp exact               nsim
-------------------------------------------------------------------------------------
 1:K&B|C#      22.2  16 0.136 0.247 (0.214-0.284) -0.40 0.041 0.056 (0.040-0.078) 1000
 2:K&C|B#      40.8  22 0.009 0.007 (0.003-0.018) -0.53 0.002 0.003 (0.001-0.012) 1000 x  -
-------------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.025 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
-------------------------------------------------------------------------------------

Excluded: B - Mod.act

-------------------------------------------------------------------------------------
                         p-values                p-values (2-sided)
Hypothesis       X²  df asymp exact              Gamma asymp exact               nsim
-------------------------------------------------------------------------------------
 1:K&C|#       41.9  18 0.001 0.001 (0.000-0.008) -0.62 0.000 0.000 (0.000-0.007) 1000 xx --
-------------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.050 for FDR = 0.05
                          and p <  0.010 for FDR = 0.01
Significance of
X²        xx : FDR = 0.01    x : FDR = 0.05
Gamma  ++/-- : FDR = 0.01  +/- : FDR = 0.05
-------------------------------------------------------------------------------------


Remaining biased items(s):  C - Liftgroc
```

Figure 2.4.6 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. It turns out, however, that K and B are conditionally independent given the score *and* B for which reason the original evidence of DIF is regarded as spurious.

```
-------------------------------------------------------------------------------
                        p-values                p-values (2-sided)
Hypothesis       X²  df asymp exact             Gamma asymp exact              nsim
-------------------------------------------------------------------------------
 1:#&K         28.2  18 0.059 0.042 (0.028-0.062) -0.22 0.000 0.000 (0.000-0.007) 1000 x  -
-
-------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.050 for FDR = 0.05
                         and p <  0.005 for FDR = 0.01
Significance of
X²        xx : FDR = 0.01    x : FDR = 0.05
Gamma  ++/-- : FDR = 0.01  +/- : FDR = 0.05
-------------------------------------------------------------------------------

K has a marginal effect on the score

-------------------------------------------------------------------------------
                        p-values                p-values (2-sided)
Hypothesis       X²  df asymp exact             Gamma asymp exact              nsim
-------------------------------------------------------------------------------
 1:#&L        206.2  54 0.000 0.000 (0.000-0.007) -0.47 0.000 0.000 (0.000-0.007) 1000 xx -
-
-------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.050 for FDR = 0.05
                         and p <  0.010 for FDR = 0.01
Significance of
X²        xx : FDR = 0.01    x : FDR = 0.05
Gamma  ++/-- : FDR = 0.01  +/- : FDR = 0.05
-------------------------------------------------------------------------------

L has a marginal effect on the score

2 variables with an effect on the score: K L

-------------------------------------------------------------------------------
                        p-values                p-values (2-sided)
Hypothesis       X²  df asymp exact             Gamma asymp exact              nsim
-------------------------------------------------------------------------------
 1:#&K|L       61.6  57 0.315 0.267 (0.233-0.304) -0.22 0.000 0.000 (0.000-0.007) 1000    -
```

Comments:

This is a two-step procedure. 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 2.4.7 Analysis of associations between the score and the exogenous variables

2.4.2 Model search

Item screening disclosed strong evidence of local dependence and DIF and therefore strong evidence against the Rasch model and instead proposed the GLLRM shown in Figure 2.4.2. The purpose of this model is only to serve as the starting point for a more careful search for a GLLRM and it is not expected to be the final model. It *is*, of course, expected to be close to an adequate model (and often is very close), but do not complain if the screening missed some important relationships between item and exogenous variables.

In one sense, the analysis following the creation of the screen model is a standard loglinear modeling exercise except for two reasons.

1. It is technically more complicated because we are doing a high-dimensional loglinear analysis conditional given the total score on all items.
2. The analysis in DIGRAM is never automatic. It is up to you to decide what happens after each step. To help you make this decision, DIGRAM provides information on significance of test statistics, estimates of parameters and the measures of partial association obtained during screening (when it is available). But do not forget subject matter considerations and contents 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.

The empirical evidence for or against interaction terms comes from Kelderman's conditional likelihood ratio test of local dependence and or DIF. These can be obtained in four ways that always refer to the "new" model (the model that you are trying to develop) by selection of options in the GRM dialog box Figure 2.1.5.

1. Select "reduce model" to test the terms of 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. A '*' in an interaction term is a wildcard referring to all variables. "B*" in the "Test model terms" field means that all interaction terms involving B will be tested. If X is an exogenous variable then "X*" tests DIF against all items and also includes Andersen's global CLR test of DIF.

You have already tried the first possibility in Section 2.3.3, where these tests were used for confirmatory tests of local dependence and DIF, but here you use them to search for a simpler model. Figure 2.4.7 shows the result of the selecting the second and third possibility, where the output includes the partial gamma coefficients calculated during item screening as shown in Figure 2.4.7 when available.

```
Check assumptions of local independence

B & D:   lr =   11.60  df =   4  p = 0.0206  -0.17  0.09
B & E:   lr =    9.42  df =   4  p = 0.0514
B & F:   lr =    3.86  df =   4  p = 0.4256
B & G:   lr =    5.74  df =   4  p = 0.2197
B & H:   lr =   10.75  df =   4  p = 0.0295  -0.02 -0.16
B & I:   lr =    7.35  df =   4  p = 0.1185
B & J:   lr =    6.34  df =   4  p = 0.1751
C & D:   lr =    9.06  df =   4  p = 0.0597
C & E:   lr =   13.43  df =   4  p = 0.0094  -0.14 -0.36
C & F:   lr =    4.96  df =   4  p = 0.2912

       Output has been deleted here

H & J:   lr =    9.17  df =   4  p = 0.0571

Check assumptions of no DIF
Gamma coefficients will be reported for significant test results


B & K:   lr =    5.22  df =   2  p = 0.0737
D & K:   lr =    2.58  df =   2  p = 0.2746
E & K:   lr =    1.74  df =   2  p = 0.4181

       Output has been deleted here

G & L:   lr =   15.54  df =   6  p = 0.0164 gamma =  -0.04
H & L:   lr =    4.61  df =   6  p = 0.5953
I & L:   lr =   28.46  df =   6  p = 0.0001 gamma =   0.07
J & L:   lr =   30.81  df =   6  p = 0.0000 gamma =   0.22

