DEPARTMENT OF LEARNING, INFORMATICS,
MANAGEMENT AND ETHICS
Karolinska Institutet, Stockholm, Sweden

A ‘3 step’ IRT strategy for evaluation of the use of sum scores in small studies with questionnaires using items with ordered response levels

(subtitle: ‘To make a lot out of nothing’, from the Swedish locution: ‘Att koka soppa på en spik’)

Ulf B. Brodin

Stockholm 2014

All previously published papers were reproduced with permission from the publisher.
Published by Karolinska Institutet. Printed by Universitetsservice US-AB

ISBN 978-91-7549-709-9

All models are wrong but some are useful.
[Everit & Dunn, 1991]

Abstract

This study is focused on a strategy for a basic evaluation of a questionnaire at an early state (i.e. only a limited sample from the target population is available).
Several questionnaires are constructed within the medical research to investigate phenomenon which cannot be explicitly measured (latent variables). In many cases, these questionnaires are applied on a limited sample (less than 100 subjects), without any proper evaluation of its basic characteristic.
This thesis presents a '3- step' strategy for evaluation of questionnaires, where only a limited sample is available for the investigation. Only questionnaires, containing items with a common set of answer alternatives, are considered. The answer alternatives are in terms of an ordered scale to measure the underlying, latent, variable. In all cases, the intention is that a sum score will be a relevant measure of the status of a subject. The higher the score, the more of the latent characteristic is attached to the subject. This implies that all items should represent one common dimension.
The investigation is carried out in three steps and is focused on whether the following criteria, at least approximately, correspond to the intentions which the researcher was supposed to have in mind, at the construction of the questionnaire:

1. All items cooperate.
2. All items work together towards a common dimension.
3. Non-coherent/irrelevant items, as well as non-coherent answer profiles can be identified.
4. The subjects can be reasonably ranked, based on the sum score, on the latent scale.
5. The sum score can be transformed, via a statistical model, to a relevant interval scaled measure.
6. The set of items covers reasonably the intended population.
7. The item set is sufficient for an estimate of a person's position on the latent scale.
8. Defined subgroups perceive the questionnaire in the same way.
9. A straight forward sum score is a sufficient measure (sufficient statistic). Otherwise, an elaborated model, with item specific discrimination, is considered.

Step 1: Non parametric statistical analysis according to 'The Mokken scale analysis' (1 - 4).
Step 2: A parametric model according to the Rasch approach (3, 5 - 8).
Step 3: Can further information be gained from an extended model? (9).

This strategy was applied on 5 studies, all carried out with the intention to use the sum score as a relevant measure of persons' status on the underlying latent scale. Each study involves more than one questionnaire. The '3 – step' strategy was applied on 13 questionnaires within these 5 studies, where the intention was to use a sum score.

Results: Already Step 1 reveals most of the basic weaknesses of the questionnaire, such as weak or contradictory items, deficient correlations between items and a violation of an invariant ordering of the item across low to high scoring persons. These findings are also confirmed in later steps, where further characteristics can be revealed. It also turned out that a well behaved questionnaire according to the 'Mokken requirements' is a prerequisite for a reasonable parametric approach. Furthermore, in most cases the number of items appears to be too small and that the item set has an insufficient coverage to reasonably estimate a person measure for all subjects belonging to the intended population. But, the '3 steps' can constitute a comprehensive message for a basic improvement of the questionnaire.

Sammanfattning på svenska

Denna studie inriktar sig på en strategi för att på ett tidigt stadium (d.v.s. där endast ett begränsat antal personer ur målpopulationen ligger till grund för utvärderingen) utvärdera basala egenskaper hos ett frågeformulär.

Åtskilliga frågeformulär (enkäter) konstrueras inom den medicinska forskningen för att studera fenomen som inte explicit kan mätas (latenta variabler). I de flesta fall tillämpas sedan dessa formulär på ett litet antal personer/patienter, utan att någon egentlig utvärdering av formulärets egenskaper har genomförts.

Detta arbete presenterar en '3 – stegs' strategi för utvärdering av enkäter med en för alla poster (items) gemensam uppsättning svarsalternativ. Svarsalternativen utgörs av en graderad skala med syfte att mäta den underliggande (latenta) variabel som studeras. I samtliga fall är avsikten att summascore skall vara ett relevant mått på en persons status. Ju högre summascore, ju högre grad av den latenta egenskapen anses personen ha ('lida av'). Detta innebär att samtliga items skall samverka och representera en gemensam dimension.

Utvärderingen inriktar sig på huruvida följande kriterier, åtminstone approximativt, motsvarar de intentioner som försöksledaren (forskaren) antages ha avsett vid konstruktionen av sin enkät:

1. Alla items samverkar.
2. Alla items samverkar till en gemensam dimension.
3. Icke relevanta/bidragande items, såväl som inkongruenta svarsprofiler kan identifieras.
4. Personerna kan rangordnas, med hjälp av summascore, på den latenta skalan.
5. Summascore kan transformeras, via en statistisk modell, till ett relevant mått på en intervallskala.
6. Uppsättningen items täcker den population som avses.
7. Uppsättningen items är tillräcklig för en rimlig skattning av en persons läge på den latenta skalan.
8. Definierade undergrupper uppfattar enkäten på samma sätt.
9. Summascore är ett tillräckligt mått. Alternativt beakta en utvecklad modell där items måste viktas för att erhålla ett relevant mått.

Steg 1: Icke parametrisk analys enligt 'The Mokken scale analysis' (1 – 4).
Steg 2: Parametrisk modell enligt Rasch (3, 5 – 8).
Steg 3: Kan ytterligare information hämtas från en utvidgad modell? (9).

Ovanstående strategi prövas på 5 studier, alla genomförda med avsikt är att använda summascore som ett relevant mått på de ingående personernas grad av det latenta problemet. Varje studie omfattar mer än en enkät. Denna '3 – stegs' strategi har prövats på sammanlagt 13 frågeformulär i dessa 5 studier, där avsikten är att använda summascore.

Resultat: Redan Steg 1 påvisar ett frågeformulärs basala svagheter, såsom svaga eller motsägande items, bristande korrelation mellan items och en icke invariant ordning mellan items från personer med låga score till de med höga. Dessa fynd verifieras också i Steg 2 och 3, där också ytterligare egenskaper kan klarläggas.

Det visar sig också att ett väl konstruerat frågeformulär enligt 'Mokken- analysen' är en förutsättning för en rimlig parametrisk ansats. Vidare, i de flesta fall visar sig antalet items vara otillräckligt och uppsättningen items ha bristande täckning för att skapa ett rimligt intervall-mått för varje person.

List of publications

I. Brodin Ulf, Fors Uno, Bolander Laksov Klara. The application of Item Response Theory on a teaching strategy profile questionnaire. BMC Biomedical Education 2010, 10:14.

II. Nissell M., Brodin, U., Christensen K., Rydelius P-A. The Imperforate Anus Psychosocial Questionnaire (IAPSQ): Its construction and psychometric properties. Child and Adolescent Psychiatry and Mental Health 2009, 3:15 (14 May 2009).

III. Brodin U., Fors UGH, Olsson Gunilla M. Adolescent Adjustment Profile - revised and investigated by means of an Item Response Theory approach. (Manuscript).

IV. Adler M., Hetta J., Isacsson G., Brodin U. An Item Response Theory evaluation of three depression assessment instruments in a clinical sample. BMC Med Res Methodol. 2012 Jun 21;12(1):84

V. Adler Mats, Brodin Ulf. An IRT validation of the Affective Self Rating Scale. Nordic Journal of Psychiatry, 2011, Vol 65 No 6, p 396-402.

Contents page

The three step strategy

Appendices

Abbreviations

AIC = Akaikes Information Criteria
AISP = Automated Item Selection Procedure
BCa = Bias corrected C.I. for the percentile method.
BIC = Bayes Information Criteria
C.I. = Confidence Interval
Cor(u,v) = Correlation between variables u and v
CTT = Classical Test Theory
d.f. = degrees of freedom
DIF = Different Item Functioning
GRM = Graded Response Model (Usually with a common set of category thresholds)
E(Y) = The statistical expectation of the stochastic variable Y.
I = Item information, or the number of Items in an enumeration i= 1 … I
IIO = Invariant Item Ordering
IRF = Item Response Function
IRT = Item Response Theory
LCL = Lower Confidence Limit
LID = Local InDependence
LR = Likelihood Ratio
LRT = Likelihood Ratio Test
MNSQ = Mean Square
P(A) = The probability of event A
PCA = Principle Component Analysis
RSM = Rating Scale Model
RSS = Raw Sum Score
S.E. = Standard Error of measurement
S.D. = Standard Deviation
UCL = Upper Confidence Limit
ZSTD = Standard deviation expressed as z-score.
(a | b) = 'a conditioned on b'.
1st Qu. = First quartile, 25% of the observations are ≤ the stated value.
3rd Qu. = Third quartile, 75 % of the observations are ≤ the stated value

Indices in the Mokken analyses of monotonicity and non-intersection

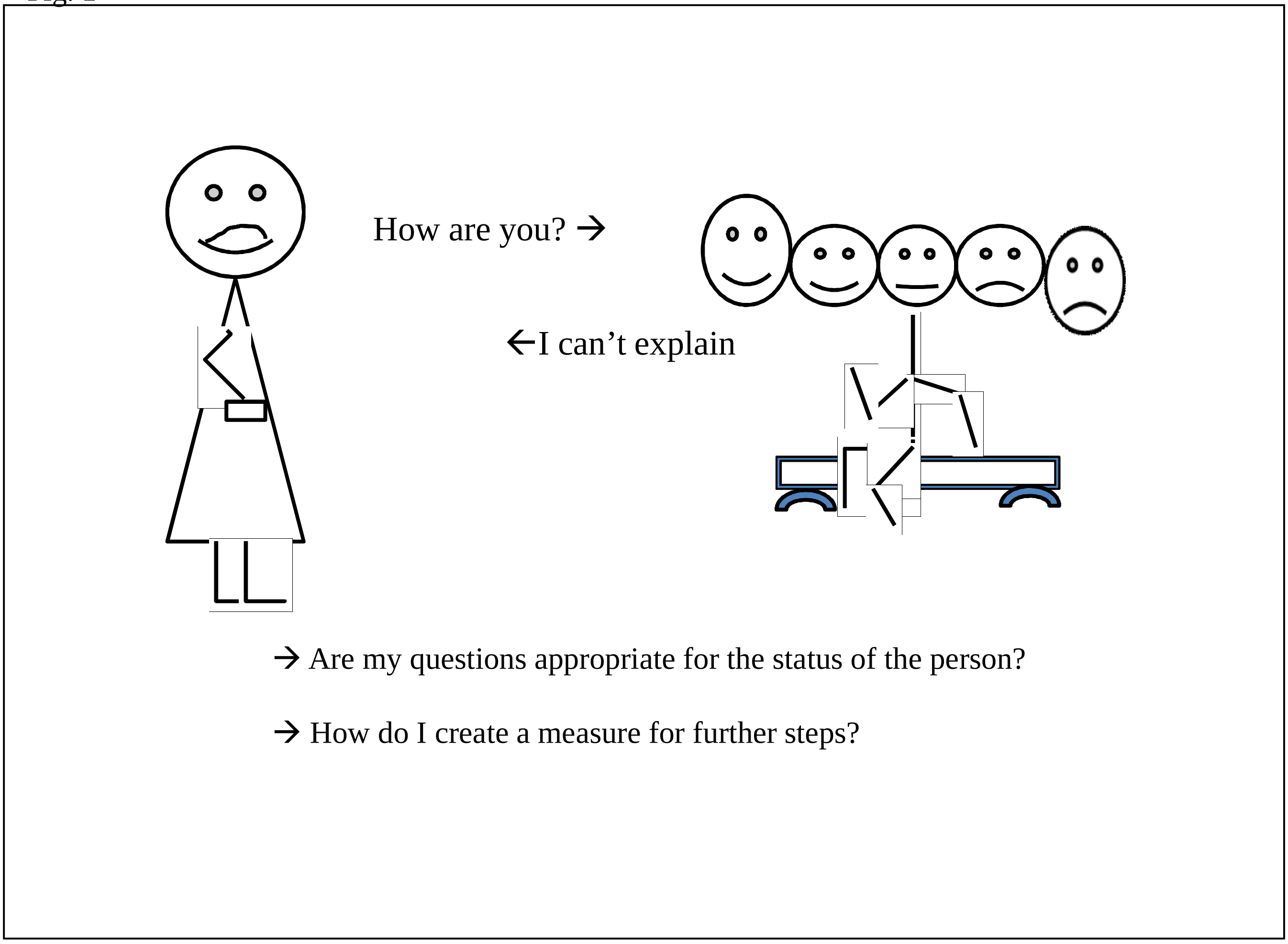
# = number of
#ac = the number of active pairs (comparisons) for each response level.
#vi = number of violations greater than a specified minimum. Default = 0.03
sum = sum of violations
maxvi = the maximum violation
zmax = the maximum violation expressed as z-score
minsize = minimum group size in forming rest score groups
#zsig = number of significant violations

Background

(Frustration is often the origin of good ideas)

Getting information about person characteristics, of a type which is not directly measurable or can be observed, such as depression, attitudes, social functioning etc. (i.e. latent variables), is a common problem in many research areas. Much attention has been devoted to how to capture this type of variables and the measurement theory has steadily developed since the 1930's, particularly in medical, social and psychological research. The most frequent way of carrying out such a measuring procedure is to set up a questionnaire with relevant questions, collect the answer profile, create some aggregated measure and let this measure represent the individual on a latent scale, a constructed dimension of interest. A common approach is the construction of a questionnaire where the individuals are asked to reply to a set of questions or to decide on statements in terms of more or less agreement to what is stated. Thus, the researcher often might ask her-/himself questions like what is illustrated in figure 1 below.

Fig. 1

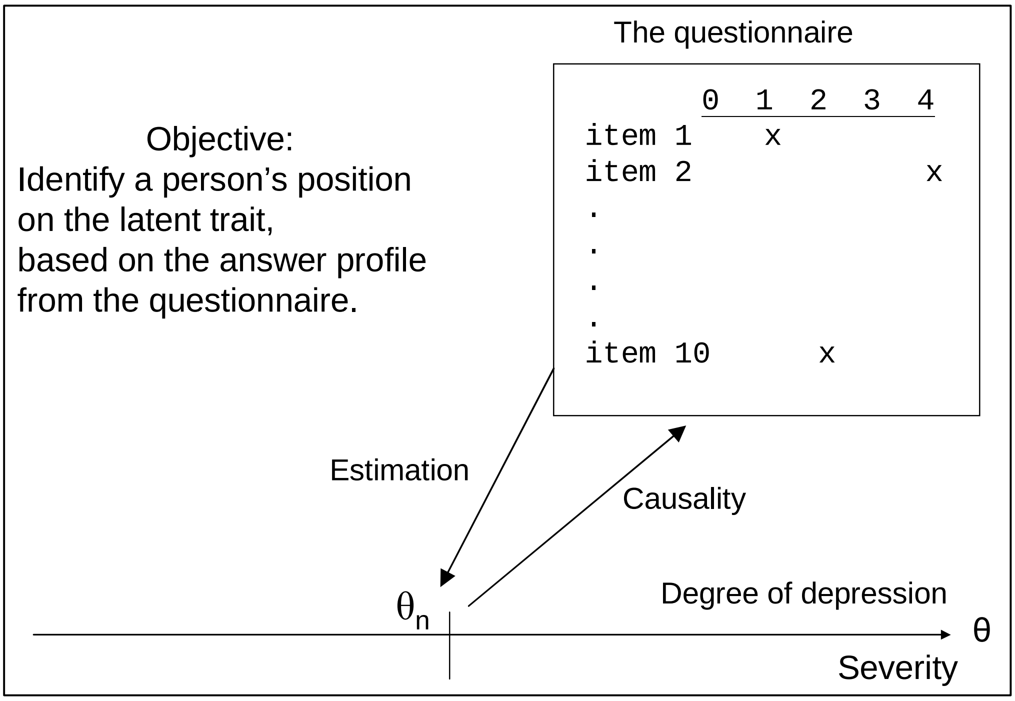


→ Are my questions appropriate for the status of the person?

→ How do I create a measure for further steps?

An overwhelming majority of such studies are based on straightforward questionnaires with a set of items where the response alternatives are ordered according to the degree of the actual phenomenon. The questionnaire can be self-administrated or interview directed.
Usually, such a questionnaire is focused on one single underlying latent variable and all items have the same (or slightly adapted) wording – ‘strongly agree’, ’agree’, ‘neutral’, ‘disagree’, ‘strongly disagree’, -‘never’, ‘occasionally’, ‘quite often’, ‘always’ or similar sets of ordered answers, often referred to as Likert scales. The intention behind such a construction is to create a vector of response variables, where each individual gets its own response profile. In a next step, an aggregated measure is formulated, in some way or another, by summing up the respondents ‘reaction’ to the items, i.e. turning the set of responses into one comprehensive measure. In order to manage the collected data in a simple form from such a study, consecutive numbers are attached to the response alternatives. ‘strongly agree’, …, ‘strongly disagree’ are replaced by 0, 1, 2, 3, 4 or 1, 2, 3, 4, 5. The relation between the individual and the response profile is outlined in fig. 2.

Fig.2.



The respondent’s actual state of depression, the degree of depression in fig. 2, is the (hidden) source (Causality) when answering the questionnaire. The researcher then tries to find the source (Estimation) by collecting information from the questionnaire and ‘transform back’ to a scale representing the depression.
Attaching consecutive numbers to ordered response categories makes it tempting to treat the numbers as metric measures. A metric interpretation of the sum of the item scores is also appealing when the response categories have the same wording of the categories. This is idea behind the Likert scale, assuming a linear relationship between the response probability and the underlying trait, which is not particularly plausible. Even if there is an order, the distances between the numbers are not inherent. The distance is determined from the latent scale, of which we do not know anything, so far.

As a medical statistician I have come across with a large number of questionnaires. My experience is that the researchers' intention as well as their expectation, when constructing a straightforward, one dimensional questionnaire with items of the Likert type, can be summarized as follows:

1. The raw sum score, RSS, is a reliable person measure.
2. For each item: Higher score ➔ more of the latent trait (Monotonicity).
3. The sample of individuals is supposed to be homogenous with respect to how the persons perceive the questionnaire. I.e. the importance, or at least the ranking of items, is the same for all respondents or subgroups of respondents (Invariant Item Ordering, IIO).
4. The items are thought to contribute equally to the aggregated measure, i.e. they have an equal capacity to discriminate persons on the latent scale. (Unit discrimination)
5. The questionnaire is designed for one underlying latent trait (Unidimensionality).
6. The instrument should be able to reliably measure all individuals belonging to the target population, i.e. at all levels of the intended dimension (Good coverage).
7. The questionnaire is aimed for general application in clinical practice, but a first evaluation will be carried out on a limited sample. The reason might be that the researcher wants to evaluate its basic characteristics at an early state or that the target population is small, difficult to reach or there are economic or time restrictions.

Even if the intention is to create a sum score, or any other aggregated measure for further use as a metric variable (possibly after some transformation), this has to be investigated and validated. It should be noted, that forming an aggregated measure based on a set of items, such as a sum score, implies setting up a model like any other statistical model such as regression, ANOVA etc. The model has to be verified and its applicability investigated. It is far from evident that summing up the items in a questionnaire automatically constitutes a variable ready for use and interpretation as if it were a valid continuous metric variable. However, if a reasonable metric variable can be constructed, there is a large variety of suitable methods available within the Classic Test Theory (CTT). Most of these methods can easily be found in statistical packages like SPSS, SAS, STATISTICA etc. Sometimes non parametric methods are applied on a sum score, mainly due to concerns about the normal distribution, but this approach does not circumvent the basic problem - a sum score is not a proper metric variable.