Benjamini & Hochberg rejects at 0.01042

Suggested additions to the model:

Positive LD:  EI EJ
Negative LD:  CE DI DJ EH GI GJ
DIF:          IL JL
```

Figure 2.4.7 Checking local independence and missing DIF

Figure 2.4.7 summarizes the results at the end of the list with test statistics. The analysis provided evidence of both local dependence and DIF. At the end of the day, adding IL and JL to the model and retesting local dependence and DIF showed that this was all that needed to be done.

And that is all. Stepwise model selection may be an unusual technique if you are used to IRT and Rasch analyses in general, but has little experience with multivariate statistics. Learning to use and appreciate these methods may therefore take longer than going through the first three DIGRAM tours. Fortunately, things simplify after model search where we continue to do what we are used to do in exactly the same was as before we introduced model search: we calculate over-all fit statistics, item fit statistics, person estimates, and assess test information and the targeting of the items to the study population 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.

Finally, partly to convince you that the procedures described in this section actually work, but also to warn you of some technical problems that you may run into, Figure 2.4.8 shows the over-all fit statistics for the GLLRM with IL JL (DIF of walking one block and bathing relative to Age) included.

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

             CLR   df   p        delta
-------------------------------------
scoregroups   43.5  49 0.694     0.045
K:      SEX   49.6  45 0.295     0.043
L:      AGE  101.7  99 0.406
```

Figure 2.4.8 Over-all fit tests for the (BC, DE, GH, HI, IJ, CK, IL, JL) model

The problem referred to is indicated by the Delta values of Figure 2.4.8. The problem is that the model structure may be so complicated that the iterative procedure DIGRAM uses to find the estimates of the parameters runs into numerical problems. DIGRAM claims that estimation did not converge. That is not exactly true in this case, but the convergence was so slow that DIGRAM lost its patience and stopped before it was happy with the solution to the estimation equations.

If this happens during the initial estimation of parameters DIGRAM will tell you so (after 1000 iterative steps) and ask (Figure 2.4.9) whether you want to continue.

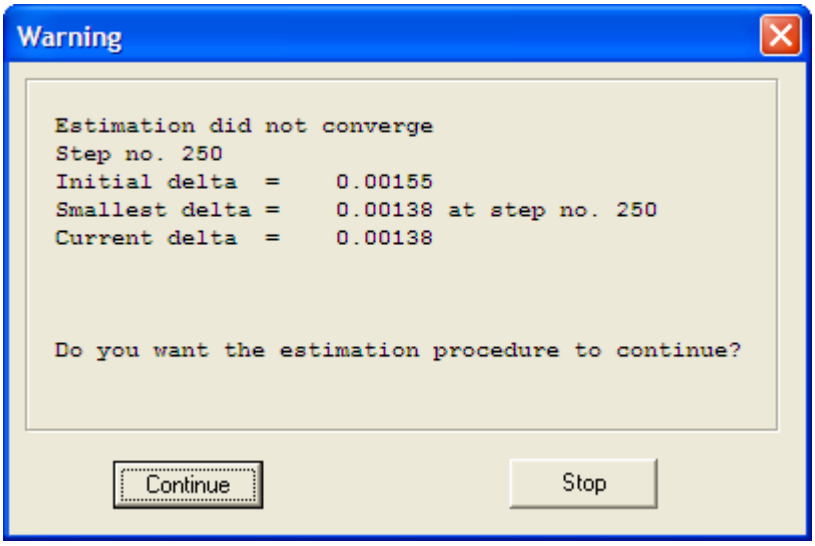


Figure 2.4.9 Warning when the iterative procedure is slow[17]

It is, in other word, up to you to decide how long time you want to use on the estimation. If you stop before DIGRAM is satisfied with the solution, DIGRAM will tell you so and ask you whether you want continue finishing with the other things you have asked it to do (Figure 2.4.10)[18].

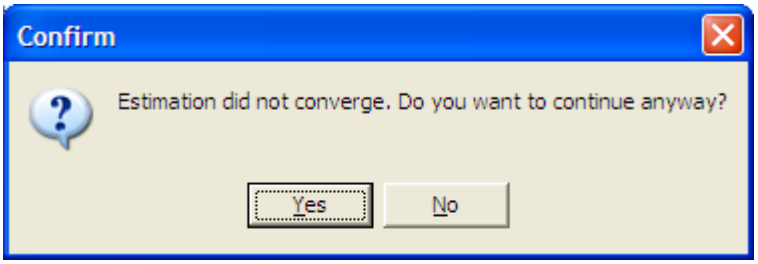


Figure 2.4.10 Warning after unsuccessful estimation

The Delta referred to in Figures 2.4.9 and 2.4.10 is the maximal difference between the fitted and observed margins of the model. The difference should ideally be equal to zero, but the default option implemented in DIGRAM is to stop when Delta is less than 0.0001. The default is arbitrary and you can change it if you want to (there is an editable field in the GRM dialog, where you can do that). The default is also to use 5000 steps and ask you every 1000 steps whether you want to

[17] Figure 2.4.9 was produced by an earlier version of DIGRAM where iteration was interrupted after 250 iterative steps. In the current version, this only happens after 1000 steps.

[18] For some reason, this problem occurs more frequently when the score distribution is skewed with a pronounced ceiling effect. If this is the case, it is therefore often helpful to stop the analysis and flip the items so that the score has a floor effect before you continue.

continue (you can take another 1000 steps if you are not happy after the first steps)[19]. In the very large majority of cases estimation stops much earlier than that, but the plan is also to make these limitations editable.

In addition to giving these warning when things do no work as fast as expected, DIGRAM also reports the value of the delta in one of the status fields at the bottom of DIGRAM's main form so that you can see whether the estimation procedure has serious problems. This is not the case here, and since inadequate estimates usually result in extremely large and highly significant test statistics, there is no reason to be concerned about the results shown in Figure 2.4.8.

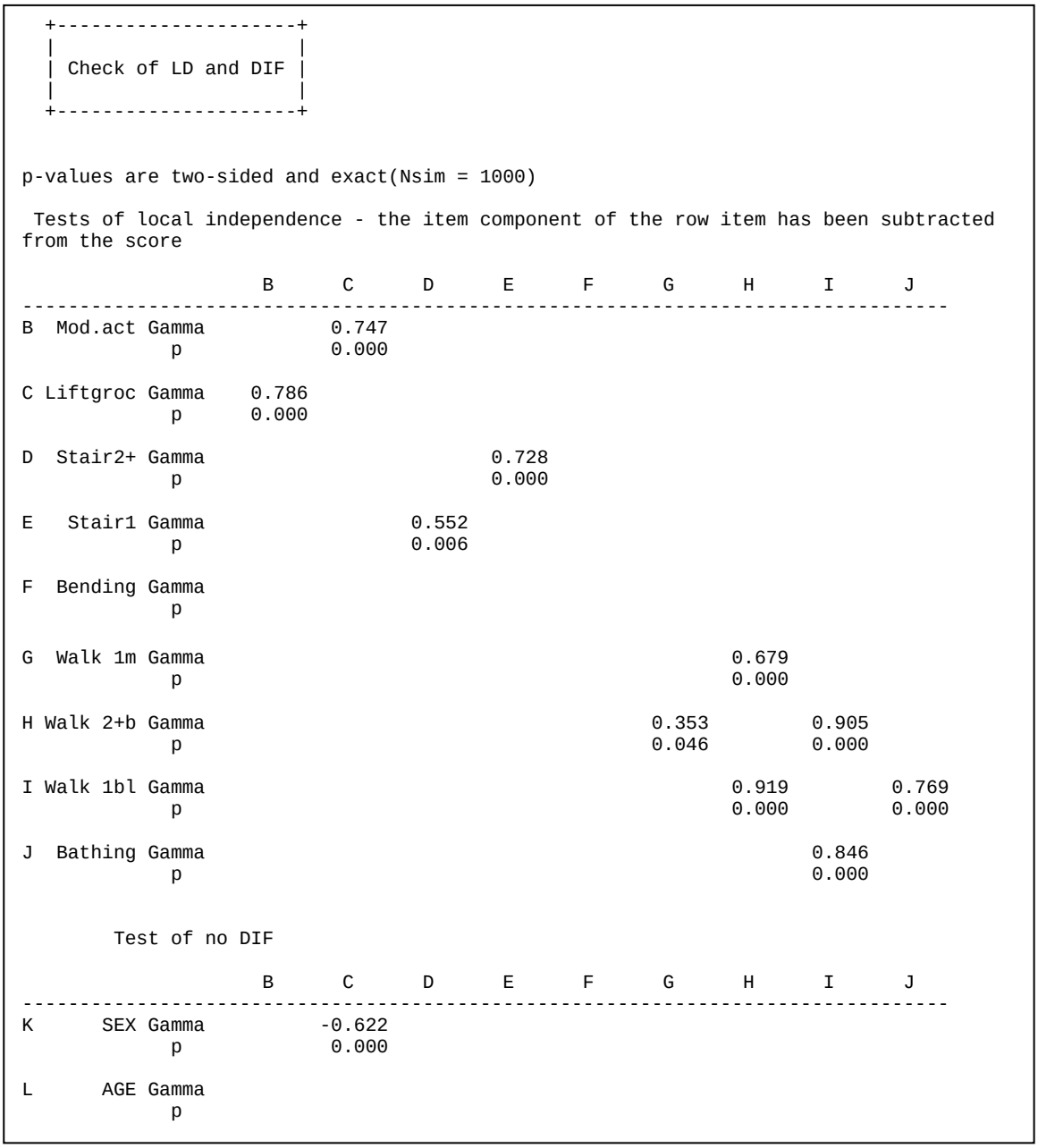
2.4.3 Checking the global Markov properties[20].

The first part of the item screening where DIGRAM tests for local dependence and DIF (Figure 2.4.3) consists of test of the so-called global Markov properties of the Rasch models. The current model GLLRM also has global Markov properties that can be tested. To do this you have to return to DIGRAM's main form. Remember to save the new GLLRM if you want DIGRAM to regard it is the current model before you exit the GRM dialog.

To test the global Markov properties of the current model, you must invoke a "**CHECK I**" command. Output from this procedure is very extensive, because DIGRAM produces details on both the Markov properties and the tests of the hypotheses and provides a summary of the analyses at the end.

In most cases, you only need to look at the summary. This is partitioned into two sets, one with test results of the local dependence in the model and another with test results relating to local independence and no DIF. These results are shown in Figures 2.4.11 and 2.4.12 for the model defined by item screening.

[19] During calculation of test statistics where DIGRAM has to estimate parameters in different groups or under different models, the default is to stop the iterative procedure after 250 steps. This limit will also be editable in the future.
[20] You should consult Kreiner & Christensen (2011a) before you try this to be sure that you understand what goes on.



```
  +--------------------+
  |                    |
  | Check of LD and DIF |
  |                    |
  +--------------------+


p-values are two-sided and exact(Nsim = 1000)

 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      J
-------------------------------------------------------------------------------
B  Mod.act Gamma           0.747
             p             0.000

C Liftgroc Gamma    0.786
             p      0.000

D  Stair2+ Gamma                         0.728
             p                           0.000

E   Stair1 Gamma                  0.552
             p                    0.006

F  Bending Gamma
             p

G  Walk 1m Gamma                                              0.679
             p                                                0.000

H Walk 2+b Gamma                                       0.353         0.905
             p                                         0.046         0.000

I Walk 1bl Gamma                                              0.919         0.769
             p                                                0.000         0.000

J  Bathing Gamma                                                     0.846
             p                                                       0.000


        Test of no DIF

                      B      C      D      E      F      G      H      I      J
-------------------------------------------------------------------------------
K      SEX Gamma          -0.622
             p             0.000

L      AGE Gamma
             p
```

Figure 2.4.11. Test of local dependence and DIF in the (BC, DE, GH, HI, IJ, CK) model