There are two main reasons, as evident as elementary:

1. A sum of ordinal variables (if possible) does not automatically results in an interval scaled variable.
2. A sum score variable does not have a fixed scale parameter.
3. The alternatives within an item are defined by analogy, not by a metric unit.

The naive approach is however understandable, as nothing or very little is said about proper handling of questionnaire data in most of the basic courses in statistics.
It is astonishing that so many studies are carried without a least bit of adequate verification. My impression is that the researchers are not aware of that they are creating a model. In general terms, even a sum score is a model with its assumptions and qualities.

In medical, as well as psychiatric and pedagogic journals, questionnaire studies are frequently reported. A considerable amount of these studies is based on small samples, often < 100 individuals. In many cases, the evaluation is ambiguous or unsatisfactory due to the improper use of sum scores together with unmet assumptions, at least approximately required for the applied methods. In the available literature, very little attention is paid to the problem of how to evaluate questionnaires with only a small sample at hand.

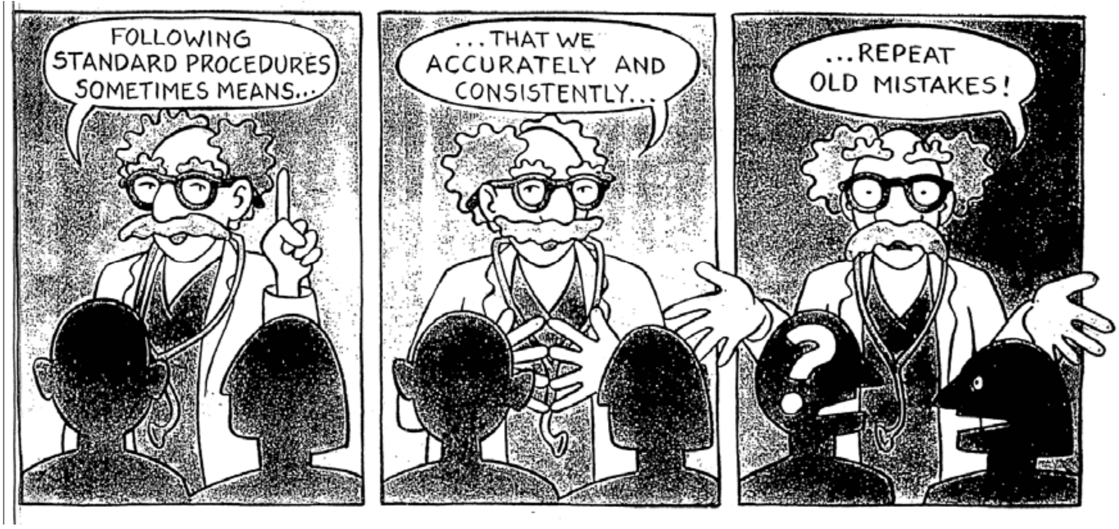
There are statistical methods, based on Item Response Theory (IRT), which are especially developed for evaluation of questionnaires with ordered items, and fully presented in the literature [Bezruczko 2005, Bond & Fox 2001, de Ayala 2009, Embretson 2000]. In conformity with de Ayala 2009 and Embretson, I consider Rasch models as a branch of IRT although many statisticians will keep them apart. Rasch and other models used in thesis is further explained in Appendix B.
'In Item Response Theory (IRT), a person's trait level is estimated from responses to test items' [Embretson, 2000]. In claire: (And to my experience) The researchers use a questionnaire to place a respondent on an interval scaled latent trait, representing the dimension of interest, such as personal attitudes, depression, social function etc..
However, IRT methods are mostly proposed for application on large studies with the intention to find a representative model and to determine parameters for further use when evaluating individuals (patients, examinees etc.). de Ayala 2009, Edelen & Reeve, 2007 and Reeve & Fayers, 2007, give some advice about the required sample size, but the recommendations are directed towards finding suitable models, for which rather large samples are required. Some interesting plans for investigation of models are presented in Bot et al., 2004 and Reeve et al., 2007, but again, with large samples in mind. "Most IRT models are slanted towards understanding IRT application within context of large-scale educational assessments" [Embretson 2000]. Linacre, 2004, proposes rather large samples for 'stable model estimates' in a questionnaire with polytomous items. As much as 25 observations per category is discussed but, as a guideline, he proposes at least 10 observations per category. However, only Rasch models are considered.
Bot et al.,2004, is focused on CTT methods but mention in the discussion that 'The evaluation of shoulder disability questionnaires may be improved by using IRT'? Goodman & Scott, 1999, compare questionnaires with a moderate sample (n=132), but they use receiver operating characteristic (ROC) and structural equation modelling rather than an IRT approach.
Schumaker, 2004, has reported a small sample study where a 12 item 'Mobility Scale' is administered to elderly independent living individuals, n= 108. Although the features of the items are analysed in some detail, the questionnaire is restricted to dichotomous items and a Rasch model. Smith, 2005, analyses a study with just 32 individuals in two groups. A Rash model is considered and the analysis relies on, what is seemed to be, an established questionnaire, entitled Post Traumatic Stress Disorder. According to R.J. de Ayala, Wright, 1977, stated that "useful information can be obtained from samples as small as hundred". In some research areas, such as pedagogy, n=50 is considered a rather large sample (e.g. Study I). However, this standpoint is based on pragmatic conditions rather than statistical issues.
Where the sample size is discussed to some extent, it is focused on the required samples size from a statistical point of view rather than how to proceed with a limited sample available.

The proper use of IRT and its advantage is frequently overlooked (Embretson 2000). Many researchers are completely convinced that a sum score works as a proper metric variable. Their conviction is often based on the large amount of published papers where the misuse/misunderstanding of statistical methods' application on questionnaire data is common practise. This is sometimes explicitly discussed [Bond & Fox 2001; Embretson, 2000; Good 2009, Jamieson, 2004]. It is also a common misunderstanding that a questionnaire can be objectively validated. If a questionnaire is translated from a foreign language and/or applied on different a population, a new evaluation of its quality is required.

If not “It would be like visual observation using eyeglasses borrowed from someone else. It is bound to produce unclear or suboptimal results” [Sijtsma & Molenaar, 2002].
The reason is that the characteristics of a questionnaire instrument, including a possible scale structure, are set by the sample of persons which is given the questionnaire. Even the possibility to sum up items is decided by the answer profiles, not by the constructers even if it is their intention. This is in contrast to a metric instrument, where the characteristics and the scale unit is decided beforehand by the constructers, independent of any target population. As an example, the construction of a ruler to measure the height of person is validated, including the scale unit, without any reference to a certain population or environment.
IRT-methods have been suggested as a tool for the development of better instruments for the evaluation of depression [Bagby et al., 2004]. Wilson, 2005 suggests detailed strategies for construction of questionnaires.

There is a tendency to repeat statistical analyses carried out in references or in articles written by near colleagues. This practise ought to be changed. There is a need for some strategy, which can convince the medical and psychological researchers about the advantages of IRT methods and encourage them to actually carry out some basic investigation of the quality and characteristics of their questionnaires, preferably at an early stage, see figure 3.

Fig. 3


FOLLOWING STANDARD PROCEDURES SOMETIMES MEANS...
...THAT WE ACCURATELY AND CONSISTENTLY...
...REPEAT OLD MISTAKES!
?

With permission from Gunnar Kaj & Ragnar Levi

Why do not researchers use IRT methods to a larger extent?

- IRT methods are usually not considered in basic courses in medical statistics. Most literature in medical statistics does not even mention the existence of IRT.
  There are traditions within scientific institutions where the researchers tend to follow ‘standard procedures’. Following standard procedures sometimes means that we accurately and consistently repeat old mistakes. ”If an erroneous method is applied frequently enough, it will be considered correct.”

- In case IRT methods are known, they are perceived as complicated and not appealing as they are not in the agreement with the researchers' perception of statistical thinking, which is based on CTT.
- Most of the IRT literature is about large studies, where a relatively large sample is seen as a prerequisite of getting reasonable IRT models and thereby good estimates of person measures on the latent trait, which as such, is a relevant attitude when looking for a model for further use. Very little is said about the relation between a small sample and a primary pragmatic evaluation of a questionnaire. I have not found anyone taking care of this particular problem.
- "The few available textbooks are not easily accessible to the audience of psychological researchers; the books contain too many equations and derivations and too few familiar concepts" [Embretson, 2000].

Why not the CTT approach?

It is well documented in the statistical literature that an item score or a 'raw sum score' is not a valid interval scaled variable. [Embretsson E. Susan, Reise Steven P., 2000; Smith Richard M., 2003]. Thus, CTT methods on sum scores are not relevant and potentially misleading. The analyses are generally wrong or have no meaning as these methods are based on assumptions that are not met by questionnaire data. 'Rasch measures is the only way to convert ordinal observations into linear measures', [Fischer, 1995].

The aim of the thesis

The aim of this thesis is to develop a pragmatic, but still scientifically sound, strategy for investigation of small questionnaire studies, where a raw sum score, RSS, (or a transformation of it) is thought to constitute an aggregated measure of an individual's position on the underlying latent trait.
The thesis is directed solely towards questionnaires formulated as a set of questions/ statements with a common set of ordered response alternatives (often referred to as a Likert scale).
A further aim is to find hints for a 'redesign' or 'design corrections' of an instrument at an early stage before it is generally applied in clinical practice.

The thesis is focused on quantitative studies where the sample sizes, in most of the literature, are considered inadequate (too small) for proper evaluation. In practice, this means in the range 50 – 150. No firm limit can be drawn but less than 50 persons in a study usually imply too many 'empty cells' and very loose conclusions. If the upper limit is enhanced, further evaluation about the persons estimated position on the latent scale, as well as other characteristics might be of interest – and the study might no longer be entitled 'small'.
The fundamental approach:

- To evaluate basic characteristics of a questionnaire when only a small sample of respondents is available.
- To use IRT methods to investigate the relationship between an underlying latent trait and the observed answers to the response alternatives of an item.
- The respondent's choice of a particular answer category of an item with ordered categories is a probability function, based on an individual's position on an underlying latent trait.

Item Response Theory approach, IRT, is mainly associated with large studies. Its use in small studies is sparsely presented. There is a considerable distance between the researchers' small studies and the large-scale national studies, where efficient and relevant methods, suitable for evaluation of

questionnaire data, are applied. By re-analysing some studies already carried out, and where the individual RSS is designed to be an aggregated person measure ‘according to standard procedures’, I hope to be able to show how an IRT approach can be used to gain enhanced insight in a questionnaire, although the study is quite small. Characteristics such as the relevance of a sum score, bad or non-co-operating items, incoherent answer profiles and much more can be revealed. The aim is not to find a final or an applicable model for further use, as in most studies where IRT methods are involved, but rather focus on the questionnaire as such. The aim is an ‘advisory message’ about the questionnaire, rather than a solution. In this sense, the suggested strategy should be considered as an explorative process, where a ‘significant’ result from a statistical test should be perceived as a strong indication how to proceed rather than a firm decision.
The theory is not new but I will encourage researchers to use a combination of IRT methods. My intention is to show that a well thought-out strategy of informative IRT- analyses can be performed even in small studies, and thereby gain valuable information in order to correct and improve the questionnaire. Ideas about constructing and improving questionnaires are also discussed by Wilson, 2005.

After developing the strategy, a set of reasonably accomplished studies are needed to evaluate how the proposed pragmatic strategy works. The chosen studies are not identical. They have, except the small sample size, their own characteristics and deficiencies. I thought it would be a great advantage to collaborate with the researchers, responsible for the studies. They are experienced in their domains, familiar with the target populations and have well founded motivations for the choice and formulation of the items. In such a way, the procedure could be anchored in their environment, and consequently, the Study I – V were written in collaboration for an audience of practising clinicians. This also led to avoidance of a lot of mathematics in favour of explanatory reasoning.

I started with Study I as it had a straight forward layout with just three items, just 59 persons and an supposed fairly homogeneous sample. The result was encouraging, so I continued with studies of typical sizes within domains where this type of questionnaire studies frequently are carried out.

A brief description of the five studies

All five articles are theoretical methodological studies, where modern statistical methods are applied in constructing and analysing questionnaires. The methods are applied on already collected and analysed data. The data from the studies, or applications of the questionnaires, are in some cases used in already published articles where conventional methods according to CTT have been used.

Table 1. Characteristics of the five questionnaires, included in this thesis

Study	The latent variable/ dimension	No. of items	No. of categories/item	No. of persons, N, and in subgrups, n_i.
I	Teacher's teaching practice towards the 'activating' of - application of knowledge (AA) - meaning of knowledge (AM) - reproduction of knowledge (AR)	3 3 3	5 5 5	N= 57
II	Psychological dimension Social dimension	23 12	5 5	N= 87 n_1 =25, n_2 =30, n_3 =32
III	Attention deficit Externalising behaviour Internalising behaviour	13 11 8	5 5 5	N= 131 n_1 =65, n_2 =66
IV	Depression – based on PHQ9 Depression – based on AS-18-D Depression – based on MADRS	9 9 10	4 5 7	N= 61
V	Depression - AS-18-D Mania - AS-18-M	9 9	5 5	N= 231*

*A large subset scored zero to all items and were 'outside the range' of the questionnaire. The effective sample was reduced to n=174 (Depression), n=151 (Mania).

The designs of the questionnaires, where the items within a questionnaire have a common structure in terms of ordered answer alternatives, is presented in table 2.
The questionnaires are different in terms of the number of items and response categories. For further details about the questionnaires, see Appendix A.

Table 2. The typical design of a questionnaire with 10 items, ordered response alternatives and a common set of answer categories. The questionnaire is designed to capture one single underlying trait.

	Response alternatives to questions about problems					Code
	No (0)	Just a little (1)	Moderate (2)	Large (3)	Very large (4)	Y_i
Item 1				X		3
Item 2			X			2
…						…
…						…
Item 10	X					0

In Study V, the MADRS questionnaire does not follow exactly the structure outlined in table 2. The wordings of the answer alternatives are item specific, but the layout is similar to the other questionnaires, with a common item scale 1 to 6 and the author's intention form a straightforward sum score.

The 3 step strategy and the basic theory

Why is a specified strategy desirable?

Like all instruments, a developed or translated questionnaire has to be validated before it is taken in regular use. Such a procedure has its special difficulties when the available sample of respondents is limited. A questionnaire might appear simple and straightforward. However, it is composed of set of items (ordinal variables) with a complex structure of relationship, for which the usual means of evaluation, according to CTT, are not suitable. In general, an evaluation of a questionnaire is focused on the sum score and its relation to the individual items. We might say that the sum score is an approximate rating of the respondents, but by no means an interval scale. The ultimate goal is to estimate a person's position, on the dimension of interest, with as a high precision as possible. In general, this requires responses that improve the power of person fit indices. Ideally, this implies longer questionnaires (e.g. more than 30 items), a wide range of item location parameters and highly discriminating items [Embretson & Reise, 2000].

This also requires an approach by methods specially developed for questionnaire evaluation. I suggest, that in a small sample situation, it is essential to start 'from scratch' and evaluate very basic characteristics of the items, and from there, possibly proceed to a more elaborated modelling procedure. During more than 40 years of experience, as a medical statistician, I have worked with a number of studies involving questionnaires. This experience led to the idea, and formulation, of a 'strategy', which hopefully would make the life easier for statisticians as well as for researchers.

The hypothesis is that the following '3-step' strategy will be a sound and pragmatic tool, readily available for clinicians and other scientific practitioners, avoiding too much theory and with, to at least some extent, existing computer programs available. The proposed strategy is designed to find pragmatic hints for an efficient redesign of a questionnaire, where certain characteristics are required. The pragmatic execution of the 3 step strategy can be attained by a combination of existing computer programs. In this work I have used the Mokken (non-parametric) scale analysis in R [Andries van der Ark 2007], Winsteps, [Linacre 2008] and An R package for Latent Variable Modelling and Item Response Theory Analyses [Rizopoulos 2006]. Other combinations of existing programs are of cause possible. Complementary analyses and illustrations are achieved by specially written R- programs and by the general purpose statistical program package STATISTICA [STATISTICA, 2011].
In the literature, the Rasch approach is frequently considered not belonging to the IRT domain. I consider the Rasch approach as a step among others within the IRT area, even if it is based on somewhat different estimating procedures and assumptions. As will be seen, the '3-step' strategy will go through a series of statistical models, and their variations, in an explorative search for informative structures of the questionnaires. However, we have to remember that *'all models are wrong, but some are useful'* [Everitt & Dunn, 1991]

The three steps are:
Step 1 evaluates, by means of a non-parametric approach, the capability of the set of items to cooperate towards a common aggregated measure, representing a respondent, for ranking the individuals.

Step 2 evaluates the possibility to use a sum score as a 'sufficient statistic' to create a reasonable 'person measure' on an interval scale. The Rasch approach is a first parsimonious step.

Step 3 evaluates whether a model in step 2 suffice for a reasonable description of the characteristics of the questionnaire or if a further extended model is needed. The item discrimination, item and total information is of special interest.

A motto for this strategy: Try to keep the interpretation of analyses as simple as possible and keep hold of parsimonious models unless they are strongly contradicted.