Comments: Figure 2.4.11 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.

```
  +------------------------+
  |                        |
  | Check of LI and no DIF |
  |                        |
  +------------------------+


p-values are two-sided and exact(Nsim = 1000)

 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      J
-------------------------------------------------------------------------------
B  Mod.act Gamma                  0.171  0.101 -0.163  0.037  0.276 -0.133 -0.403
             p                    0.450  0.633  0.251  0.667  0.336  0.667  0.189

C Liftgroc Gamma                  0.197  0.213 -0.270 -0.072 -0.008 -0.427  0.288
             p                    0.186  0.351  0.104  0.720  0.982  0.122  0.462

D  Stair2+ Gamma    0.112                       -0.031  0.258        -0.586 -0.622
             p      0.560                        0.667  0.076         0.026  0.012

E   Stair1 Gamma   -0.291 -0.315                0.270 -0.280  0.143  0.462  0.383
             p      0.393  0.152                0.128  0.196  0.644  0.094  0.130

F  Bending Gamma          -0.236 -0.122  0.500         0.196  0.101 -0.091
             p             0.183  0.359  0.016         0.146  0.857  0.824

G  Walk 1m Gamma    0.141 -0.408  0.132 -0.083                       no     no
             p      0.381  0.014  0.667  0.730                      test   test

H Walk 2+b Gamma    0.351 -0.194 -0.478  0.543 -0.072                       no
             p      0.098  0.310  0.012  0.002  0.684                      test

I Walk 1bl Gamma    0.232 -0.090 -0.642  0.782 -0.206    no
             p      0.323  0.571  0.010  0.000  0.326   test

J  Bathing Gamma   -0.436  0.333 -0.763  0.739  0.074    no     no
             p      0.100  0.126  0.000  0.000  0.789   test   test


        Test of no DIF

                     B      C      D      E      F      G      H      I      J
-------------------------------------------------------------------------------
K      SEX Gamma   -0.402         0.091 -0.190  0.059  0.029 -0.397  0.123  0.289
             p      0.041         0.577  0.416  0.717  0.875  0.154  0.639  0.348

L      AGE Gamma   -0.082 -0.041  0.027  0.419 -0.082 -0.007  0.415  0.000  0.212
             p      0.564  0.735  0.805  0.020  0.444  0.966  0.101  1.000  0.365
```

Figure 2.4.12. Test of local dependence and DIF in the (BC, DE, GH, HI, IJ, CK) model

Comments: Figure 2.4.12 tests the global Markov properties of all the claims of local independence and no DIF in the model defined by item screening. There is some evidence against the model (e.g. evidence of negative local dependence between items D and J) and some hypotheses that cannot be tested because of the complicated model structure, but the overall impression is that the model looks pretty good.

In addition to the tests of the global Markov properties, this procedure also provides tests of the assumption that local dependence and DIF is uniform. These results are not summarized, so you

need to look at the test results for the separate hypotheses to see whether this assumption is satisfied. Figure 2.4.13 shows the analysis of the DIF of item C relative to K.

```
         +---------------------------------------+
         |                                       |
         | Check of DIF for C-Liftgroc relative to |
         |                                       |
         | DIF source(s): K-SEX                  |
         |                                       |
         | No other biased items                 |
         |                                       |
         +---------------------------------------+
-------------------------------------------------------------------------------------
                        p-values                p-values (1-sided)
Hypothesis       X²  df asymp exact             Gamma asymp exact               nsim
-------------------------------------------------------------------------------------
 1:C&K|#       41.9  18 0.001 0.005 (0.002-0.015) -0.62 0.000 0.000 (0.000-0.007) 1000 xx --
-------------------------------------------------------------------------------------
Benjamini Hochberg rejects if p <  0.050 for FDR = 0.05
                          and p <  0.010 for FDR = 0.01
Significance of
X²        xx : FDR = 0.01    x : FDR = 0.05
Gamma  ++/-- : FDR = 0.01  +/- : FDR = 0.05
-------------------------------------------------------------------------------------

Fit of 2-factor model:    Dev =    25.4 df =   16 p =  0.0634
  +-----------------------------------------------+
  |                                               |
  | Analysis of fitted Gamma coefficients for CK |
  |                                               |
  +-----------------------------------------------+

Fitted partial Gamma:

Observed =  -0.622
Expected =  -0.624
    s.e. =   0.123
       Z =   0.015
       p =   0.988

Local values

Strata defined by #
                                                        cumulated
                                                     down          up
 # observed expected   s.e.      z      p       z      p      z      p
-------------------------------------------   --------------------------
 2   -1.000   -0.785  0.620  -0.346  0.729   -0.346  0.729  2.193  0.028
 3   -0.333   -0.785  0.399   1.132  0.258    0.555  0.579  2.372  0.018
 4   -1.000   -0.785  0.502  -0.428  0.669    0.206  0.837  2.143  0.032
 6   -0.778   -0.785  0.319   0.024  0.981    0.190  0.849  2.367  0.018
 8    0.333   -0.785  0.451   2.481  0.013    1.280  0.201  2.475  0.013
 9   -1.000   -0.785  0.525  -0.409  0.683    1.002  0.316  1.782  0.075
10    1.000   -0.600  0.713   2.243  0.025    1.775  0.076  2.034  0.042
11    1.000   -0.600  0.702   2.278  0.023    2.466  0.014  1.327  0.185
12    0.000   -0.683  0.397   1.719  0.086    2.898  0.004  0.503  0.615
13   -0.875   -0.714  0.229  -0.700  0.484    2.528  0.011 -0.217  0.828
14   -0.447   -0.625  0.222   0.800  0.423    2.651  0.008  0.107  0.915
15   -0.742   -0.600  0.267  -0.532  0.595    2.385  0.017 -0.339  0.735
16   -1.000   -0.619  0.207  -1.839  0.066    1.781  0.075 -0.038  0.969
17   -0.018   -0.600  0.326   1.784  0.074    2.193  0.028  1.784  0.074

Chi square =   29.1  df = 13  p =  0.0064
14 strata. U-test compares no 1-5 and 10-14   u =  1.059
```

Figure 2.4.13. Test of DIF of item C relative to K

Comments on Figure 2.4.13. The test of DIF of C relative to K require a test of conditional independence of C and K given the total score S. This test is presented first.

To check that DIF is uniform we have to test that there is no third order interaction in the three-way table with C, K and S. The loglinear test of this hypothesis is presented next. The hypothesis is accepted.

Since there is reason to be concerned about the power of loglinear tests of higher order interactions, DIGRAM also calculates the expected gamma coefficient in each strata under the two-factor model and compares them to the observed coefficients, in order to se whether there is a trend in the differences between then observed and expected coefficients. There is no apparent trend, but a χ^2 test summarizing the differences disagrees with the hypothesis that there is no three-way interaction and therefore casts some doubt on the adequacy of the GLLRM defined by screening.

2.4.4 Analysis of person fit

The analysis of person fits attempt to identify all persons with improbable responses according to the conditional distribution of item responses given the total score on all items. The test of person fit has to be initiated from DIGRAM's main form and therefore assumes that you want to test the person fit under the current model and that item parameter estimates under this model is available.

Use the "**PERSONFIT**" command if you are sure that this is what you want and if you are confident that the estimates of the parameters of this model are available.

The analysis is a two-step procedure. In the first step DIGRAM calculates the conditional probabilities of all response patterns for all combinations of outcomes on DIF sources and saves the result on a text file called "Responsepatterns.txt". We won't show it 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 a set of tables with the most probable response patterns for each combination of outcomes on DIF sources. Figure 2.4.14 shows one example.

```
K:       SEX = Female
L:       AGE = 60 - 69

Max probability patterns
 0 1.0000000   pattern =  0 0 0 0 0 0 0 0 0
 1 0.3319113   pattern =  0 0 1 0 0 0 0 0 0
 2 0.1862790   pattern =  0 0 1 0 1 0 0 0 0
 3 0.1801821   pattern =  1 1 1 0 0 0 0 0 0
 4 0.2244477   pattern =  1 1 1 0 1 0 0 0 0
 5 0.1648985   pattern =  1 1 1 0 1 1 0 0 0
 6 0.1055161   pattern =  1 1 1 1 1 1 0 0 0
 7 0.0738719   pattern =  1 1 2 1 1 1 0 0 0
 8 0.0458253   pattern =  1 1 1 0 1 1 1 1 1
 9 0.0484315   pattern =  1 1 1 1 1 1 1 1 1
10 0.0548641   pattern =  1 1 1 1 1 2 1 1 1
11 0.0726418   pattern =  1 1 2 1 1 2 1 1 1
12 0.0545729   pattern =  1 1 2 1 2 2 1 1 1
13 0.1080148   pattern =  2 2 2 1 1 2 1 1 1
14 0.1328614   pattern =  2 2 2 1 2 2 1 1 1
15 0.1619054   pattern =  2 2 2 1 2 2 1 1 2
16 0.1850167   pattern =  2 2 2 2 2 2 1 1 2
17 0.3815043   pattern =  2 2 2 1 2 2 2 2 2
18 1.0000000   pattern =  2 2 2 2 2 2 2 2 2
```

Figure 2.4.14 Overview of the most probable response patterns for females, age 60-69

The results of the pattern analysis are less than important here but we return to it below to explain what it can be used for.

In the second step, DIGRAM, calculates the conditional probability for all persons and assess the significance of the response pattern using the exact test proposed by Martin-Löf (1977)[21], and prints all response patterns that are significant at a 5 % level. The results can be seen in Figure 2.4.15

[21] 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

```
+---------------+
|               |
| Person review |
|               |
+---------------+

K  L B C D E F G H I J score  Prob  count     p

2  2 0 0 2 0 0 0 0 0 0   2   0.004      1    0.042
2  4 1 1 1 1 0 0 1 2 2   9   0.000      1    0.015
2  1 1 0 0 0 0 2 1 0 0   4   0.001      1    0.043
2  2 0 2 0 0 0 0 0 0 0   2   0.003      1    0.027
2  3 0 1 0 0 0 0 0 0 1   2   0.004      1    0.040
2  1 0 0 1 1 0 2 1 0 1   6   0.000      1    0.039
1  3 0 0 1 0 0 1 0 1 0   3   0.002      1    0.043
1  3 2 0 1 1 0 0 0 0 0   4   0.001      1    0.046
1  4 0 0 0 1 0 0 0 1 1   3   0.000      1    0.001
1  2 0 0 0 0 1 0 0 0 1   2   0.005      1    0.040
2  4 2 2 0 0 0 0 2 0 0   6   0.000      1    0.012
1  1 1 2 1 2 1 1 0 0 2  10   0.000      1    0.006
2  2 1 2 0 1 2 0 2 2 2  12   0.000      1    0.010
1  3 1 1 1 1 2 0 1 2 2  11   0.000      1    0.023
1  4 1 1 1 1 0 0 1 2 2   9   0.000      1    0.017
2  4 1 1 0 2 0 0 1 1 2   8   0.000      1    0.005
2  4 0 1 1 1 1 0 0 1 2   7   0.000      1    0.025
2  3 0 1 0 0 0 0 0 0 1   2   0.004      1    0.040
1  4 0 0 0 1 0 0 1 1 2   5   0.000      1    0.001
```

Figure 2.4.15 Overview of persons with improbable responses

The analysis found 19 persons with improbable response patterns out of a total of 319 persons with non-extreme responses. This is close to what one would expect under the model so from that point of view there is nothing to be concerned about here. To be extra careful, one should look at the association between misfitting response patterns on one hand and the total score relative or exogenous variables on the other. The current version of DIGRAM produces the table shown in Figure 2.4.16 containing the frequencies of improbable response patterns in different score groups.

The frequencies of unexpected patterns are larger among persons with high scores, which could suggest that physically impaired persons have aberrant responses compared to persons with a normal physical functioning.

score	n	misfit	%
1	87	0	
2	63	5	7.9
3	31	2	6.5
4	34	2	5.9
5	19	1	5.3
6	16	2	12.5
7	9	1	11.1
8	7	1	14.3
9	14	2	14.3
10	5	1	20.0
11	6	1	16.7
12	8	1	12.5
13	6	0	
14	4	0	
15	5	0	
16	2	0	
17	3	0	
Total	319	19	6.0

Figure 2.4.16 Overview of persons with improbable responses

A closer look at why the response patterns in Figure 2.4.15 are improbable suggests, however, that the reason is typing error because the improbable response patterns are illogical. Consider, for instance, the last pattern in 2.4.15: “0 0 0 1 0 0 1 1 2” containing responses of an elderly male with a total score equal to 5. The most probable response pattern of such a person is “1 1 1 0 1 1 0 0 0” corresponding to a person with no limitations walking one flight of stairs, walking several blocks and bathing or dressing. According to the last response pattern in Figure 2.4.15, the person experienced no limitations walking more than a mile, but some limitations walking one and/or several blocks which obviously do not make much sense. Also, considering the relative few problems this person has, it is also very surprising that he should experience lot of limitations *due to his health* while bathing or dressing himself. The original questionnaires with the responses are, unfortunately not available, so we cannot do what would be the natural thing to do, namely go back to see whether data had been correctly typed.