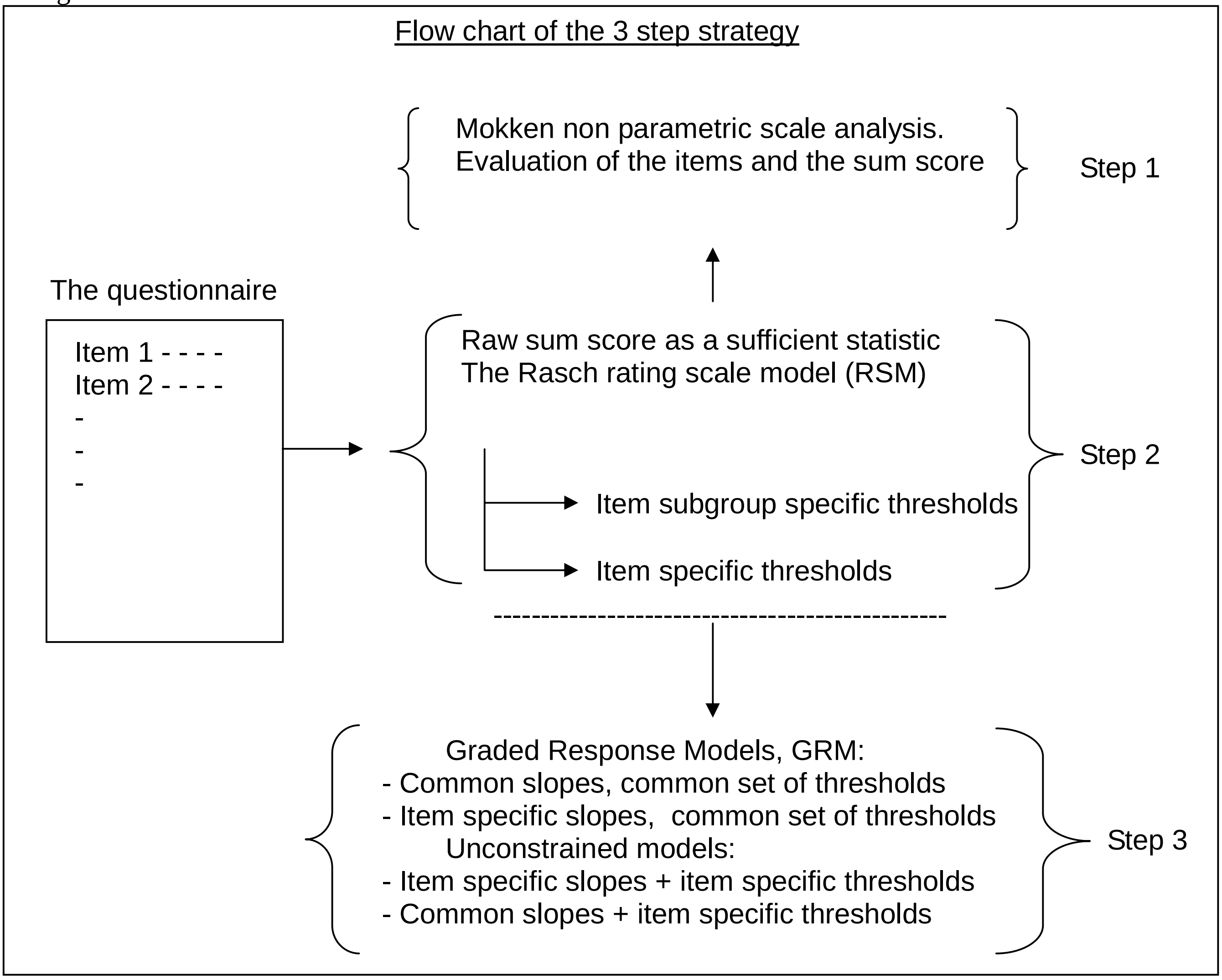
In most clinical studies, where CTT is the relevant approach, observed data usually represent a set of variables, of which some are considered dependent, or result variables, while others are considered independent (treatment, placebo, prognostic factors, baseline data etc.). The statistical analysis is essentially a search for a relationship, where the dependent variables (usually called Y- variables) are modelled as functions of the independent variables (usually called X- variables), i.e. treatment effect as a function of treatment, age, sex etc..
In studies based on questionnaires, the questions/items are the Y- variables and the responses are observations of these variables. The main variable of interest, which is the source of the observed responses to the items, is the unknown latent trait (the X- variable). I.e. we observe a lot of Y-variables, but have no observations of the independent X- variable – a situation quite different from a typical clinical study - and quite different from statistical procedures applied for the evaluation of such a study. In this setting we get the following from respondent 'n': $(Y_{n1}, Y_{n2}, \ldots, Y_{nI})$, n=1,…,N where N is the sample size (the number of respondents) and the Y_i:s are the category responses (the attached number code of the chosen category) to items i=1,…, I. The response Y_{ni} is thought to be a function, F, of the characteristics of the item i, say δ_i, and the individual's position, θ_n, on the unknown underlying dimension, i.e. $Y_{ni} = F(\theta_n , \delta_i)$.
θ_n is the unobserved and not measurable X- variable on the latent dimension and F is a suitable probability function, relating the response vector with its source. δ_i is a parameter determining an item's position on the latent scale. It should be noted that the latent variable is a construction which is more or less successful in capturing the actual phenomenon.

Let us choose an aggregated measure, T_n , representing person n. T_n is thought to be an efficient and reliable message from the person via the questionnaire:

T_n 🡐 The questionnaire 🡐 Person n, where $T_n(Y) <= \sum a_i Y_{ni}$, i= 1,2,…,I
(🡐 means 'a function of')

This is the most simple model, i.e. the raw sum score (RSS). a_i is a possible item weight, but usually it is assumed to be equal for all items in the RSS model (in general, a_i is set equal to 1)
The authors of a questionnaire usually have the RSS in mind when they construct the questionnaire. The usefulness of RSS is not evident and we have to start our evaluation by taking a step backwards and investigate whether the RSS is a reasonable measure, and even more basically, if a summation of the item scores is a relevant procedure. The procedure of the '3 step strategy' is outlined in fig 4.

Fig. 4.


Flow chart of the 3 step strategy
Mokken non parametric scale analysis.
Evaluation of the items and the sum score
Step 1
The questionnaire
Item 1 - - - -
Item 2 - - - -
-
-
-
Raw sum score as a sufficient statistic
The Rasch rating scale model (RSM)
Item subgroup specific thresholds
Item specific thresholds
Step 2
Graded Response Models, GRM:
- Common slopes, common set of thresholds
- Item specific slopes, common set of thresholds
Unconstrained models:
- Item specific slopes + item specific thresholds
- Common slopes + item specific thresholds
Step 3

A brief description of the models, slopes and thresholds is presented in Appendix B. It is further illustrated in fig. 6 on page 27.

Full mathematical formulation of IRT models for ordered data, together with a conceptual description of the estimation methods, can be found in de Ayala, 2009. A complete description of Rasch models can be found in Smith EV and Smith R.M, 2004, while a more pragmatic description can be found in Bond & Fox, 2001. A comprehensive compilation of polytomous item response models is presented in Ostini & Nering, 2006.

In evaluation of the questionnaires by use of the IRT methods, outlined in the ‘3 step’ strategy, some recurrent concepts are essential in all three steps and call for brief descriptions (more details are added in the text when required). These are described below:

Unidimensionality

The questionnaire is intended to measure only one underlying latent trait. In practice, after fitting an IRT model, there is still a certain amount of unexplained variation which cannot be related to the intended dimension. This 'residual' variation is supposed to be 'random noise' and small compared to the variation explained by the model. If there are only negligible systematic components in the residual variation, we might say that the questionnaire measures one main underlying trait, however with varying quality. The effect on parameter estimation of small departures from unidimensionality remains undemonstrated. 'In fact, some research indicates that IRT model parameter estimation is fairly robust to minor violations of unidimensionality, especially if the latent trait dimensions are highly correlated [Embretson & Reise, 2000]. In a small sample study, I suggest that the dimensionality is evaluated in step 1, before any particular parametric model is applied.

Local InDependence (LID) vs Local Dependence (LD)

Local independence is a condition for a meaningful sum score approach as well as for reasonable estimation of more elaborated IRT models. This means that, conditioned on an individual's position, θ, on the latent trait, the items are assumed to be independent. Observed relationship between item responses is supposed to stem from the fact that one individual, with a fixed level of level θ, has responded to all items. Thus, all dependence is regressed on the individual. Local independence is considered a prerequisite for meaningful likelihood functions, as they are composed by multiplicative probabilities for estimation of the parameters. LD affects the estimation of test information and item discrimination parameters, making both of them larger than they should be [Yen, 1993] according to Embretson & Reise, 2000.

Conditional independence can be investigated by correlating residuals after applying a model. The residuals for an item are the differences between the individuals' response to an item and their expected responses according to a suitable model.

$Res_{ni} = Y_{ni} - M_{ni}$ where i=item i, n= person n. Y_{ni} is the observed person score and M_{ni} is the person score as estimated by the model. The set of Res_{ni} for all i:s and n:s contains all information about the chosen model.

The residual correlation $cor(Res_{ni}, Res_{nj})$ between any two items, i and j, and calculated over the sample n=1,…,N is assumed to be zero on the assumption that we have found a suitable model. A set of one or more large correlations is therefore a sign of incompatibility with the assumption of LID. However, even if the model is perfect, the pairwise correlation between the residuals for two arbitrarily selected items will be biased, $E(cor(i,j))\neq0$, as the original observations are included in the estimated model. A correction of this bias, as suggested by Yen, 1993, is possible but might be hazardous in a small sample situation, with unknown consequences. It will influence an already very approximate estimation and is therefore avoided in this work. For further explication, see Appendix B.

As parsimonious models, such as the Rasch or the GRM model, should be preferred in a small sample situation, we have to accept some remaining dependence between the residuals (more complicated models such as a model with item specific discriminations and thresholds, might solve the problem but will probably be too data driven, i.e. 'over parameterized').

How large a correlation between pair wise item residuals can be accepted? We know that the real correlation is $\neq 0$, so hypothesis testing has no meaning. A $|r| \leq 0.3$ might be acceptable, which means $0.3^2 = 0.09$, just about 10% of the variation for one of the variables is explained by the other variable. This moderate amount of dependence is not very influential on an estimating procedure which assumes LID.

Monotonicity
This characteristic of the questionnaire means that for increasing degree (severity) of the respondent on the latent trait, the probability of endorsing higher categories of an item is non-decreasing. In the nonparametric approach, where the person measures θ are not estimated, different levels of θ are represented by rest score groups (further explained in Step 1 and Appendix C).

Non intersection and IIO
For the type of questionnaires considered in this work, Invariant Item Ordering is a desirable characteristic. This means that the order of the item locations on the latent trait is the same for all levels of the person measure, θ, as well as for subgroups. This can be investigated by 'non intersection' graphs, where rest score groups are used (as described in Step 1). A brief description can be found in Appendix C.

Differential Item Functioning, DIF.
The questionnaires are usually constructed with the intention of a general applicability for the target population. However, subgroups might perceive the questionnaire differently, usually seen as group differences in item locations, δ_i. If the difference is of the same magnitude over all items, in the sense of high or low scoring, it will be housed by the aggregated measure from the item scores and the measure can be classified according to an external criterion, such as sex, status of health etc., (difference in degree). If the DIF is concentrated to one or a few items, the groups are classified on the basis of these specific items and the aggregated measure (such as the sum score) is more or less invalidated when groups are to be compared (difference in kind). This is particularly troublesome in a case where the DIFs are reversed for some items. Although DIF are connected to parametric models it also has it relevance in step one, where large differences between groups, in terms of item scalability, might be seen as a DIF in a broader, more unspecified sense.

The relation between item location and the thresholds
The basic characteristics of an item is its location δ, the centre of the item on the latent scale, and the thresholds (category boundaries), which are the distances, τ_j j=1,…,m-1, from the item location. The boundary for category j is then described as $\delta + \tau_j$. The first and the last categories are open, see fig. 6 on p.27.
The primary interest is in item locations. If the items are too close to each other, a difference between items are moved to differences between thresholds, particularly if these have a large span such as ultimate wordings like 'Never' and 'Always' for a question about 'How often?'. On the other hand, starting with an unconstrained model (item specific thresholds) may force the item locations to be close to each other if the thresholds represent a large span.

Reliability
The basic idea is: The degree of stability of a respondent's score across independent replications of an administration of the questionnaire. This idea is also applicable for the items. The structure and the perception of the items should be the same at such a replication. The concept of reliability is outlined in Appendix B.

The estimation procedure, the 'Likelihood approach'
The likelihood approach is a computer intensive procedure to maximize an estimation function in such a way that the result of the function becomes as close as possible to what is actually observed. This is

achieved by optimizing the function's parameters by means of a 'maximum likelihood estimation procedure', ML.
As an example – consider a three item questionnaire where the respondent 'n' has the answer profile $(1,3,3)_n$. Such an observation is modelled as $P(Y_{n1}=1)*$ $P(Y_{n2}=3)*$ $P(Y_{n3}=3)$ with the parameters θ_n, δ_1 ,δ_2 ,$\delta_{3,}$ the set of τ_k:s (the set of thresholds in terms of the distances from the item locations) and a suitable function for the probability P. An assumption of local independence is required for a simple multiplication of the probabilities. The analogous expressions for all respondents are multiplied together (as the persons are assumed to be independent of each other). The resulting expression (the likelihood) is then maximized with respect to the parameters. In theory, this could be done by non linear models in some of the ordinary statistical computer packages, but the large set of parameters make it impossible in most cases, at least in case of small samples. Person parameters (θ_n) and item parameters have to be alternatively estimated and refined by the ML method. For further description, see Appendix B. The Rasch model and it's extensions relies on the person's total sum score as a sufficient statistic which implies that the estimation procedure is based on the total sum scores and the frequency of every observed sum score. This results in a somewhat different use of the ML method. Embretson, 2000, gives a well formulated explanation of the different ML-methods.
Further model specific details are outlined under the respective headings, Step1, Step2 and Step3.

Total and item information
Information is a statistically defined indicator of measurement quality and varies as a function of θ, the person's position on the latent scale. The informative value for a specific item, the item information, is of special interest and can be inspected at each point along the latent variable. It serves as an indicator of the usefulness of an item, mainly from two aspects. By looking at the item information, compared to the other items, its relative importance for the questionnaire can be evaluated. Poor item information might be used as a basis for exclusion or reformulation of an item. The second aspect is an item's sphere of activity, which, in general terms, should be concentrated to the vicinity of the item location. The item information is closely related to the discriminating power and inversely related to the measurement error.
The total (or test) information is the collected information from the complete set of items and can, as well as the item information, be inspected at each point along the latent variable. Its main interest lies in the outline of the information curve and in general it should rise at the beginning of the person measure range and remain at approximately the same level over the range. However, the ideal curvature depends on the purpose of the questionnaire.

In the IRT literature, the concept of information is drawn from the general statistical theory. Having created a continuous probability function, assumed to be twice differentiable, the information is based on the 'second derivative of the 'LogLikelihood function'(LogL), based on the so called 'Fisher's information'. The definition and use of the 'Likelihood function' is fully outlined in de Ayala, 2009 for dichotomously scored items, and extended to polytomous items. Basically, the total information is the sum of the items' information as the items are assumed to be locally independent. However, this approach is developed under the large sample theory and I suggest it to be handled with care in a small sample environment. The second derivative (with respect to the slope) tells us how rapidly the slope (the discrimination) is changing around the maximum likelihood estimate. A large second derivative corresponds to a good estimate, and thus a small standard error. In case of dichotomously scored items, the item information is the squared discrimination multiplied by the probabilities of positive and

negative answers, $a_i^2(P_i(1 - P_i)$. In the polytomous case, the calculation of the item information is not straightforward as there are ≥ 2 category thresholds, not independent of each other.
The behaviour of this information is not evident, mainly for two reasons:
1, The likelihood function is based on strictly locally independent items.
2, It's relevance for small samples is unclear.

The IRT literature recommends examination of item information before examination of person information [de Ayala 2009 p. 55]. However, this recommendation is based on a large ratio of the number of persons to the number of items, that is, with large samples in mind. When such a ratio becomes small, let us say #persons/#items <10, the influence from individual persons might be substantial. Even if 'Inspecting item information functions for each item is an essential item analysis procedure' [Muraki, 1993], is a valid recommendation [Van der Linden & Hamilton, 1997, ch. 9], it is not obvious how to proceed with a small sample. In such a situation I suggest taking care of an incoherent subject answer profile before further investigation of the model.
In a basic Rasch model, with a common set of category thresholds, all items are postulated to have equal information available for estimating person measures. In this case, the total information is of greater interest. However, in extended models, the item information is strongly, related to the discrimination coefficient.
In an unconstrained model, where we have item specific category thresholds and discrimination coefficients, the total information is the sum of the items' information (due to the independent estimation of item thresholds). Thus, $I_k / \sum I_k$, (k=1,…,K items) will be the relative information of item i. A more thorough description can be found in Ostini & Nering, 2005. Estimates can be calculated by the R program for Latent Variable Modelling (ltm), [Rizopoulos, 2006].
Such a model will be too complicated for a small sample study, but might be worthwhile to consider as the information will be released from the constrained estimation of a common set of thresholds. In a small sample environment, the item information will vary substantially between samples (I would say wildly). However, the strongest items, as well as the weakest, tend to keep their relative importance from sample to sample. When the main interest is in a GRM (see Appendix B for a description of the model), with a common set of thresholds, the interpretation of total and item information is not straightforward. The a_i:s in a GRM still reflect the items' informative value, but not in a simple way. But conceptually, strong and weak items might still be identified by this approach, as they tend to keep their relative importance. This can be illustrated by plotting a_i(GRM) vs a_i(unconstrained).
We have to take care of two situations in evaluating the item information:

1. The GRM is considered sufficient (it cannot be statistically rejected).
2. The GRM appears not be sufficient (strongly rejected in favour of an unconstrained model).

For item i in a GRM, I suggest $I_k/\sum I_k$, k=1,…,K, as a rough measure of the relative item information in spite of it's 'quasi correct' calculation.

The problem of non-response

When, for some respondents, there is a 'non response' to a particular item, this has to be taken care of in the analysis. At a first glance, imputation of values might be considered. A simple and reasonable method is to look for colleagues with similar profiles as the participant with a non-response. The median or most frequent value in this set might then be imputed for the non-response. Such a procedure can be refined by an iteration procedure, in case of a set of 'non response'. A disadvantage of this method is a bias towards a more homogenous sample (which means a more favourable sample) than could be expected from a complete sample. A more complicated situation arises when the non-responses are not due to missing at random but rather that the question is interpreted as irrelevant for the respondent, or she/he did not understand its meaning (something we seldom know). If we consider that the person has actually chosen ‘not to respond’, this should not be called a ‘missing value’, and thus, no value should be imputed. A striking example may be sited from Study II. One of the items is as follows: “How much does your father love you? “ If there is no father? There will be a non-response, and obviously, an imputation is not relevant. However, in exceptional cases, a single missing might be imputed for a simplification of the analysis.

If imputation is found relevant, a reasonable strategy would be to replace a missing value only when there is a strong scalability (correlation) in the item set. Otherwise only noise is added, without any valuable information. Most of the interest in this matter is about “how to replace missing values?”. Very little, if else to my knowledge, is discussed about the appropriateness of such an imputation. There is also a more theoretical reason for no imputation. The aim of this study is to evaluate the questionnaires. If models are used to impute data values, the questionnaire characteristics are already used – thus the evaluation of the questionnaire will be dependent on these imputed data.

The Rasch approach can handle incomplete data sets (permitting incomplete answer profiles), while the Mokken analysis requires complete answer profiles. Programs handling GRM with item specific discriminations usually require complete answer profiles. My general opinion is that a ‘non response’ should not be subject to imputation. This means that Step 1 and Step 3 use only complete answer profiles.

Empty categories

‘Empty categories’ are frequent in small studies, particularly when the items have an enlarged number of response alternatives. The Rasch and Mokken analyses accept empty cells, while many other IRT-programs don’t. Of cause, the estimates will be affected and less reliable in presence of empty cells. Collapsing categories might be considered, and is also recommended when estimation of parameters are of importance, i.e. in large studies. In a small sample situation, collapsing categories might be a step to investigate, and hopefully verify the stability of the evaluation procedure, rather than a procedure to find a suitable set of categories. Empty categories will probably be endorsed when a study is enlarged.

How large is a ‘small sample’?

This topic is addressed, at least to some extent, already in the introduction.

There is no consensus of what is a small sample. From a clinical trial’s perspective, an adequate sample size varies with the generalisation of the result. In a medical laboratory experiment using rats, n=10-20 is often sufficient. In an experimental clinical trial, where the object is to show that a treatment effect exists, n= 50-100 is usually sufficient. In a pragmatic clinical trial, investigating a treatment’s general effect on the target population, some hundreds will usually be OK. In an observational study, where no randomisation or experimental stratification is possible, n> 1000 is certainly required. In this setting, a questionnaire study is a clinical trial where the variable, causing the’ treatment effect’, is not measurable, but may be reached by a thoroughly constructed evaluation instrument. Treatment groups are then compared based on information from the questionnaire. In that sense we may identify the actual studies as ‘experimental clinical trials’. The great difference is that we, in the first place, have to evaluate the aggregated measure, such as the sum score, as a relevant outcome measure. If we then ask the IRT literature: ‘How large a sample?, we will get quite different answers. In the IRT literature, the problem is stated as ‘How large a calibration sample?’ De Ayala, 2009, gives some examples. For a questionnaire with 30 items and a four-category response scale, >250 subjects is recommended. A uniform distribution of respondents across the response categories is assumed, which is not likely to occur in practice. For the GRM he refers to a study with 25 items and with five response categories, saying that at least 500 subjects are needed. Bond & Fox, 2001, present an example with 26 six-category items, where n=372 is mentioned as a 'large sample'. Linacre [Smith EV and Smith R.M, 2004, ch. 11] recommends at least 10 subjects in each category for a rating scale type questionnaire. There are certainly a lot of reasons why there are many 'small sample studies' around, which is in conflict with the requirements from a statistical point of view.