2.4.5 All the other options.

Look again at the GRM dialog form Figure 2.3.2. There are, as you can see, several options that we have not looked at yet, and a few that are not available yet, but included to show what we intend to add to DIGRAM in the future.. Taken from the top they are

1) Analysis of local homogeneity and DIF
2) Analysis by the rating scale model of Andrich (1977)
3) Tables with the sufficient margins of the models.
4) Tables with estimates of the parameters of all the models considered during the analysis.
5) Export of data for latent regression analysis based on the current model.
6) Inclusion of response patterns during item parameter estimation and analysis of model fit.
7) Extended output during person parameter estimation and calculation of item fit statistics
8) Creation of text files with person parameter estimates that can be merged with the original data file.
9) Extra output in case of non-convergence during item parameter estimation.

All these options will be described in the next subsections. The description will in some cases be brief because these options are rarely used and probably most useful to the programmer, but there are a few of the options that deserve a more comprehensive treatment.

2.4.5.1 Analysis of local homogeneity and DIF.

The global conditional likelihood ratio (CLR) tests of homogeneity and no DIF compares item parameter estimates in different groups to the item parameters for the complete samples and conclude that item parameters are not the same in all groups when the CLR tests are significant. When this happens, the next question to ask is whether item parameters are different in all groups or whether there is local homogeneity[22] and/or local no DIF because there are some groups where item parameters are the same.

DIGRAM addresses this issue by pairwise comparison of groups followed by stepwise collapse of groups if there is no evidence against the hypotheses that item parameters are the same in both groups. The procedure is a fully automatic p-value based procedure. A similar non-automatic procedure has been implemented for analysis of local homogeneity of dichotomous items (see Section 3.3 for an example). If the variable defining the groups is ordinal (which, of course, is the case for the score groups) DIGRAM only compares and collapses adjacent categories.

[22] DIGRAM uses score groups rather than raw scores for analysis of local homogeneity.

We illustrate the procedure with analysis of local DIF. Sex is binary for which reason DIGRAM skips this variable. Age, on the other hand, has four ordinal variables. According to the moel, Age is a source of DIF for two items. The DIF interaction terms relating to Age are consequently disregarded during the analysis, but apart from this, the model is unchanged. Figures 2.4.7 and 2.4.8 show some of the results.

```
   category count
-----------------

 1  16 - 29    47
 2  30 - 44    60
 3  45 - 59    97
 4  60 - 69   115

Acceptance of hypotheses subject to FDR <= 0.05
at each step.

p-values for pairwise comparisons

      1     2     3     4
---------------------------
  1  .    0.068   **    **
  2 0.068  .    0.000   **
  3   **  0.000  .    0.108
  4   **    **  0.108  .

3 p-values. The Benjamini Hochnerg procedure
accepts if p greater than  0.0167

Equivalence rejected for the following groups: 2 and  3 p =  0.000

Summary of coherence adjusted Benjamini Hochberg analysis

Collapsed groups:

     1+2
     3+4
```

Figure 2.4.17 Initialization of analysis of local no DIF

Comment on Figure 2.4.7:

The initialization includes a list of the groups to be examined including the number of persons in each group and a table of CLR based p-values comparing adjacent groups. No evidence is found against the hypotheses that the item parameters are the same in the first two groups and in the last two groups. Item parameters in groups 2 and 3 are, however, significantly different.

```
  +-----------------------+
  |                       |
  | Stepwise model search |
  |                       |
  +-----------------------+

Number of groups   = 4
Number of p-values = 3

stepno.    CLR  df      p     collapsed groups
-------------------------------------------------------
     1    44.5  34  0.108     3 and 4
     2    42.3  36  0.068     1 and 2
     3    65.7  37  0.003     1+2 and 3+4

  +-----------------------+
  |                       |
  | MCA analysis finished |
  |                       |
  +-----------------------+


Analysis stops. Largest p is equal to 0.0025


Collapsed groups after stepwise model search:

     1+2
     3+4

Final p-values:

 1+2 and 3+4 p =   0.0025
```

Figure 2.4.18 Stepwise analysis of local no DIF

Comment on Figure 2.4.8:

The stepwise analysis first collapses groups 3 & 4. During the next step DIGRAM compares the item parameter estimates in group 2 to the item parameter estimates in the collapsed groups 3+4. The p-value is less than the p-value of the test comparing groups 1 & 2, for which reason groups 1 & 2 are collapsed in the second step. Finally, the CLR test comparing the collapsed groups 1+2 and 3+4 rejects that item parameters are the same in the two groups. The analysis therefore stops concluding that there is local “no DIF” in two pairs of groups.

Finally, the parameter estimates for the two groups are presented in the same way that item parameter estimates are always shown in DIGRAM. These estimates are not shown here, but take a look at them to see if you understand the result of the analysis.

2.4.5.2 The rating scale model

The rating scale model by Andrich (1978) is on the list, but is not yet available.

2.4.5.3 The sufficient margins

The sufficient margins of GLLRM include the total score over all items, the item margins, and all the two-way tables describing the marginal association between locally dependent items and the two-way tables with DIF item and DIF sources. During estimation of parameters, DIGRAM searches for parameters given which the observed sufficient margins are equal to the expected margins. If estimation fails, it may be useful to take a look at the observed margins to see whether there is anything unusual about them. You can ask for these margins for the model you are currently estimating and for all the models that are fitted during calculation of conditional likelihood ratio tests. They are not shown here because the output is extensive, but you should have no problems reading these tables.

In addition to the tables, DIGRAM produces list of cases with incomplete responses and distinguish between useless cases where person parameters cannot be estimated and useful cases where the person parameters can be estimated from the incomplete response patterns.

2.4.5.4 Item parameter estimates for all models

Conditional likelihood ratio tests require item parameter estimates for several models. DIGRAM reports a number of statistics that can help you understand what the differences mean (see Figure 2.1.9 for an example), but does not print the estimates of the separate models unless you specifically asks it to do so.

2.4.5.5 Latent regression

A good analysis of the association between the latent scale and the exogenous variables require a proper latent structure analysis. Such analyses are not implemented in DIGRAM, but SAS macros

exist (Christensen & Bjørner, 2003) that will do the job and DIGRAM can generate the files with the information for such analyses.

2.4.5.6 Analyses of incomplete response patterns

The analyses of the adequacy of the models only use data on persons with complete information on item responses and exogenous variables. In most cases where the frequencies of persons with incomplete data are small, such analyses are preferable to analyses with incomplete data, because there is a wider range of methods for the complete data and because conditional inference have no problems with sampling of persons with complete responses even though item responses may not be missing completely at random.

There are situations where the number of persons with missing responses is large and where item parameter estimates and fit statistics that take person with incomplete responses into account. For this reason we are working on methods for conditional inference that takes these persons into account. The methods are unwieldy, but within reach and they will appear in the program in the future. For now, we suggest that you use the RUMM program instead, if you are in a situation where you have to include the incomplete item responses in your analysis. Section 3.8 in Part II of these notes describes how to generate RUMM files from within DIGRAM if you want to do this.

2.4.5.7 Extended output during person parameter estimation

The defaults during person parameter estimation include ML and WML estimates, assessment of reliability and a DIF equation summary describing the effect of DIF on the total scores. There are, however several other options to select from. These options are presented in the dialog shown in Figure 2.4.19 that turns up, if you select “Extended output for person estimation and item fits” on the GRM dialog.

Gamma coefficients

Use the first options if you want to see the score parameters (called gamma coefficients) of the power series distribution of the total score. In the Rasch models for dichotomous items these score parameters are the symmetrical polynomials. In GLLRMs for polytomous items where items can be locally dependent and/or function differentially, these parameters are similar, but more complex functions of the item parameters and loglinear interaction parameters.

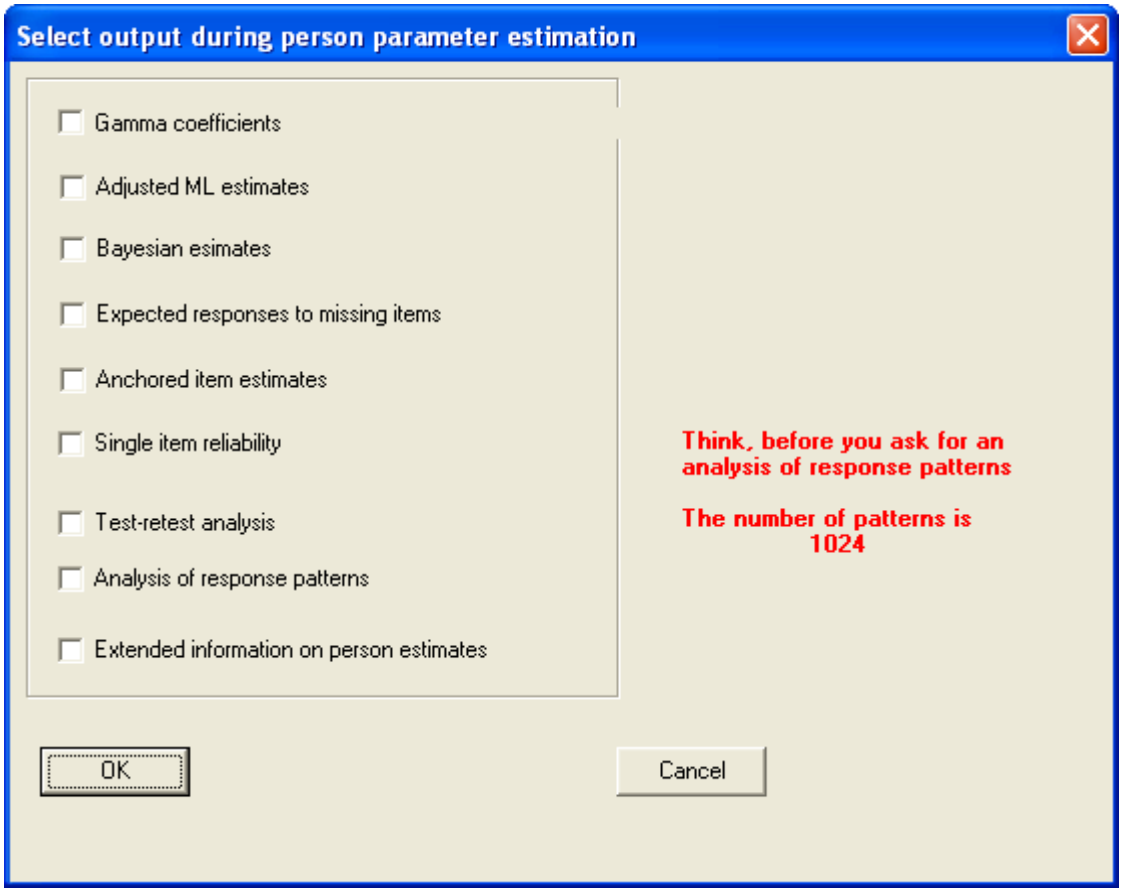


Figure 2.4.19 Output options during person parameter estimation

Adjusted and Bayesian estimates

In addition to the ML and WML estimates, DIGRAM also calculates so-called adjusted person ML estimates[23] (AML) and Bayesian and Bayesian mode estimates and present them in the same way as the ML and WML estimates in Figures 2.1.14 and 2.1.15 with information on bias and measurement error.

If you care about person parameter estimation we suggest that you try these options and compare the performance of your scale on your own data. According to theory, the WML estimates outperform the ML estimates in terms of bias and measurement error. Our experience with these estimates suggest that the adjusted ML estimates control bias even better than the WML and that Bayesian estimates do worse than the ML estimates. However, there is no reason for you to rely on

[23] The adjusted estimates are described in Section 4.2 of part II.

our experience. The point is that DIGRAM permits you to assess the performance of all these estimates on your own data.

Expected responses to missing items

Assume that the response to one of the items is missing whereas responses to the other items are available. For this situation DIGRAM calculates the expected item score on the missing item given the (rest)score on the other items if the missing item is locally independent of the other items. This facility will eventually be extended to cover any number of missing items whether or not they are locally independent.

Estimates related to a set of anchor items

Two things are produced when you select this option.

First, DIGRAM adjust the main effect item parameters so that the sums of the item thresholds for the anchor items are equal to zero.

Second, DIGRAM calculates and reports the expected score over the anchor items for a range of person parameter values.

Note: we are currently developing these facilities, but they are not highly prioritized. For this reason these results are only reported for the reference group of persons and not for all combinations of values of DIF sources.

Single item reliability

Test-retest reliability can be evaluated for single items by the same Monte Carlo methods that DIGRAM uses to calculate the test-retest of the total score. If you select this option DIGRAM will let you choose one item for an analysis of item reliability.