In conclusion, it is easy to realise that with less than about 50 subjects, answering a 10 item questionnaire, the estimation procedure will be very hazardous due a troublesome number of empty or too sparsely scored categories. We have to realise that numerical problems might occur.

Let us conclude, quite arbitrarily, that a 50 – 150 subjects is considered a small sample.

Testing hypotheses, the use of p-values and Confidence Intervals (C.I.)

The general hypothesis in each step is: 'The stated model is a reasonable representative of the general structure of the questionnaire.' However, a formal test of such a complex hypothesis is not meaningful. A p-value, related to such a hypothesis, is difficult to interpret and is of minor interest as we are more concerned (or focused) on how each item contributes to the proposed aggregated measure as stated by the model. This means that p-values appear in batches connected to the set of items in the questionnaire. In the suggested '3 step strategy', which is an explorative investigation, the p-values, or the Z- statistics, should be used as indicators in combination with other 'messages' from the analyses. This means that a parsimonious model should not be abandoned due to just one or a few 'significant' p-values. They should rather lead to suggestions how to improve a questionnaire. As an example, the item AttDef2 in the evaluation of the Attention Deficit questionnaire in Study III is found to be a weak and questionable item in all the three steps. This does not mean that we should disprove the investigated models due to a single small p-value, but rather suggest a reformulation or a replacement of this particular item. However, alternatively, if there is a set of messages (such as a set of small p-values) contradicting an agreement with the stated model, a more elaborated model than the one under investigation might be an alternative. This means that the stated model, limited by some restrictions, is found not able to reasonably capture the structure of the questionnaire. The more elaborated model might give further information about the structure of the data but should not necessarily be considered the best choice, even if the simpler model is statistically rejected. Its message might rather guide a reformulation of the questionnaire.

In a small sample evaluation we should not make too much attention to p-values $0.01 < p < 0.05$ as there is a relatively large probability of a 'false significance' when, in a test, there is one p-value for each item. Furthermore, the tests are seldom strictly independent and the distributions of the test statistics behind the p-values may not be as expected, even if some tests, such as the t-test, are fairly robust.

As is earlier stated, the aim is to 'evaluate basic characteristics' of a questionnaire. This implies that there are probably both special and complex characteristics which will not be revealed. In that sense, they will be hidden behind a number of false negative significances ($p > 0.05$), due to the small sample size.

Confidence intervals are of more interest as they give pragmatic information about the actual estimates. A C.I. is usually presented as a symmetric interval around the actual estimate, saying that with a probability of 95%, the C.I. will cover the true parameter value in the target population. In a small sample situation, a bootstrapped C.I. is probably more reasonable (if feasible) as it does not rely on any prespecified distribution. Furthermore, it might well be non-symmetric. In most cases, a 90% C.I. is considered as only the lower or the upper bound is of interest, i.e. with 95% confidence.

The Bootstrapping procedure

Only a few sample estimators, such as the mean, have exact formulas to estimate their associated sampling variability. Both exact and approximate expressions for estimating variability depend on knowing or postulating specific properties of the sampled distribution.
A bootstrap estimation approach utilizes computer based methods to provide estimates and their confidence intervals without theoretical models, extensive mathematics, or restrictive assumptions about the structure of the sampled population. Results are potentially misleading if inappropriate assumptions or simplifications have been made. ‘When sampled values come from an asymmetric parent distribution, a bootstrap confidence interval usually reflects the asymmetry in the confidence limits. An approximate procedure based on the symmetric normal distribution would be expected to work less well under these circumstances’ [Selvin 1998].
There are no particular restrictions about the sample size when applying bootstrap methods. Many examples in the statistical literature show that the bootstrap approach works well, and often better than other methods, when applied on small samples. Selvin, 1998 ch. 5, demonstrates an example where the bootstrap procedure works very well with a sample of 30 units from a normal population, where the parameters are known and the sample estimates are calculated in the ordinary way. Davison & Hinkley, 1997, show that the method works even with small samples. They use n=12 and n=49 in two examples. Efron and Tibshirani (ch. 13), 1993, show that even with a sample size of n =7, the bootstrap works well. Manly (ch. 3), 1997, on the other hand, has a convincing example when it does not work.. Bootstrap methods should not necessarily be assumed to work with small samples. However, when estimating ‘not too complicated’ parameters, it usually works well even for small samples. However, the estimates have to be ‘smooth’ (a continuous first derivative). ‘Naive bootstrapping’ questionnaire data from a small sample is the only reasonably procedure, as we virtually do not know anything about the correlation structure of the items. This simply means that we are resampling the respondents answer profiles ‘as they are’, that is, treating them as single observations.

Basic characteristics of the bootstrapping approach

An empirical distribution is obtained from a sample, which is assumed to be representative for the population. This is a prerequisite for every statistical approach and statistical inference. All information is embedded in the sample values and their distribution. Provided the sample reasonably represents the distribution in the population, we may use the empirical distribution as if it were the distribution in the population. Thus, we can extract a number of new data sets by random resampling with replacement.
A sample statistic can be extracted from each of these samples. We then get an empirical distribution of this statistic, without any assumptions of a distribution function. However, there are mathematical requirements about the statistic, such as its existence as a parameter in the population distribution, a ‘smooth’ estimate, which in essence implies a continuous derivative. Direct calculation of the statistic, based on resampling individual observations, is often called ‘naïve bootstrapping’. A 90% C.I. is achieved from the distribution of the bootstrapped statistics as the interval percentile range 5% - 95%. This interval is in a way ‘quasi-independent’ of the bootstrap mean estimate. The bootstrap mean should be in agreement with the original sample estimate of the statistic. The bias of a bootstrap estimate is indicated by the deviance of the bootstrapped estimate from the original sample estimate. Bias correction is not straightforward when percentiles are considered. There are however methods for correction of a bias, but they should be avoided unless the bias is large and they might be questioned in

small samples (Davison & Hinkley, 1997). A comprehensible description of some bootstrap and bias correction methods can be found in Manly, 1998 ch.3. Our goal is not any exact estimates of the C.I.s but rather guidance for further conclusion about the role of an item or the utility of a scale. The percentile method is used in this work (referred to as 'the first percentile method' in Manly's book). Bias correction is considered just on a few occasions.

Restricted bootstrapping
A naive, unrestricted bootstrapping might exaggerate the variability in a small sample due to repeated sampling of a few extreme answer profiles in certain samples. Such extreme profiles are likely to be seen in questionnaires where there are only a few items. Just one or two misinterpreted items can make the sample estimate very odd. An example how this will affect the estimate is illustrated in study I. Such an overrepresentation can be moderated by a limitation of the number of times a person might appear in a sample – restricted resampling. The effect of restricted bootstrapping is not easy to investigate but the method might be considered. Another form of restricted bootstrapping, often called balanced/stratified bootstrapping, might be used, for example if the male/female proportion should be maintained in the resampling procedure. This procedure is used for comparison between subgroups.

Why not simulation?

It is sometimes suggested that the reliability of an IRT-model can be further investigated by simulating new samples from estimates based on the observed sample (it is included as an option in the Winsteps program). However, this should be avoided in a small sample setting because:

- The model is not well founded.
- We will be likely to produce answer profiles which are not observed, profiles which may not exist even if a large sample could be observed.

That is, we shall use our observed profiles and not force artificial answer profiles into the population.

The '3- steps' strategy

Based on the previous reasoning I suggest the following strategy, which is also earlier outlined in fig. 4. Before going into details about various parametric models I suggest an unconstrained (as few assumptions as possible) investigation of some basic ideas, underlying the construction of the questionnaire. This is the idea of step 1.

Step 1: The non parametric approach, the Mokken scalability analysis [Sijtma & Molenaar 2002]

The model: $T_n(Y) <= \sum Y_{ni}$, the raw sum score for person n, based on item i=1,2,…,I..

This analysis is focused on the relation between the items and the items' relation to the raw sum score. The analysis is carried out with reference to the basic sum score model $T_n(Y)$. The primary question is whether T_n can be used as a representative for person 'n' when ranking the individuals on the intended latent trait. It is then required that the items co-operate in the 'same direction', i.e. that they are positively correlated at three levels –pairwise scalability H_{ij} , item scalability H_i, and the 'over all' or total scalability H. The word 'scalability' (instead of correlation) is used as the correlation between ordered categorical data is somewhat different from what we usually perceive as a correlation coefficient. Let Y_i and Y_j denote the observed responses on items i and j. Let $R_{(i)}$ denote the sum of the responses, called the rest sum score, across all items except item i. On the person level, let Y_{ni} denote the response from person n on item i. The procedure is as follows:

H_{ij} = the pairwise scalability between the items i and j, given the data, i= 1, …, I, j= 1, …, J where i≠j.
H_{ij} = Cov(Y_i, Y_j) divided by Cov_{max} (Y_i, Y_j), given the marginals in the cross table.

H_i = the scalability of item i related to the rest of the items. All H_i :s should be >0 and preferably >0.3 to be contributing items. H_i = Cov(Y_i, $R_{(i)}$) divided by Cov_{max} (Y_i, $R_{(i)}$), given the marginals.

H = the scalability of the total set of items. H = ∑ Cov(Y_i, $R_{(i)}$) divided by ∑Cov_{max} (Y_i, $R_{(i)}$), where i= 1,…, I, the number of items.

H is used to judge the overall quality of the RSS to represent the information from the questionnaire. $0.3<H <0.4$ indicates a weak but reasonable set of informative items, $0.4\leq H <0.5$ a moderate set an $H \geq 0.5$ a strong set of items. If the scalability is found not sufficient, $0< H <0.3$, the sum score approach might be rolled out as not suitable, saying that the questionnaire produce little or negligible information (Sijtsma & Molenaar, 2002). A $H_i <0$ is considered counterproductive and indicates deletion of the item.

Furthermore, to get a sum score to work, which means 'ordering the persons on the latent trait according to their personal total sum score', a non-parametric MHM (Monotone Homogeneity Model) is required. MHM – a measurement model to order persons on a latent trait.
Three assumptions constitute such a model:

1. *Unidimensionality:* All items represent the same underlying latent trait. This is virtually impossible to achieve in practice. We have to accept that it means one dominating dimension, contaminated with minor 'negligible' dimensions. That is, all H_{ij} :s ought to be positive, a few negative (but in the neighbourhood of zero) can be tolerated.
2. *Monotonicity*: The IRFs are monotonically related to the latent trait.

Formaly: For an arbitrary fixed value of θ, say c, $P(\theta > c| S_1) < P(\theta > c| S_2)$ for all person sum scores $S_1 < S_2$.
This signifies in essence a non-negative relationship between responses Y_i to an item and the rest score $R_{(i)}$. This is verified by the rest score method (Appendix C). Persons with low $R_{(i)}$:s are expected to score low at item i, while persons with high $R_{(i)}$:s are expected to score high on item i. This relationship is investigated for each item.

3. *Local independence*: This is, as defined on page 15, a basic assumption, but is not easy to verify in a non parametric setting, as a person's position on a latent trait construction is not estimated.

A desirable feature of the set of items is 'non intersection', or 'Invariant Item Ordering', IIO. An IIO implies that the item ordering is the same across subgroups along the scale, i.e. the ordering of item locations is the same for low, medium and high scoring persons. As no item locations are estimated in the non-parametric approach, the rest score method is used. The relation between Y_i and Y_j are compared to the rest score $R_{(ij)}$, i.e. the total sum score without items i and j.
An IIO is an important property when individuals or groups are compared but is not required when the only purpose is to rank the individuals. Typically, items with low scalability have relatively flat item response functions and tend to obstruct the IIO. Alarmingly many violations against non-intersection might easily occur in a small data set, as the rest groups are relatively few and (by necessity) of small sizes. Rarely endorsed categories as well as too many categories might also be destabilizing factors. By varying the minimum rest group sizes within a reasonable range, we can get a hint of which item(s) will consistently show many violations. IIO is a desirable condition of Step2 and Step3.

The scalability analysis is a fundamental step in the three step strategy. The estimates of the scalabilities (H and Hi :s) constitute a basis for decisions about the instrument. These estimates are subject to variation which might incidentally influence a decision. An analysis of the sampling variation is therefore of vital importance, especially for items found suspicious by the non-parametric analysis in this first step, in order to not suggest exclusion of potentially contributing items.
We therefore need reliable confidence intervals. Conventional methods for establishing confidence intervals require distributional assumptions, which are not available in this case. Therefore boot-strapping methods, based on the actual empirical distribution, are more suitable and have also been shown to work on small samples [Davison & Hinkley, 1997]. Whether exclusion of an item will significantly improve the H can be addressed by a test of the difference in H(all items) vs. H(one item excluded). An empirical distribution of the difference is achieved by resampling a large number of samples and calculating the difference in scalability for each sample. A C.I. or a direct test can then be constructed [Davison & Hinkley, 1997]. The percentile bootstrapping approach is used for creating confidence intervals. A large variability of bootstrapped scalabilities can be expected when there is a weak scale, $H< 0.4$.

Investigation of unidimensionality
In the CTT regression analysis, it is recommended to start with the complete set of variables (items in IRT). By analysis of the residuals, where all information is gathered, variables are rejected as long as they do not contribute to the model more than is expected besides the random variation. This is not the case for a non-parametric approach where there are no residuals. In contrast, the 'bottom-up strategy' is recommended [Sijtsma & Molenaar, 2002]. A small set of items are selected, by the researcher, or on mathematical grounds, as a start. Items are then selected one by one and added to the already selected

set. This procedure is mostly carried out to verify that the sample of persons and the designer of the questionnaire agree concerning the unidimensionality. For further details, see Appendix B.

Investigation of influential answer profiles
Influential answer profiles are of particular concern in a small sample study. Such profiles may deteriorate an investigation of a good questionnaire by decreasing the scalability estimates and falsely yield bad fit statistics when parametric models are tried. However, in case of a weak questionnaire (weak scale) such a person is likely to not cause much harm as it will be embedded in the heterogeneity of incoherent profiles. When there are few items in a questionnaire, there is little 'room' for a deviant profile as the number of possible profiles is very limited. On the other hand, floor and ceiling answer profiles tend to over-estimate the scalabilities in a questionnaire with few items.
In Step 1, influential persons are identified by using the 'jackknife' method [Efron Bradley, Tibshirani, 1993]. Leaving out one person at a time yields scalability estimates, $H_{(n)}$ where (n) denotes deletion of person n, n=1,2,…,N, when the item set scalability is calculated. The procedure is carried out on the analogy of the jackknife procedure for simple and smooth statistics.
Influential persons can be directly identified by the corresponding jackknife value $H_{(n)}$ by simply looking at top (or lower) end of the jackknife distribution. The search for influential individuals by this method is not perfect but if such individuals are found already in Step 1, they are likely to be 'disturbing' whatever parametric model is used (as in Step 2 or 3).

Investigation of weak items in terms of scalability
The scalabilities reveal much of the items' relevance and importance in a questionnaire. Values around or below 0.3 [Sijtsma & Molenaar,2002] will raise the question of a recommendation of deletion or at least reformulation of an item. Therefore, the precision of the estimated scalability has to be addressed. If a H_i <0.3 while a reliable 90% C.I. covers the limit 0.3, it would be too risky to point out the item as 'not contributing'.
The investigation of item 'i' should be performed without any assumed distribution of the answer profiles. Three approaches are available:

1. The confidence interval approach:
A C.I. can be achieved by resampling observed answer profiles (bootstrapping).
Restricted resampling might be considered.

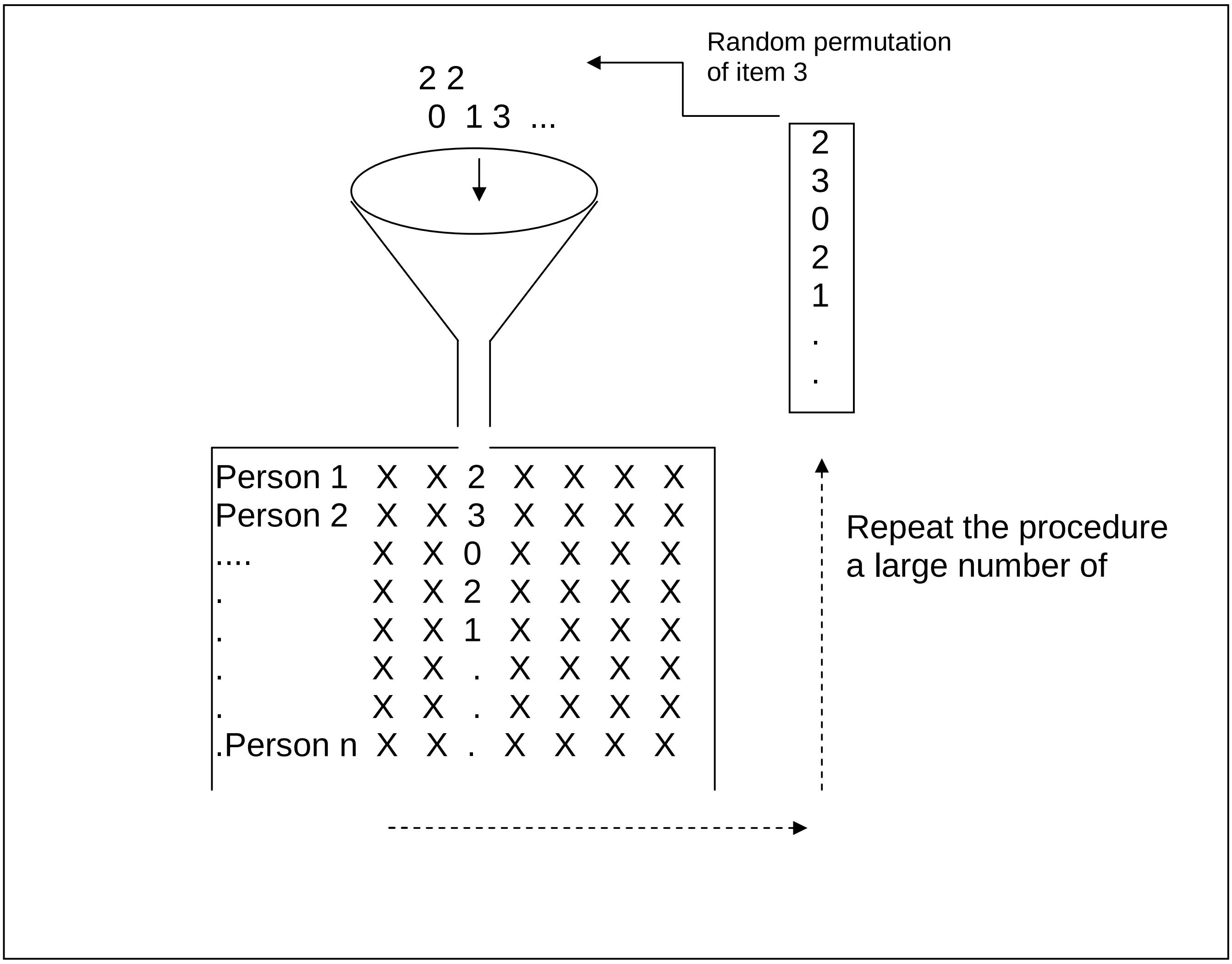
2. The testing approach:
A low H_i can be investigated by hypothesis testing. Is the observed H_i reasonable under the null hypothesis H_0 : $H_i = 0$? H_0 is evaluated by random permutation of item 'i', while keeping the rest of the answer matrix as observed. The permutation procedure does not automatically meet H_0 , but the bias can easily be corrected. The procedure is illustrated in fig. 5.
When we randomly put back the 'set of observed categories', in this case item 3 in figure 5, we do not expect any correlation with the rest of the items. Repeating this procedure, we get a distribution of H_3 based on no correlation, i.e. H_3= 0. We can then investigate the actually observed H_3 and judge whether it is a reasonable member of the generated distribution.

3. The 'utility' approach:
This approach is a direct test of the utility of an item.
Does item 'i' (in case of a low H_i) significantly deteriorate the overall H?

The null hypothesis H_0 : $H_{(i)} = H$, where $H_{(i)}$ is the item set scalability without item i. Only the one sided test is of interest. The test is performed by resampling (bootstrapping) a large number of samples. $H_{(i)}$ and H are then calculated pairwise, based on samples in common. A test statistic, $D^s{}_i = [H - H_{(i)}]^s$ based on sample s, s=1, 2, …, S yields an empirical distribution of the effect of deleting item i.
In case of more than one questionable item, one at a time, in a stepwise manner, should be evaluated.

Fig. 5. Random permutation of item answer profiles.

Permutation of an item: Random replacement of observed categories



Strong item pair scalabilities

Even if good item pair scalabilities are desirable, strong positive scalabilities might conceal an item dependence in addition to what can be regressed to the person level. This problem can, at least to some extent, be clarified in Step 2 and 3 by an investigation of residual correlations.

Differences in H and H_i between subgroups (Different Item Functioning)

Although the focus is on small samples, there might be large differences in scalabilities between subgroups, such as male/female, patients/controls etc. The possibility of a systematic difference might be investigated by the difference in scalability, such as H(male,i) – H(female,i) for item i. A C.I. for the difference, by using stratified bootstrapping will guide to further conclusions. A systematic difference in item scalability between groups has no straightforward interpretation but says something about how the items are perceived in the sense of the items' capacity to capture the intended underlying trait. The questionnaire might be more or less suitable for specified subgroups. This raises the question about subgroup specific items. The problem is similar to 'different item functioning' when parametric models are considered. The same can be said about the item set scalability, H. If there is a general difference in scalabilities over the item set, stratified bootstrapping for investigation of the difference H(subgroup 1) - H(subgroup 2) can be applied.

Questionnaires divided in specific parts to catch more than one dimension

In some cases, the questionnaire is set up with more than one latent dimension in mind. The authors of such a questionnaire have specified in advance which items belong to which dimension. Even if such a specification is well founded, the respondents might perceive certain items quite differently.

In the Mokken non parametric analysis, there is a possibility to investigate the dimensionality, under very few restrictions, by use of item scalabilities. See Appendix B and the 'Investigation of unidimensionality on page 22. However, in a small sample, the result might depend on the starting point as well as on a few incoherent answer profiles. This method might yield some information, but we cannot expect it to be very efficient. Usually, already identified weak items tend to fall outside an identified first dimension. In case of more than one dimension identified, weak items tend to move back and forth due to just slight changes in the sample.

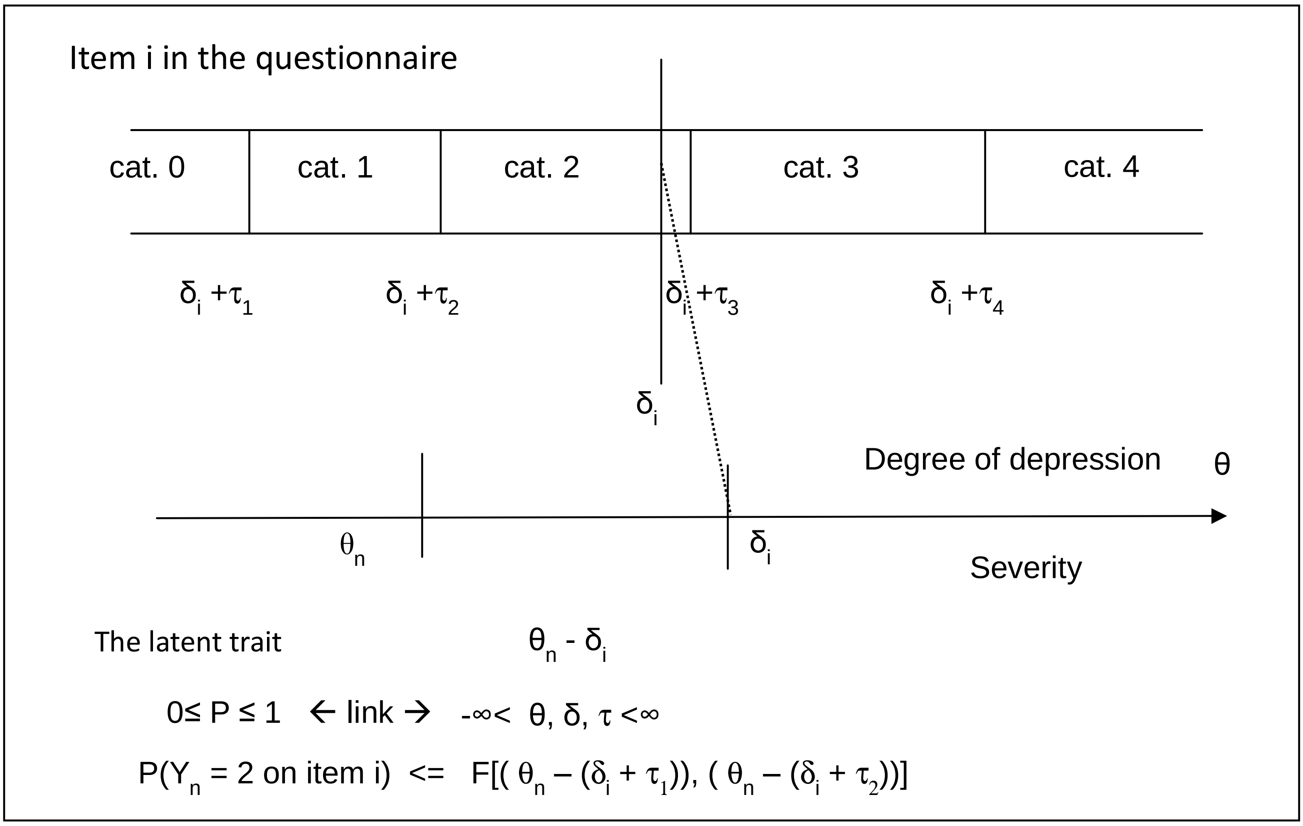
Step 2. A basic IRT model for ordered categorical response data

This step moves the evaluation from ordering the subjects to investigate the capability of the questionnaire to find a parsimonious estimation of an interval scaled value for each subject.

'Rasch measurement is the only way to convert ordinal observations into linear measures' [Fischer 1995, according to Smith & Smith, 2004]. This implies a logistic transformation, where a person's position on a latent trait is related to a probability. This will be further explained as follows:

The following descriptions relates to a questionnaire as outlined in table 2.
Consider an item, i, where the ordered response alternatives are coded k= 1,2, ..,m, which is thought to reflect the degree of severity on an underlying trait. An individual's position on the latent trait scale is directly related to the log odds of answering at different levels of the item, and thereby also to a probability statement. The higher the position, the larger the probability to answer on high levels in a positively ordered item set. A respondent n, with an unknown measure θ_n on the dimension, is thought to generate an answer, Y_n, in category k according to the probability $P(Y_n = k$ on item i) as a function of θ_n and the location δ_i of the item 'i' on the latent trait. The threshold for answering in category k of item i is represented by $\delta_i + \tau_k$, where τ_k is a distance from the location of item i to the actual threshold**,** see fig. 6**.**

Fig.6 . Item i with location δ_i and a set of thresholds = $\delta_i + \tau_1, \tau_2, \tau_3, \tau_4$



In fig.6, the latent scale can be thought of as lying behind the category scale. If the person's status of depression, θ_n, is close to the item location δ_i, we can expect an answer in category 2 or 3. If the status of depression is much lower than the item location, as is indicated in the figure, an answer in category 0 or 1 is more plausible. Thus, the probability of answering in a particular category depends on the persons distance from the item location, $\theta_n - (\delta_i + \tau_2)$. We also realise that this is a stochastic process. θ_n as it is indicated in figure 6 does not automatically generate an answer in category 0 or 1. An answer in category 3 or 4 is possible but unlikely. Thus, a probability function as a link, or correspondence, between the ordered categorical scale and the constructed interval scale is a suitable choice. Saying that $P(Y_n = 1$ on item i, given $\theta_n) \geq P(Y_n = 3$ on item i, given $\theta_n)$ now becomes a natural statement, which also can be expressed in mathematical terms.

We then get the relation $P(Y_n = k$ on item i$) <= F(\theta_n, \delta_i$ and τ_j, j=1..k). As all these measures are placed on a common construction of an interval scaled dimension, representing the underlying trait, we are able to formulate F as a function transforming values from the unbounded latent dimension, $[-\infty, +\infty]$ onto a probability, $[0 \leq P \leq 1]$. See Appendix B. The most suitable formulation of such a transformation is by $F= z/(1+z)$, where $0 \leq F \leq 1$ and $-\infty \leq z \leq +\infty$.
The formulation $z= \exp(\theta_n - (\delta_i + \tau_j))$ is for mathematical convenience and relates a person's measure directly to the location of item i (sometimes called item difficulty). z is interpreted as a transformation of the distance between the position on the latent trait for person θ_n and the position of the category threshold $(\delta_i + \tau_j)$. When a person's position is far below the threshold F approaches zero, when it is on the same position as the threshold F approaches ½ and when it is far above the threshold F approaches 1. The set of distances from an item location to the item m-1 category thresholds are chosen to be the same for all items and $\sum\tau_{j,\ j=1,\ldots,m-1}$ is set =0. This corresponds to the structure of a questionnaire with all items having the same wording of the categories as well as the same number of categories.
The estimation of the parameters is based on the likelihood between the model and what is actually observed, i.e. the answer matrix. Restrictions and assumptions, together with a crosswise estimation procedure between person and item parameters, are used to solve the problem. This is accomplished by computer programs, aimed at IRT models, with somewhat different estimation procedures within the framework of maximum likelihood methods.

The agreement of the Rasch estimates and an intended idea, where the increasing difficulty of the item steps corresponds to what is actually observed, can be evaluated from the output from the Winsteps program. "Disordering of these estimates (so that they do not ascend in value up the rating scale), sometimes called "disordered deltas", indicates that the category is relatively rarely observed, i.e., occupies a narrow interval on the latent variable, and so may indicate substantive problems with the rating (or partial credit) scale category definitions" (Linacre). With a small sample in mind we should not take moderate violations against a proper ordering too seriously. Categories too close to each other in combination with sparse data might well cause these violations.

The basic Rasch Rating Scale Model, RSM

The basic Rasch model postulates the person raw sum score as a sufficient statistic for the person (all relevant information is collected in the raw sum score), a model where all items are equally informative and a common structure for the categories within item.

The Rasch formulation of P(Person n answers in category k on item i) is based on model (1) in Apendix B.

A reasonable result from step 1 is a favourable condition for a parametric IRT- model to be informative. As all items, in this type of questionnaires, have a common set of ordered response alternatives, a Rating Scale Model (RSM) is the most suitable. Such a model is convenient for small studies as the number of parameters is at a minimum. In its most parsimonious formulation, the number of parameters is relatively low. All items are postulated to have equal discriminations (weights), i.e. equal ability to discriminate between persons in the vicinity of δ_i. They also share a common set of category thresholds, τ_j, j=1, …, m-1, and m categories, with the restriction $\Sigma\tau_j = 0$. However, each item has its own location δ_i , i=1,2, …, I. Each person is assigned a measure θ_n, n=1,2, …, N, where N is the number of respondents.

The Rasch model is based on successive transactions between categories, a procedure which means stepwise decisions along the ordered categories of an item. To answer in category k of item i, the respondent has to 'pass' k lower categories, i.e. pass k thresholds (category boundaries in the Rasch vocabulary). Then P(Y_n = k on item i) is built on a transformation of $\sum(\theta_n - (\delta_i + \tau_j)$, j=1,..,k. The Rasch model assumes that an individual's sum score is a 'sufficient statistic', which means that all information from the answer profile is housed in the sum score. This implies that persons with different answer profiles, but with the same sum score, get exactly the same Rasch person measure. This is a strong restriction, but if approximately fulfilled, the model has a number of advantages. Incomplete answer profiles and subgroup specific items can conveniently be handled by the Rasch approach. However, the probability model, as well as the estimation procedure, is somewhat different from those of other IRT models.

If the model is found to approximately represent (fit) the data, the respondents' sum scores are readily transformed to an interval scaled variable, with a specific value on the latent trait. However, both the person and the item reliability should be > 0.9. Items with bad fit (infit MNSQ> 1.5) are candidates for being reformulated. MNSQ is the chi-square statistic divided by its degrees of freedom and serves as an indicator of 'the value' of the item in the Rasch modelling procedure. The built-in help in the Winsteps program provides good explanation much advice. Items with very low MNSQ, indicated by a negative z-score, might be questioned due to dependence on other items, or redundancy. They are not of any immediate concern unless a shortening of the questionnaire is of interest. The unidimensionality and the fraction of explained variance can be inspected.

Bad fitting items

The most poorly fitting item(s) should be put aside and a new estimation procedure should be performed. Although the items are considered locally independent, they share a common set of category thresholds and are of equal weight. These restrictions imply that a 'bad fitting' item forces the other items to an unrealistic compromise, with a possibly misleading message as a consequence. A new estimation, with the actual item deleted will give more adequate information about the questionnaire. In case of more than one bad fitting item, exclusion of one item at a time will be a cautious procedure. Such an item may deteriorate the relation between the questionnaire and the Rasch model. An MNSQ>

1.5 indicates that the item contributes more noise than valuable information. An MNSQ>2.0 ‘degrades measurement’ [Linacre J., Winsteps 3.66, 2008]. Such an item has to be excluded in the first place. Further ‘pragmatic’ information can be extracted from the ‘matching %’ column in the ‘model table’ in the computer output from the Winsteps program. Obs% ≈ Exp% for an item indicates random variation about what is expected and Obs% > Exp% indicates more random variation than expected. Obs% < Exp% indicates less random variation than expected. This indicates that the answer of the actual item is predictable and might be an indication of redundancy.

Item discrimination

Although no discriminating coefficients (item weights) are included in the Rasch approach, the Winsteps program offers the possibility of their estimation, however outside the modelling procedure, i.e. after the model fitting is completed.

An estimated item discrimination <0.5 indicates a weak ability to discriminate between persons on the latent trait. A low discrimination is usually connected with a high MNSQ. Removal of a ‘disturbing’ item may bring the rest of the items set into a better agreement in constituting the model. A removal may also lead to a positive change in person and item reliability. However, such an item is probably detected already in step 1 A high discriminating coefficient (>1.5) is a further indication against ‘equal weights’ and points to step 3.

‘Outliers’ in terms of incoherent answer profiles

Before any drastic action is carried out in the item set, poorly fitting persons should be identified. This is in contrast to what is usually recommended, but the reason is that in a small sample situation, single persons might have a large influence on the evaluation. Persons, lacking in concentration, not having understood the questions or thinking about something else when scoring, might destroy the process and give misleading results. This is of particular concern as our analysis is based on a small sample. These persons are identified by a difference between their answer profile and what is expected from the model. This difference can be turned into a fit statistic similar to those for the items, or identified graphically. Even if a fundamental principle in statistics says that no person should be excluded if they belong to the target population, these persons should be put aside in the evaluation procedure to investigate whether they distort or conceal the structure of the questionnaire. Our primary aim is to investigate the questionnaire. Only noticeably ‘outlying’ persons should be excluded. Usually they are just a few. If they represent a substantial part of the sample, the problem probably might be due to the formulation of the questionnaire, the composition of the sample or the conditions, under which the questionnaires are answered.

Different Item Functioning

There might be a pronounced difference between groups how they perceive the questionnaire. This is not an easy task in small studies, but obvious subgroups, like gender or groups with different degrees of a disease, might perceive the questionnaire differently. This indicates that the subgroups need to be judged separately. A possible Different Item Functioning, DIF, has to be investigated. In case of a marked DIF, questions might be reformulated into a ‘subgroup neutral formulation’. There is also a possibility, within the framework of Rasch models, to set up ‘group specific’ questions to solve the problem. This is a task when our purpose is to keep the questionnaire valid over subgroups.

Item specific thresholds
It might be the case that the basic Rasch model, with the common set of thresholds, is not able to house all the information from the questionnaire. Then, the common set of thresholds might be released so that items, or groups of items, get their own set of thresholds, but this will be at the cost of an increased number of parameters. This might be questioned due the small sample size but it can sometimes explain a bad fit or any other anomaly of particular items.

Item and person reliability
Item and person reliability is a concept primarily connected to the Rasch approach. The person reliability is measuring the capability of a questionnaire to reasonably separate the respondents in subgroups on the latent trait. Or, as stated in [Bond & Fox, 2001], "The person reliability index indicates the replicability of person ordering we could expect if this sample of persons was given another set of items measuring the same construct [Wright & Masters, 1982]." To achieve a high person reliability we need a sample with a large person measure range and/or an instrument with many items.
Item reliability is telling the relevance of the item set and whether the actual sample has the capacity to reasonably locate the items on the constructed scale. It also "indicates the replicability of item placement along the pathway if these same items were given to another sample with comparable ability levels" [Bond & Fox, 2001]. To achieve a high item reliability we need a questionnaire with a large item location range.
'A high degree of reliability is necessary for validity' (de Ayala, 2009). It is hard to find any recommendation in the literature but I suggest a reliability ≥0.8 as a reasonable lower limit for the use of a parametric model.

Good item reliability but poor person reliability
Even with a relevant set of items, a model's ability to catch the respondents answer profile might be insufficient. There are mainly two reasons for this lack of efficiency.

- There are too few items in the questionnaire to reasonably estimate a person's position on the latent trait.
- The item locations are not sufficiently distributed over the range of person measures. I.e. the coverage is usually too narrow or poorly centred compared to the person measures. Further items, with a wider distribution, might be a remedy.

Step 3. The extended model

The person answer profile is now a sufficient statistic rather than the raw sum score. This step also evaluates whether an item weighted approach is necessary. Thus, the item information, which in Step 2 is assumed to be equal for all items, is further investigated. It can be argued that such an extended model is 'too much' for a limited sample of subjects, but valuable information can be gained and bring some light over the earlier two steps.

The model: P(Person n answers in category k on item i) <= F[α_i(θ_n – (δ_i + τ_k))].