Test-retest analysis

Assume that you are able to let a person respond to the items in such a way that you can assume that there is no local dependence among repeated responses to items and that you want to test that the value of the latent variable is the same the first and the second response time. For this purpose

DIGRAM produces a number of tables summarizing the conditional probabilities of the two scores conditionally given the sum of the two scores so that you can evaluate whether there is a significan difference between the two scores.

Analysis of response patterns

Output here includes the same list of the most probable response patterns for each score in the groups defined by outcomes on the DIF sources that were produced during the analysis of person fit (Figure 2.2.14). Recall, that person parameters only provide relative assessment of the trait that the items are supposed to measure and that a low or high person parameter value can not be interpreted in absolute terms. To help interpretation, we have therefore included the list of the most probable response patterns as a help to understanding what the scores mean.

Text files with information on properties of person estimates

If you select “Extra file output during person estimation” on the GRM dialog form DIGRAM creates a text file “projectname-persons.txt” and “projectname-persons-comma.txt”with ML estimates DIF equated scores for different combinations of values of sources of DIF and the total score. The file also include information on the number of persons with the given values of the sources of DIF and the total score on all items. These files may be useful if you want to create macros or procedure for other programs converting observed scores on all items to person parameter estimates or DIF equated scores. Figure 2.4.20 shows the beginning and the end of such a file.

```
F score theta Escore count
1 0 -3.48864 0 8
1 1 -2.09627 1 9
1 2 -1.41227 2 6
1 3 -1.02213 3 12
1 4 -0.74359 4 15
1 5 -0.51553 5 10
1 6 -0.30980 6 9
     …….
2 10 0.56932 10.227 3
2 11 0.80812 11.173 1
2 12 1.09727 12.125 4
2 13 1.48839 13.085 0
2 14 2.14253 14.049 3
2 15 3.46450 15 1
```

Figure 2.4.20. The beginning and end of DHP1-person.txt defined by the BD,DE,CF model.

2.4.5.8 Files with person estimates

You must check “save person estimate” when you estimate the person parameters if you need text files with person estimates that can be merged with your original data file.

If you do this, DIGRAM saves the person estimates for persons with complete responses, but also calculates person estimates and DIF equated expected scores on all items for persons with missing responses. Figures 2.4.21 and 2.4.22 shows some of the output after person parameters have been saved.

```
Person no. 33

Item responses: 0 3 - 1 1   Score =  5
estimated th:   -0.4136   Equated score =     5.4874

DIFsources: 2 *
```

Figure 2.4.21 Estimates of person parameter and expected score on all items for a person without respons on the third item.

```
   +--------------------------------+
   |                                |
   | Report on incomplete responses |
   |                                |
   +--------------------------------+

26 persons with incomplete responses
17 persons with completely missing responses


   +----------------------------+
   |                            |
   | Person estimates were saved |
   |                            |
   +----------------------------+

Estimates on personestimates.txt
Person files: Person1.txt (dots) and Person2.txt (commas)
```

Figure 2.4.22 Report after saving of person parameter estimates

Each person in the data set is represented by one record in the files with person estimates. This record include information on item scores the values of the exogenous variables, the total score on the observed items, a status variable (coded 0 = complete, 1 = incomplete with estimate, 2 = incomplete without estimate, 3 = missing completely), the estimate of the person parameter and the estimated DIF equated score on all items.

2.4.5.9 Extended output in connection with item fit statistics

Ask for extended output during calculation of item fit statistics if you want to understand how these statistics work. If you do that, you will get information on outfit statistics including values of residuals for different combinations of item scores and total scores and you will se the item-restscore tables that we use to calculate the item-restscore correlations. In addition to these DIGRAM also calculates and reports Molenaar's U, which may be the preferred item fit statistic for some users.

2.4.5.10 Extended output during analysis of targeting

Assume that you want to eliminate one or more items in order to create a short form of your scale. The best way to do that is by selecting items (or rather item components of locally dependent items) in such a way that the measurement error provided by the short form is as small as possible. The extended output provided during analysis of targeting is intended to help you do this by providing analysis of targeting and available test information by the separate item components and by the subscores without an item component.

Figure 2.4.23 shows the analysis of targeting for the first item and for the restscore without the first item among males in age group 18-49. The results should be compared with the results of the complete scale in Figure 2.3.10. The average error of measurement is 0.514 for the complete score and 0.545 for the restscore without the first item. In this case it looks like the nothing much is lost by eliminating the first item.

```
  +--------------------------------------+
  |                                      |
  | Analysis of item component targeting |
  |                                      |
  +--------------------------------------+

Group: SEX = Male   AGE = 18-49

  14 persons.   Mean score = 6.29   sd = 3.43  Mean Theta = -0.281   sd = 0.646


  +----------------------------------+
  |                                  |
  | Targeting Item component no. 1: A |
  |                                  |
  +----------------------------------+

Highest component score =  3
Highest restscore       = 12

Location        =       0.453   Test information     =   0.853   SEM =   1.083
Test difficulty =       0.534   Test information     =   0.864   SEM =   1.076
Test target     =       0.697   Max test information =   0.873   SEM =   1.070

Mean test information =   0.624  sd =   0.174   Target index =   0.714

Mean SEM              =   1.314  sd =   0.232   Target index =   0.814

var(true score)/var(score)     =  0.214
test-retest correlation        =  0.217
test-true score correlation    =  0.459

Probability of correct person separation =  0.477
Probability of no person separation      =  0.323

================================================================================

Restscore without Item component no. 1

Test difficulty =      -0.114   Test information     =   4.351   SEM =   0.479
Test target     =      -0.229   Max test information =   4.385   SEM =   0.478

Mean test information =   3.597  sd =   0.846   Target index =   0.820

Mean SEM              =   0.545  sd =   0.100   Target index =   0.877

var(true score)/var(score)     =  0.608
test-retest correlation        =  0.607
test-true score correlation    =  0.764

Probability of correct person separation =  0.743
Probability of no person separation      =  0.088
```

Figure 2.4.23 Output options during person parameter estimation

2.4.5.11 Extra output in case of non-convergence

This output is probably only useful to the programmer who is still working on ways to avoid non-convergence.

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.

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). Loglinear 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-restscore 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 loglinear Rasch models. *Communications in Statistics*, 33: 1239-1276

Kreiner S, Christensen KB (2007) Validity and Objectivity in health-related Scales: Analysis by Graphical Loglinear 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 Loglinear 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.