Step 1 and Step2 were focused on symmetric models corresponding to the 'symmetric layout' of the questionnaire, with a constant discrimination coefficient for all items as well as a common set of category thresholds. Even if there is a common set of category labels (the same wording for every item), we cannot be sure that the respondents will perceive the distances between the categories homogenously over items. If they are not sufficient with the symmetric approach, we have to extend the modelling in order to understand how they are interpreting the questionnaire.
If we relax the assumption of equal item discrimination and introduce item specific weights, α_i , we get a more flexible model (and leave the Rasch approach). It is then more convenient to model the probability as 'answering in category *k or higher*', from which probabilities of scoring in the specified categories can be calculated. P($Y_n \geq$ k on item i) <= F[α_i (θ_n – δ_{ik})]. This approach was suggested by Samejima [Samejima, 1969)] and is called the Graded Response Model. Muraki [Muraki, 1990] facilitated the use of such a model when analysing rating scale type of questionnaires by splitting the general category threshold into a 'centre of gravity (the item difficulty) and a set of distances from the centre to the category thresholds (as already introduced in Step 2). We then get a GRM of the form P($Y_n \geq$ k on item i) <= F[α_i (θ_n – (δ_i + τ_k)], a formulation analogous to the Rasch model. As an overlap from step 2, a GRM with a constant α may be investigated, P($Y_n \geq$ k on item i) <= F[α (θ_n – (δ_i + τ_k)]. This approach is similar to the Rasch model, but the estimation procedure is different. In a way, it can be used to verify the Rasch model 'from outside'. Person estimates from this model are usually similar to those from the Rasch RSM, apart from a scale factor. However, persons with a common sum score generally get different estimates. The similarity can be inspected by a simple graph.
GRM, as well as RSM, can house item specific thresholds, τ_{ik} , the threshold k for item i, but this will, as earlier pointed out, substantially increase the number of parameters. As this project is focused on studies based on small samples, models with item specific thresholds should just occasionally be considered. They are mainly aimed for getting an enhanced insight about the quality of the more parsimonious models. However, sometimes such a model appears inevitable.

'Outliers' in terms of incoherent answer profiles
This investigation is not as straightforward as in step 2, where persons are 'grouped' according to their sum score and can be compared to what is an expected answer for 'the group'. In principle, a GRM gives each person a unique estimate as long as her/his answer profile is unique. Of course there are situations where different profiles yield the same estimate. This may occur for respondents where the answer profiles are the same besides just a few items, which are similar in location and discrimination. Incoherent answer profiles may be identified graphically by plotting the profile, as a y-variable, against the item locations. Ideally, the plot will show a decreasing trend along the item location axis.

The disadvantages of a common set of item thresholds

The parsimonious approach of a common set of thresholds, $\sum\tau_k$ =0, is a restriction which is taken at a cost of lost information. We get no information about a possible interaction between items and response alternatives. In a small sample setting, the possibility of detecting important interactions is very limited. However, it can be investigated to some extent by applying common sets of τ_k for subgroup of items.

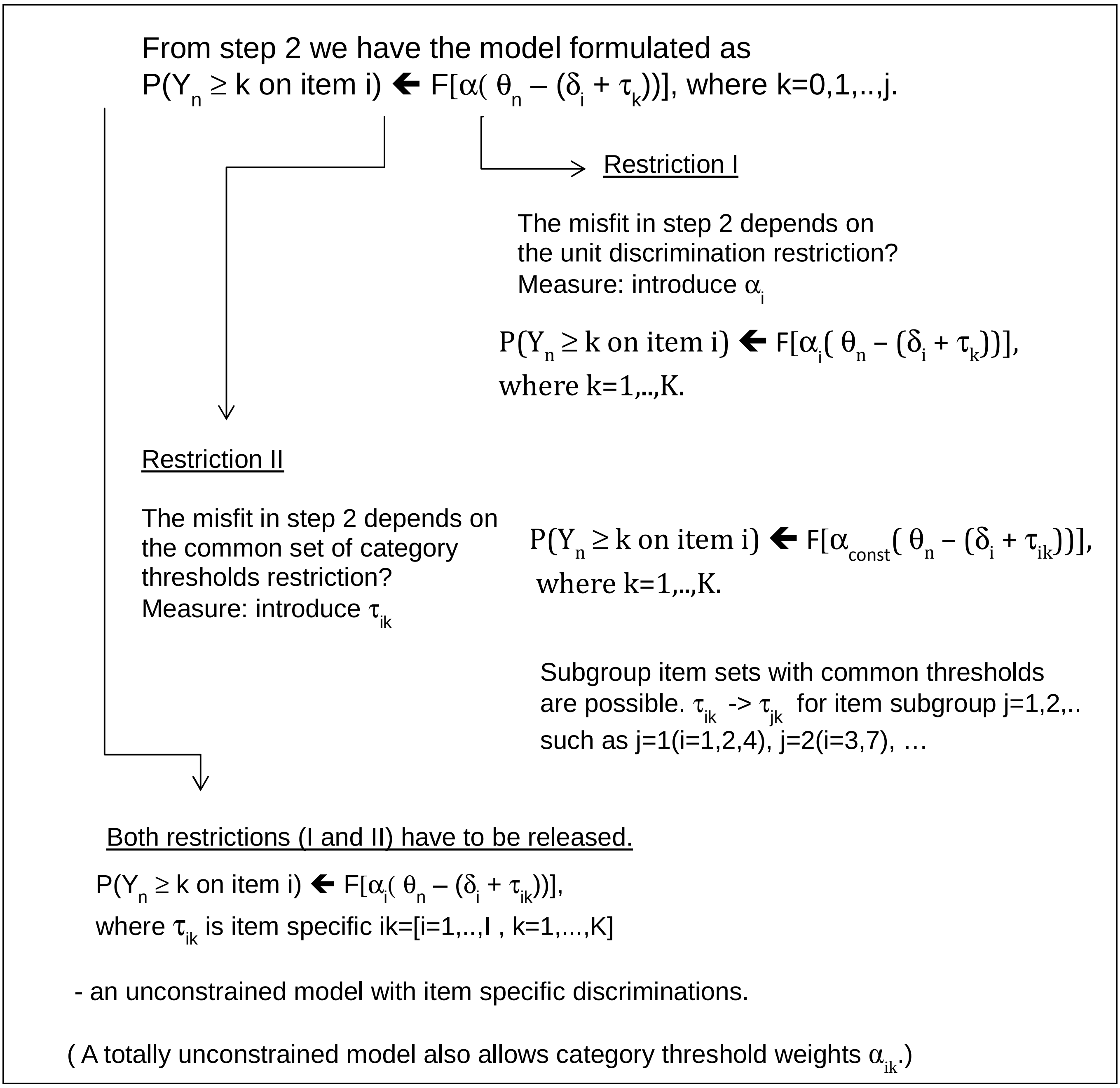
There is a relation between the item specific discriminations (often named slopes in the literature) and the choice of the set of category thresholds. If we impose the constraint $\tau_{ik} = \tau_k$, k=1,…,K, (K+1 categories) a part of the available information is moved from the estimation of category boundaries to estimation of slopes, i.e. the slopes ‘try to compensate’ for the constraint on the category boundaries, leading to an increased information variability. If the constraint is imposed on the slopes only, these estimates will work on their own and the information variation becomes more moderate.

Freeing the restriction of a common set of thresholds, τ_k , in favour of τ_{ik} , i.e. item specific thresholds, releases the available information and a more plausible variation of the discrimination coefficients can be inspected. The basic function is then constituted by $F[\alpha_i (\theta_n – (\delta_i + \tau_{ik})]$. Although far too many parameters to estimate, this procedure is a pragmatic way of evaluating the necessity of different α_i:s. In such a way, the formulation $F[\alpha_i (\theta_n – (\delta_i + \tau_{ik})]$ can be directly compared to $F[\alpha_{const} (\theta_n – (\delta_i + \tau_{ik})]$ by a likelihood ratio test, evaluating the systematic gain by moving from α_{const} to α_i .

Such a test will not be fully reliable, due to the small sample size and a possible dependence between items, which affect the degrees of freedom. However, it might well serve as an indication concerning item specific discriminations. If a common set of thresholds, τ_k, is forced on the most poorly adapted item for such a model, this item’s discrimination value is be expected to be greatly reduced.

We then get the following layout for the strategy in Step 3, see figure 7:

Fig. 7. The layout of step 3



From step 2 we have the model formulated as
$P(Y_n \geq k \text{ on item } i) \Leftarrow F[\alpha(\theta_n - (\delta_i + \tau_k))]$, where k=0,1,..,j.

Restriction I

The misfit in step 2 depends on the unit discrimination restriction? Measure: introduce α_i

$P(Y_n \geq k \text{ on item } i) \Leftarrow F[\alpha_i(\theta_n - (\delta_i + \tau_k))]$,
where k=1,..,K.

Restriction II

The misfit in step 2 depends on the common set of category thresholds restriction? Measure: introduce τ_{ik}

$P(Y_n \geq k \text{ on item } i) \Leftarrow F[\alpha_{const}(\theta_n - (\delta_i + \tau_{ik}))]$,
where k=1,..,K.

Subgroup item sets with common thresholds are possible. τ_{ik} -> τ_{jk} for item subgroup j=1,2,.. such as j=1(i=1,2,4), j=2(i=3,7), …

Both restrictions (I and II) have to be released.

$P(Y_n \geq k \text{ on item } i) \Leftarrow F[\alpha_i(\theta_n - (\delta_i + \tau_{ik}))]$,
where τ_{ik} is item specific ik=[i=1,..,I , k=1,...,K]

- an unconstrained model with item specific discriminations.

(A totally unconstrained model also allows category threshold weights α_{ik}.)

A pragmatic inspection of item information

In a model with item specific category thresholds and item specific discriminations (slopes), the overall information is the sum of the items' information. The relative item information can then be estimated. The situation becomes more complex with a restricted model but, as earlier is pointed out, $I_k/\sum I_k$, k=1,…,K, might be useful as rough approximations, although the I_k:s are not independent.

If a likelihood ratio test strongly rejects the null hypothesis, H_0: $\alpha_i = \alpha_{const}$ for i=1,…,I, given item specific thresholds, an unconstrained model should be consulted for the items' relative information.

The five studies performed

The studies, as presented in the publications, are rather restricted and shortened according to restrictions in space, the number of tables and the possibility to explain the procedure in detail. Due to these circumstances, the studies, as presented in this work, are much more detailed and many aspects, not found in the articles, are presented in order to get a more complete idea of how the ‘3 step strategy’ works. As an example, only Step 2 is presented in the published version of Study II. This is due a combination of a pragmatic investigation of questionnaire and a presentation of a reasonable strategy for such an investigation. As the studies are published in journals aimed for the practising clinicians, the investigation of the questionnaire is the main objective. The statistical aspects are forced to be expressed in fairly short terms.

The numbering of tables and figures is organised as follows:

Table (Fig.) (Study, Dimension (the intended latent trait), Step, No. within step)

As an example: Table I.AA.1.1 means
Study I, ‘Activating the Application of knowledge’ dimension, Step 1, Table 1.

Study I

Ulf Brodin, Uno Fors, Klara B. Laksov. **The application of Item Response Theory on a teaching strategy profile questionnaire.** BMC Biomedical Education 2010, 10:14

The study behind this article was to generate thoughts about one's teaching practice by revealing tendencies in a teacher's teaching practice towards the 'activating' of the application of knowledge (AA), meaning of knowledge (AM) and reproduction of knowledge (AR). There are three items for each dimension. The responses on a 5 point Likert type scale were summed up and the intention was to use the sum score as a representative for the teacher's attitude. The outcome variables were intended to work as described in 'Background' in [Brodin, Fors,Bolander, 2010]
The aim of this study was to investigate the characteristics of the questionnaire by the '3 step strategy' and evaluate whether the intentions behind the questionnaire could be reasonably fulfilled. Is the intention of the use of the sum score, based on just three items, relevant or even possible?

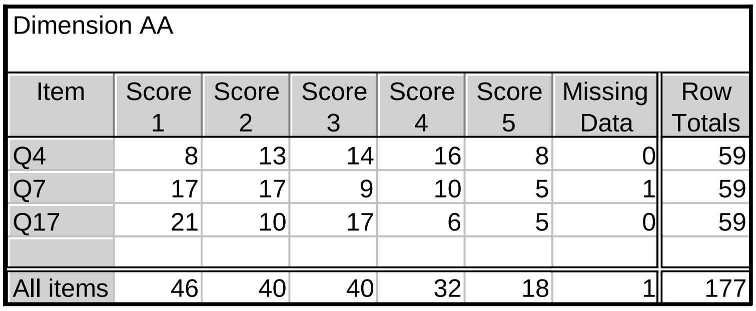
With a small sample, n=59, and just three items per dimension, a large variability of estimated bootstrap scalabilities was seen, even when a restricted bootstrapping procedure was considered.
The analyses of the three intended dimension are based on 27 females and 31 males. Gender is missing for one person.

Investigation of Activating the Application of knowledge (AA)

Step 1 AA. The Mokken scale analysis

The answers from the 59 teachers are presented in table Table I.AA.1.1. All category levels are fairly well represented and there was just one non-response. The items Q4, Q7 and Q17 are described in Appendix A.

Table I.AA.1.1. Descriptive statistics of AA,



Dimension AA

Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
Q4	8	13	14	16	8	0	59
Q7	17	17	9	10	5	1	59
Q17	21	10	17	6	5	0	59
All items	46	40	40	32	18	1	177

I made an exception from the 'non imputation principle' as there was just one non-response. The imputed value was calculated by the 'hot deck' method.

Table I.AA.1.2

H_{ij} item pair scalabilities, n=59

```
       Q4    Q7   Q17
Q4  1.000 0.264 0.381
Q7  0.264 1.000 0.510
Q17 0.381 0.510 1.000
```

H_i item scalabilities

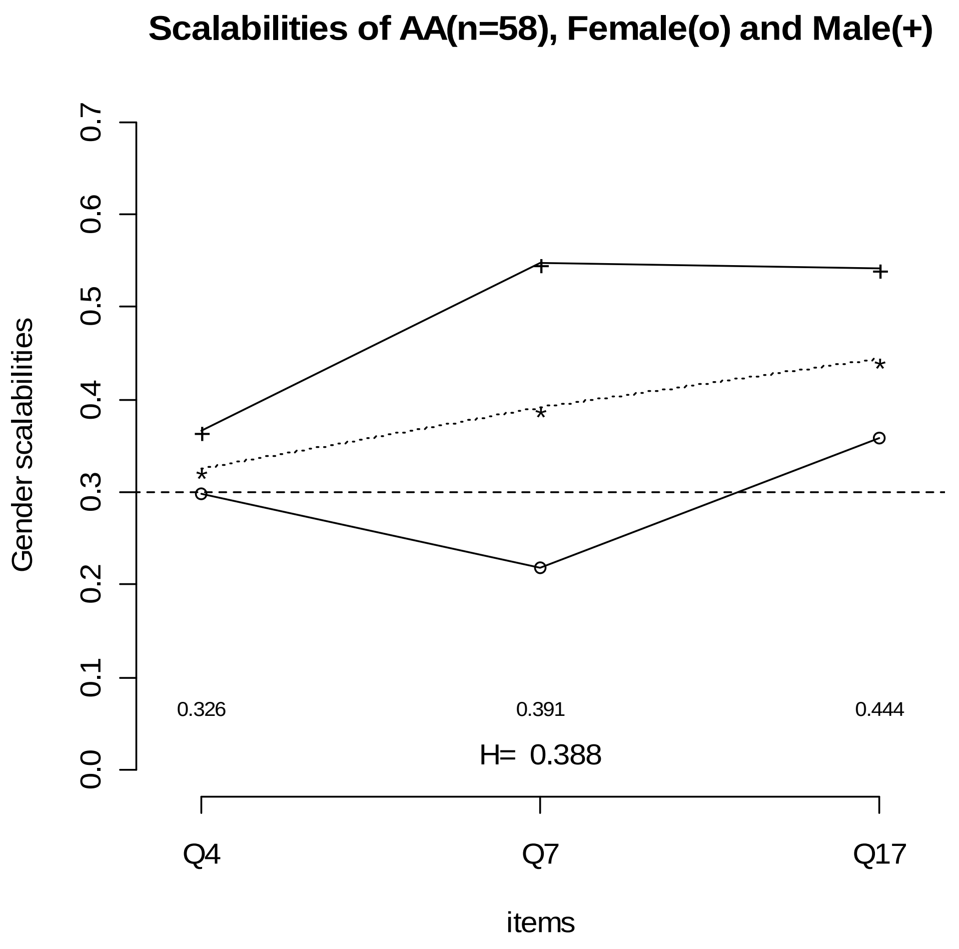
```
   Q4    Q7   Q17
0.322 0.390 0.447

The item set scalability H= 0.387
```

Table I.AA.1.2. shows that all pairwise scalabilities are positive. It can also be concluded that AA constitutes a weak scale, 0.3< H< 0.4.

Investigation of possible gender difference regarding the scalabilities

The questionnaire seems to be more suitable for males than for females in the sense that the items seem to better cooperate for the males as the males in general have larger scalabilities than the females, see fig. I.AA.1.1.

Fig. I.AA.1.1. Gender specific item scalabilities. ‘*’ represent the genders combined.



The split into gender specific scalabilities is presented in Table I.AA.1.3.The number of persons was reduced to n= 58 due to missing gender for one teacher.

Table I.AA.1.3.

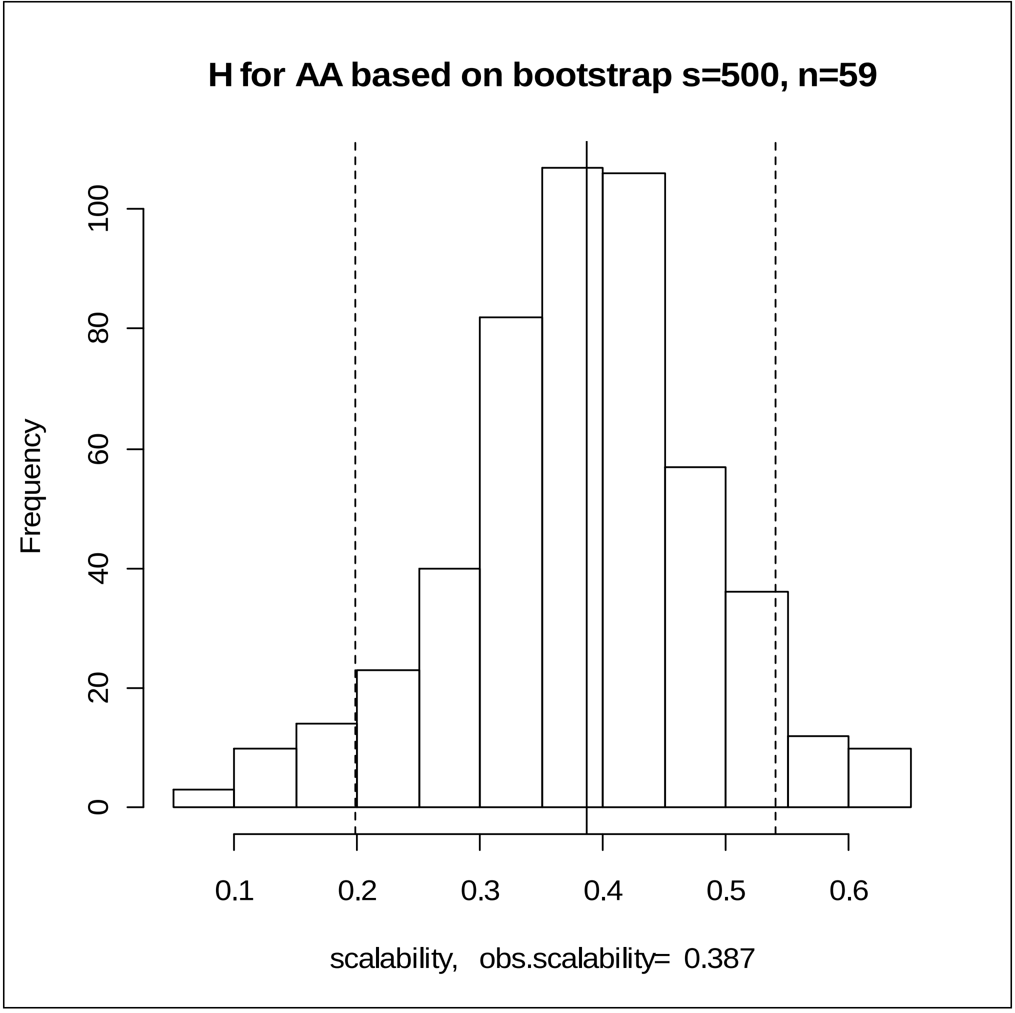
```
Scalabilities by gender, n=58
item     all  Female  Male
Q4      0.326  0.299 0.366
Q7      0.391  0.219 0.548
Q17     0.444  0.358 0.542
H= 0.388
```

Table I.AA.1.4. Investigation of the gender difference.

```
Total scalability AA: H(Female)= 0.292 , H(Male)= 0.487, diff= -0.196
Based on 500 bootstrap replicates

Intervals :
Level     Percentile                 BCa
95%   (-0.6465,  0.2362 )   (-0.6045,  0.2755 )

 Diff:  Min.  Median    Mean   Max.
      -0.797  -0.194  -0.184  0.454
```

In spite of an apparently substantial difference, an evaluation by a bootstrapped C.I. for H(males)-H(females) could not statistically demonstrate any systematic difference The 95% C.I. covers zero for the percentile as well as for the BCa method.

Fig. I.AA.1.2. The variability of the estimated bootstrap scalability H(AA)



Bootstrapping the variability of the item scalability, as is illustrated in Fig. I.AA.1.2., shows that the scalability is well on the positive side and might be as large as 0.5. This is, as such, a good sign to continue to a parametric model. The result of the bootstrapping is seen in table Table I.AA.1.5.

Table I.AA.1.5. Bootstrapping the item set scalability of AA. BCa and Percentile C.I.

```
Level          BCa
90%   ( 0.1812,  0.5403 )                           Percentile
   Min.  1st Qu.  Median    Mean  3rd Qu.    Max.    5%      95%
  0.062   0.324   0.388    0.383   0.444   0.648   0.198   0.541
```

Looking for influential answer profiles when estimating H
There were no particularly influential answer profiles when investigated by the jackknife procedure. Omitting one person with a profile= (5,5,5) decreased the H to 0.346 (minimum). Excluding a person with the profile (1,5,1) increased the H to 0.430 (maximum). None of these scalabilities can be considered extreme. ‘Jackknifing’ more than one person should be avoided in such a small sample.

One aspect, which is not mentioned in the published article, is the problem of incoherent answer profiles when estimating the variability of H.
One potentially influential person was found by the jackknife method:
Excluding ID= 15 yields a high scalability, H= 0.430

Even if there were no strong influential person profiles when estimating H(AA), the bootstrapping procedure reveals a large variability of H(AA).
When there are just a few items in a questionnaire, deviating answer profiles may largely influence the variability of the estimated scalability. Let us consider the frequency of incoherent answer profiles behind the minimum scalability (H<0.1) of 500 bootstrapped samples:

The item sums of the 59 answers in the original data set yield an estimated order of ‘difficulty’.
∑Q4= 180, ∑Q7= 144, ∑Q17= 141

Thus, the estimated order of difficulty: Q4 < Q7 ≈ Q17
5 respondents showed a marked contradicting structure and some of them were repeatedly sampled in the bootstrapped sample, for which H<0.1.

Profile of	Q4	Q7	Q17	freq	gender
1	**1**	**5**	1	4	F
(2	5	**2**	**5**	1)	F
3	**2**	**5**	3	1	M
4	**2**	**5**	4	3	M
(5	**1**	**1**	3	2)	F

Unexpected pairs of answers are marked in bold. The frequencies (4+1+3)/59 = 13.5%. >10% of the sample consists of strongly incoherent profiles (profile no. 1, 3 and 4) , resulting in the low scalability H=0.088. They are 3/59≈ 5% in the actual sample. How likely are they to be 8,as was observed, or more in a random sample with N=59?

If there are, let’s say 5% of the target population, who do strongly disagree with the order of difficulty $Q4 < Q7 \approx Q17$ by scoring as profiles 1,3 and 4, the probability of observing ≥ 8 of such respondents in a sample of 59 can be estimated: P(no. of incoherent profiles ≥ 8)≈ 0.008864. Bootstrapping N= 500 samples yields the expected number of such samples ≈ 4.4. Thus, it is quite possible to get a few such odd samples, which indicates that the bootstrapping procedure works. Even if restricted bootstrapping might be considered, it is not required in this simple approach of naïve bootstrapping.

Monotonicity

The minimum size of the rest score group was set =15, which in essence means a low, median and high scoring group.

Table I.AA.1.6.

```
    ItemH  #vi   maxvi  zmax
Q4   0.32    1   0.04   0.22
Q7   0.39    0   0.00   0.00
Q17  0.45    1   0.06   0.14
```

The analysis showed just a few and non-significant violations against the monotonicity.

Non-intersection

With only 3 items, a non-intersection evaluation is not very useful as there will be only one item forming the rest score. A rough investigation with just one item as the rest score did not indicate any dramatic violation against non-intersection.

With a weak scale and just three items, we cannot expect to obtain a sufficient interval measure when going further two a parametric model in step 2, but it might be a worthwhile exercise for getting further insight in the questionnaire.

Step 2 AA. Analysis by a Rasch RSM

Moving to a parsimonious parametric model will reveal what is possible to achieve with just three items.

Table I.AA.2.1. A Rasch RSM is applied.

A common slope (α = 1) and a common set of thresholds

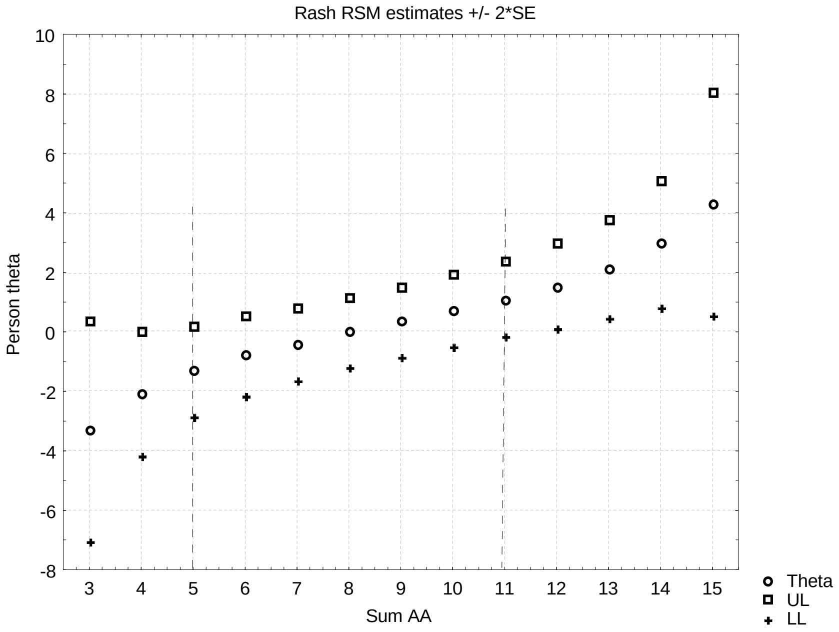
```
---------------------------------------------------
|         MODEL|   INFIT  |EXACT MATCH|ESTIM|      |
|MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|--------------+----------+-----------+-----+------|
|  -.19     .15|1.11    .6| 38.9  41.2|  .75| Q4   |
|   .67     .16|1.03    .2| 50.0  41.4| 1.05| Q7   |
|   .74     .16| .82  -1.0| 46.3  45.2| 1.22| Q17  |
|--------------+----------+-----------+-----+------|
Person reliability = 0.48, Item reliability = 0.86


     Analysis of STANDARDIZED RESIDUAL variance (in Eigenvalue
units)
                                                 -- Empirical
Total raw variance in observations     =          6.4 100.0%
  Raw variance explained by measures   =          3.4  53.4%
    Raw variance explained by persons  =          2.8  43.5%
    Raw Variance explained by items    =           .6  10.0%
  Raw unexplained variance (total)     =          3.0  46.6%
```

In the Rasch setting, it is clear from table I.AA.2.1. that the items are of reasonable reliability (0.86), but obviously too few. They explain just small part, 10%, of the total variation. Matching, obs% ≈ exp%, says that random variation is about what is expected and the estimated discrimination is not far from unity. The low person reliability (table I.AA.2.1.,person rel.= 0.48) is a clear indication that an enlarged set of items is necessary. Fig. I.AA.2.1.and I.AA.2.2. are two alternative ways to illustrate this problem. A sum score has a very low precision in estimating the respondent on the intended scale. The small sample is certainly a cause, but the small set of items is the main reason. With just three items there are only 5^3 possible answer profiles. Furthermore, fig. I.AA.2.2 clearly indicates that the items should be more dispersed (good coverage). The analysis of the 1:st contrast,1.6 in table I.AA.2.1., tells us that there is no immediate concern about any second strong dimension, but this would hardly be expected with just three items.

The person with ID= 15 was detected as an incoherent profile also by the Rasch model. An exclusion of this person did not radically change the result, even if the person and item reliabilities were changed from (0.48, 0.86) to (0.63, 0.88).

There were no violations against the ordering of the categories.

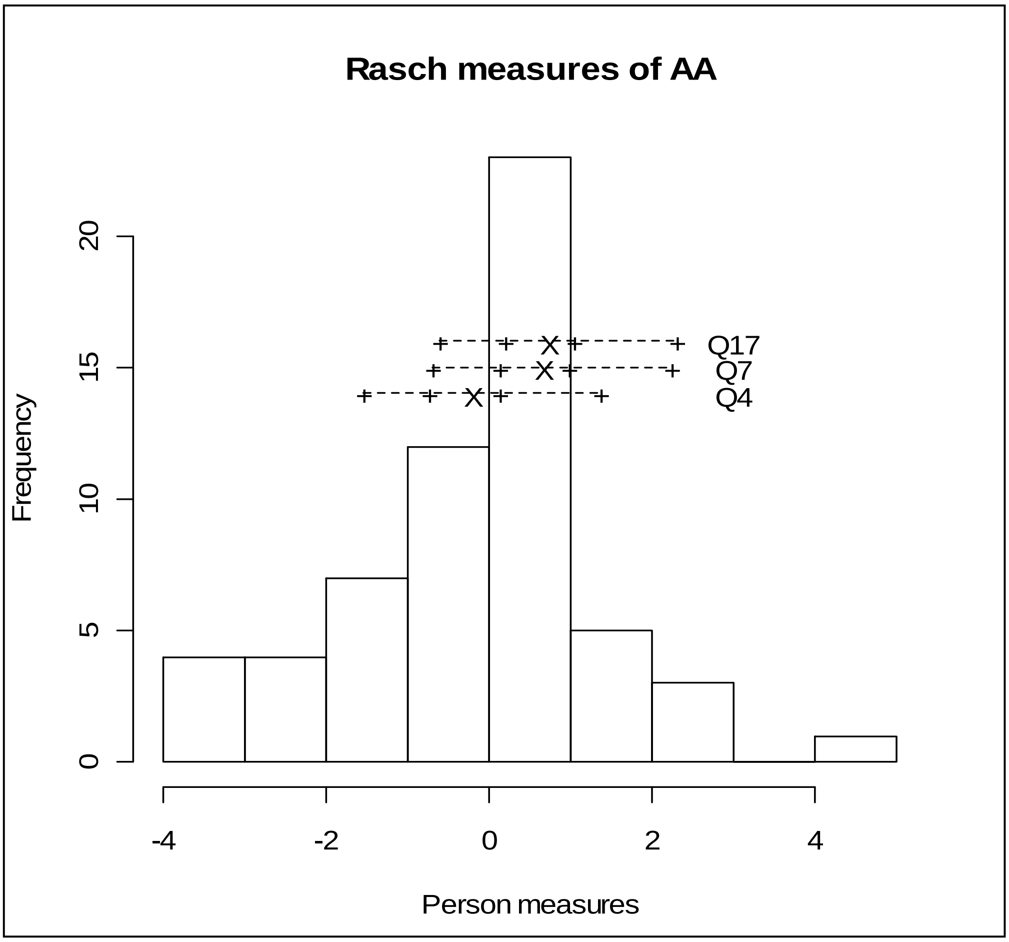
Fig I.AA.2.1. Estimated person measure (theta) from the Rasch RSM model. Approximate 95% C.I. UL= Upper limit, LL= Lower limit



The weakness of a RSM based on just three items can be illustrated by a look at a C.I. for θ. E.g. an estimated θ = 0 covers the possibility of a sum score in the range 5-11 while the total range is 3-15.

Looking at the RSM measures together with the item locations illustrates still more the insufficient coverage, see fig I.AA.2.2..

Fig I.AA.2.2. Histogram of Rasch RSM measures with imposed items and thresholds,



Step 3 AA

A GRM with a constant discrimination (equal slopes) confirms the Rasch approach. There is no sign of violation against the simple model. The item location structure is very similar to the Rasch approach.

Table I.AA.3.1. A GRM with a common slope and a common set of category thresholds

```
.CATEGORY PARAMETER  :     1.437     0.489    -0.421    -1.506
   S.E.             :     0.129     0.115     0.127     0.182
+------+---------+---------+---------+---------+
| ITEM |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+======+=========+=========+=========+=========+
| Q4   |   0.900 |   0.072 |  -0.027 |   0.221 |
| Q7   |   0.900 |   0.072 |   0.722 |   0.210 |
| Q17  |   0.900 |   0.072 |   0.771 |   0.205 |
+------+---------+---------+---------+---------+

          ITEM FIT STATISTICS
 ------------------------------------
 | ITEM | CHI-SQUARE |  D.F. | PROB. |
 ------------------------------------
 | Q4   |    7.38521 |    4. | 0.116 |
 | Q7   |    5.07823 |    5. | 0.407 |
 | Q17  |    7.11980 |    5. | 0.211 |
 ------------------------------------
```

A look at the fit statistics in table I.AA.3.1., where no objections to the model are seen, and at the discrimination estimates in table I.AA.2.1. (Step 2), which are not far from unity, indicates that an elaborated investigation in Step 3 is not needed.

Conclusion about the AA questionnaire

The items work but a sum score based on just three items is insufficient for reasonably ranking the teachers on the intended scale.
The questionnaire shows good item reliability but poor person reliability. A sum score is insufficient to be reliably transformed to an interval scaled measure.
There is an indication that the questionnaire is more suitable for men than for women, or, the women form a more heterogeneous population.

- a weak scale
- insufficient coverage
- more items are needed

Investigation of Activating the Meaning of knowledge (AM)

Step 1 AM. The Mokken scale analysis

The answer structure of the three items, designed to capture AM, is presented in table I.AM.1.1.

Table I.AM.1.1. Descriptive statistics of AM.

Dimension AM							
Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
Q6	8	17	14	15	5	0	59
Q10	39	9	5	5	1	0	59
Q15	10	11	17	13	8	0	59
All items	57	37	36	33	14	0	177

There were no ‘non response’, the score levels are fairly well represented although level 5 was sparsely endorsed for all items.

Table I.AM.1.2.

```
Hij  item pairwise scalabilities, N=59
       Q6   Q10   Q15
Q6  1.000 0.273 0.268
Q10 0.273 1.000 0.470
Q15 0.268 0.470 1.000

Item scalabilities Hi:
   Q6   Q10   Q15
0.270 0.374 0.354

Item set scalability H= 0.331
```

All scalabilities are positive, 0.2< Hij < 0.4, but we are confronted with a weak scale, H< 0.4. Looking at the scalabilities by gender reveals a certain difference, see fig. I.AM.1.1. n=58 due to missing gender for one person.

Fig. I.AM.1.1. Scalabilities by gender, n=58

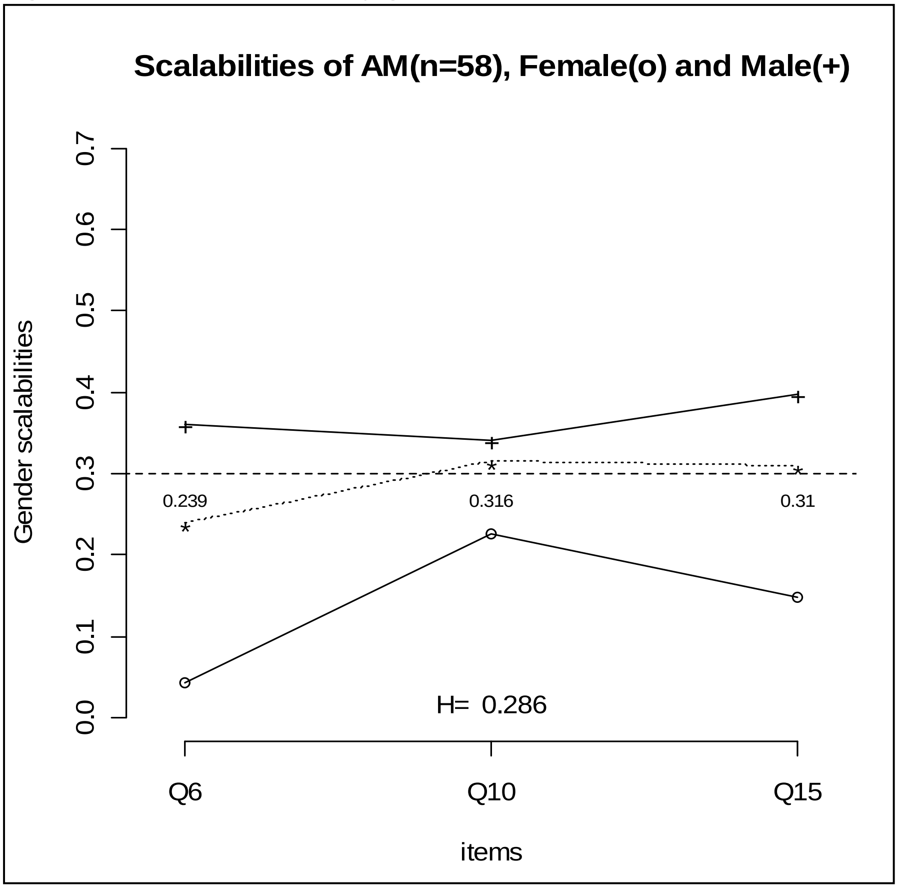


Table I.AM.1.3. Scalabilities by gender, n=58 due to missing gender for one person.

```
    item   all  Female  Male
     Q6   0.239  0.043 0.361
     Q10  0.316  0.227 0.341
     Q15  0.310  0.149 0.398

Hi  n=58
   Q6   Q10   Q15
0.239 0.316 0.310

H(Female)= 0.130
H(Male)  = 0.368

H= 0.286
```

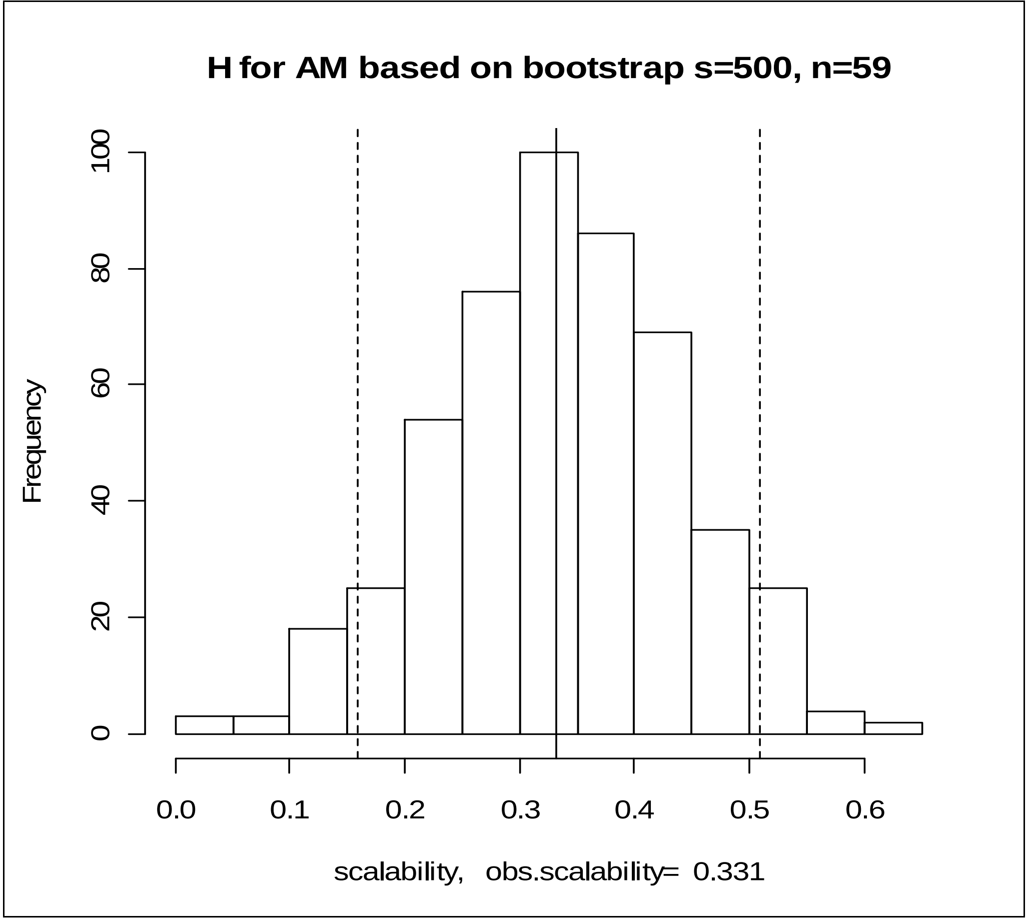
Investigation of a possible gender difference

Table I.AM.1.4. Test of the hypothesis H(Female) = H(Male)

```
Total scalability AM: H(Female)= 0.13 , H(Male)= 0.368 diff= -0.238
Based on 500 bootstrap replicates

Intervals :
Level     Percentile            BCa
95%   (-0.6179,  0.1455 )   (-0.6435,  0.1077 )

Diff:  Min   Median    Mean    Max.
     -0.776  -0.221  -0.225  0.355
```

From table I.AM.1.2.and fig. I.AM.1.1. it can be concluded that AM constitutes a weak scale. The questionnaire seems to be more suitable for males (larger scalabilities). The females do not seem to perceive the questionnaire as representing any particular dimension (table I.AM.1.3.). All item scalabilities for females, and particularly for Q6, are below the acceptable limit 0.3. In spite of an apparently substantial gender difference, an evaluation by a bootstrapped C.I. for H(males)-H(females) could not statistically demonstrate any systematic difference. The estimated C.I covers zero, see Table I.AM.1.4.

Fig. I.AM.1.2. The variability of the estimated bootstrap scalability H(AM)



Bootstrapping the variability of the item scalability, as is illustrated in Fig. I.AM.1.2., shows that the scalability is well on the positive side and might be as large as 0.5. This is, as such, a good sign to continue to a parametric model. The result of the bootstrapping is seen in table Table I.AM.1.5.

Table I.AA.1.5. Bootstrapping the item set scalability of AM.

```
Level         BCa
90%   ( 0.1528,  0.5103 )

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      5%     95%
  0.002   0.265   0.331   0.334   0.407   0.632   0.159   0.510
```

As can be seen in fig. I.AM.1.2., the estimated C.I. is well on the positive side.

Monotonicity

When looking at the monotonicity, no violation against was detected. The minimum size of the rest score groups was set =15, which in essence resulted in a low, median and high scoring group.

The item sums of the 59 answers in the original data set yield an estimated order of ‘difficulty’.
∑Q6= 169, ∑Q10= 97, ∑Q15= 175

Thus, the estimated order of difficulty: Q6≈ Q15< Q17

Influential answer profiles

Two potentially influential persons were found by the jackknife method:
Excluding ID= 72 (1,5,1) or ID= 98 (1,5,1) yields H= 0.367, which is not perceived as an extreme value. Investigation of the female profiles alone did not yield any strongly influential subject, who could be responsible for the low scalability in the female group.

Step 2 AM. Analysis by a Rasch RSM

With a scalability H(female)= 0,13 and H(male)= 0.368 a further step to a parametric model might be questioned. However, this is not an obstacle to investigate a Rasch approach.

Table I.AM.2.1.

```
----------------------------------------------------
|         MODEL|   INFIT  |EXACT MATCH|ESTIM|      |
|MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|--------------+----------+-----------+-----+------|
|   .03     .16|1.06    .4| 36.4  43.7|  .75| Q6   |
|  2.04     .20|1.07    .4| 56.4  58.2| 1.17| Q10  |
|  -.11     .16| .91   -.4| 52.7  42.9| 1.07| Q15  |
|--------------+----------+-----------+-----+------|
Person reliability= 0.55, Item reliability= 0.97
```

The persons with ID= 72 and 98, identified as potential outliers in Step 1, did not appear as extremes in the Rasch model even if they showed the worst person fit statistics.

Table I.AM.2.2. DIF: Males vs Females

```
--------------------------------------------
| SUMMARY DIF                ITEM          |
| CHI-SQUARE   D.F.  PROB.  Number Name    |
|------------------------------------------|
|     1.0980      1  .2947       1 Q6      |
|      .0153      1  .9016       2 Q10     |
|      .9085      1  .3405       3 Q15     |
--------------------------------------------
```

The slight difference between males and females, found in Step 1, can be further investigated from a Rasch model perspective. However, no systematic DIF between the genders is indicated, see table I.AM.2.2.

Besides the weak person reliability, the Rasch model looks reasonable, with estimated discrimination coefficients not far from unity. Matching; Obs% vs Exp% (Table I.AM.2.1.) tells us that Q6 is the weakest item, more random noise than expected from the model (36.4 < 43.7). Q10 and Q15 are somewhat better but still weak.

There were no violations against the ordering of the categories.

Step 3 AM

Proceeding to a GRM, even if it is not indicated by Step 2, reveals a different structure.
A test yields a statistically demonstrated significance against a common slope(discrimination) model.

Table I.AM.3.1. Item fit statistics

```
------------------------------------
| ITEM | CHI-SQUARE |  D.F. | PROB. |
------------------------------------
| Q6   |    5.94030 |    5. | 0.312 |
| Q10  |   19.68026 |    5. | 0.002 |
| Q15  |   10.24143 |    5. | 0.068 |
------------------------------------
|Total |   35.86198 |   15. | 0.002 |
------------------------------------
```

Conclusion about the AM questionnaire

The scalability analysis reveals that a sum score is a possible but not a very good measure of the AM construct, particularly not for the women.
The Rasch model (Step 2) does not clearly reveal the structure of the questionnaire. The high item reliability tells us that the items are relevant, but they are not sufficient (low person reliability).
Step 1 and Step 3 indicates that a sum score might not be a very useful measure to represent the respondents, particularly not for women. There is very little sign that the sum score is useful in ranking the women on the AM scale.
More items are needed and the three items already in the questionnaire should be more adapted for women. The indication against common slopes, found in Step3, should not be taken too seriously as there is no ‘bulk’ of items forming a group with common slopes.

Investigation of Activating the Reproduction of knowledge (AR)

Step 1 AR. The Mokken scale analysis

The answers from the AR questionnaire are well distributed over the 5 levels although there were 7 non-responses. No imputation was made.

Table I.AR.1.1. Descriptive statistics of AR.

Dimension AR							
Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
Q2	9	7	12	15	16	0	59
Q9	9	10	15	18	7	0	59
Q13	5	11	9	3	24	7	59
All items	23	28	36	36	47	7	177

The scalability analyses in table I.AR.1.2. are based on the answers from 52 teachers. The total as well as gender specific groups are investigated.

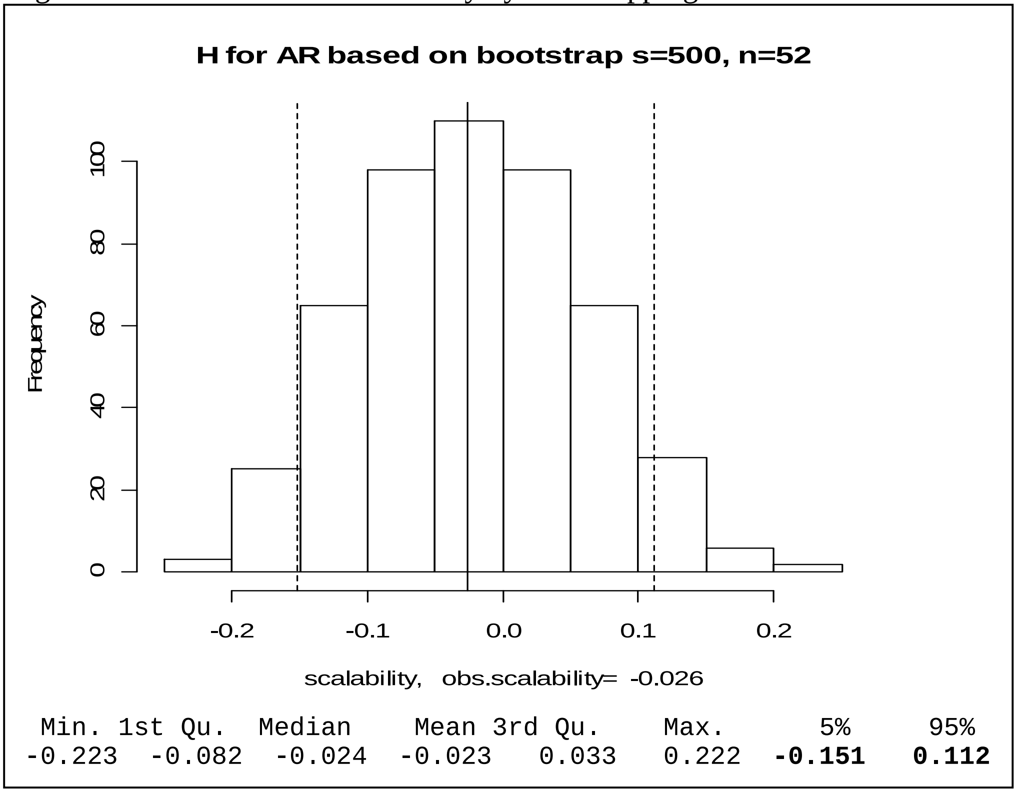
Table I.AR.1.2.

```
Hij: item pairwise scalabilities, N=52

        Q2     Q9    Q13
Q2   1.000 -0.193  0.217
Q9  -0.193  1.000 -0.140
Q13  0.217 -0.140  1.000

Item scalabilities Hi:
    Q2     Q9    Q13
 0.028 -0.166  0.049

Item set scalability H= -0.026

AR scalability by gender
    item    all Female   Male
Q2     2 -0.019  0.008 -0.027
Q9     3 -0.206 -0.321 -0.151
Q13    4  0.001  0.000  0.019
```

We can immediately notice that there is vertically no coherence between the items. The major part is even negative. Let us look at the item set scalability and its variability by bootstrapping, see fig. I.AR.1.1.

.

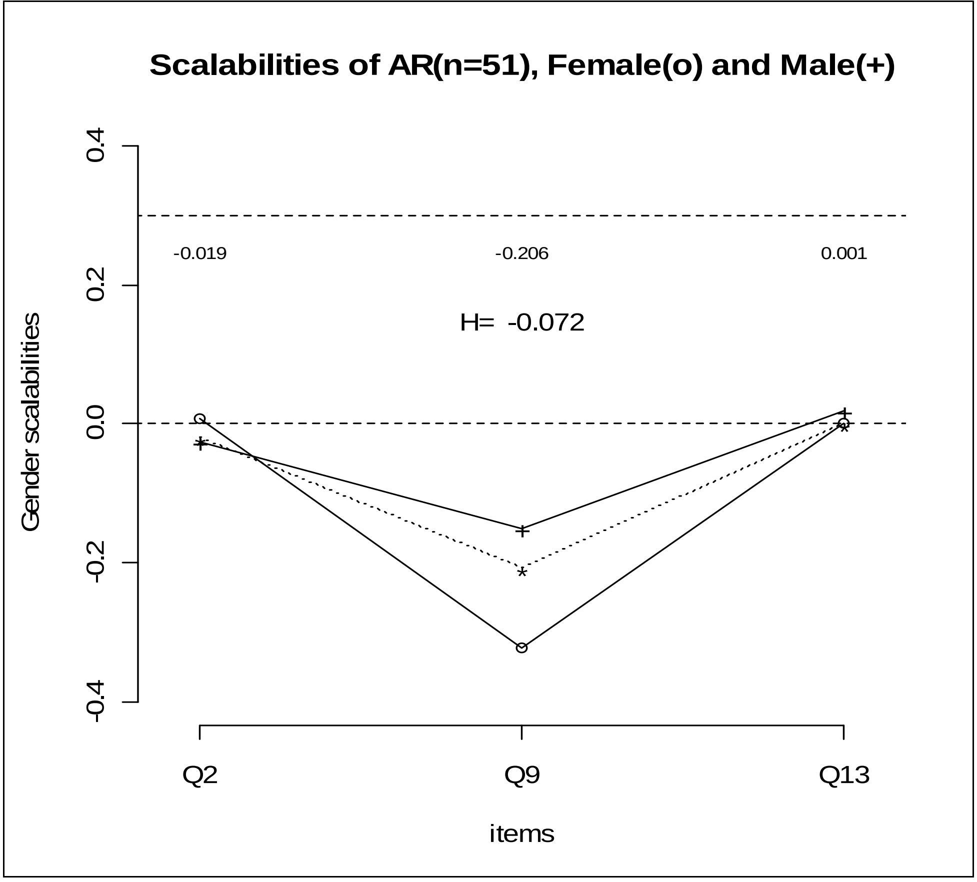
Fig. I.AR.1.1. The item set scalability by bootstrapping



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      5%     95%
-0.223  -0.082  -0.024  -0.023   0.033   0.222  -0.151   0.112
```

From fig. I.AR.1.1., we can conclude that an item set virtually equal to zero and with an upper C.I. limit just about 0.1 is a strong indication of a not meaningful sum score.

Fig. I.AR.1.2. Item scalabilities by gender



There is no obvious gender difference. The scalability for both the genders is far below what is reasonable. The low scalabilities might depend on strongly incoherent answer profiles. However, no influential profiles were found by the Jackknife method.

Thus, the sum score does not seem to be useful for ranking the respondents, and consequently no reasonable transformation to an interval scaled measure seems probable.

<u>Step 2 AR</u>. Analysis by a Rasch RSM

If we ignore the message from Step 1 and try a Rasch model, the analysis will turn out as follows:

Table I.AR.2.1. Applying a Rasch RSM

```
-----------------------------------------------------
|          MODEL|   INFIT  |EXACT MATCH|ESTIM|      |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|---------------+----------+-----------+-----+------|
|   -.29     .12| .90   -.6| 34.5  32.8| 1.23| Q2   |
|   -.04     .12|1.12    .8| 25.9  32.6|  .39| Q9   |
|   -.48     .13|1.00    .1| 19.2  32.5| 1.34| Q13  |
|---------------+----------+-----------+-----+------|
Person reliability= 0.15, Item reliability= 0.53
```

Matching; Obs% vs Exp% (Table I.AR.2.1.) indicates more random noise than expected, and Exp% is very low. ‘Estim.Discr.’ indicate a very low capacity for Q9 to separate individuals by a Rasch measure. The main message comes from the reliability estimates which, from a Rasch model perspective, are clearly insufficient. This message is in agreement with the result from Step 1.

There were, as can be expected, some violations against the ordering of the categories.

Step 3 AR

Just a look at step3, where a model with a common item discrimination indicates a bad fit.

Table I.AR.3.1. ITEM FIT STATISTICS

```
 ------------------------------------
 | ITEM | CHI-SQUARE |  D.F. | PROB. |
 ------------------------------------
 | Q2   |    4.63781 |    4. | 0.326 |
 | Q9   |   12.74306 |    4. | 0.013 |
 | Q13  |   21.91810 |    4. | 0.000 |
 ------------------------------------
 | Total|   39.29897 |   12. | 0.000 |
 ------------------------------------
```

The model does not fit the data. Further analysis is hazardous, thus we need to return to Step 1.

Let us see what happens if we proceed to a model with item specific discriminations? Of course it will be a better fit!! See Table I.AR.3.2.

Table I.AR.3.2. A model with item specific discriminations

```
+------+---------+---------+---------+---------+
| ITEM |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+======+=========+=========+=========+=========+
| Q2   |   0.206 |   0.033 |  -1.077 |   0.812 |
| Q9   |   0.258 |   0.041 |  -0.339 |   0.614 |
| Q13  |   0.142 |   0.033 |  -2.990 |   1.075 |
+------+---------+---------+---------+---------+

         ITEM FIT STATISTICS
 ------------------------------------
 | ITEM | CHI-SQUARE |  D.F. | PROB. |
 ------------------------------------
 | Q2   |    3.84092 |    5. | 0.575 |
 | Q9   |    5.96680 |    6. | 0.427 |
 | Q13  |    5.98357 |    3. | 0.111 |
 ------------------------------------
 |Total |   15.79129 |   14. | 0.326 |
 ------------------------------------
```

If we proceed to a model, presented in table I.AR.3.2., it seems as such a model will fit. A look at the estimated slopes is a clue to not to approve the model. As a rule of thumb, an item discrimination $a_i <$ 0.5 is an indication of more noise than information. This model approach is an 'over parameterization', in the sense that it is too data driven. Such a complicated model is potentially misleading. Therefore, we should rely on Step 1.

Conclusion about the AR questionnaire

- The items do not cooperate towards a meaningful measure.
- The sum score is not a relevant aggregated measure.
- The questionnaire has to be redesigned and enlarged.

The reliability coefficients, based on the Rasch model, indicate that we are doing better with the items than with the teachers, which in essence means that more items are required to actually catch the teachers' teaching practice. The teachers are too heterogeneous to be characterised by such a short questionnaire.

Conclusion about the 'Teaching strategy profile' questionnaire

Regarding the '3 step strategy', a brief result can be outlined as follows:

AA: Step 2 is sufficient for a conclusion
AM: Step 3 is needed to invalidate Step 2. Step 1 is sufficient for essential information.
AR: Step 1 is sufficient for a general conclusion.

The '3- step strategy' clearly shows that it is important to start with Step 1. Step 1 reveals most of the essential information about the questionnaire. Direct application of parametric models may be misleading or ambiguous, with obvious risks of erroneous and too data driven interpretations. The JMLE method, used in Step 2, is criticised for 'noticeable estimation bias with short tests … In practise, however, this bias has few implications because the relative ordering and placement of the estimates is maintained' [Smith & Smith, 2004, ch. 2].

Study II

Nissell M., Brodin, U., Christensen K., Rydelius P-A. **The Imperforate Anus Psychosocial Questionnaire (IAPSQ): Its construction and psychometric properties.** Child and Adolescent Psychiatry and Mental Health 2009, 3:15 (14 May 2009).

The ori$_g$in of the present study was to develop the liaison work between the disciplines of child and adolescent psychiatry and paediatric surgery and nursing, so as to improve the quality of treatment and care of a group of children with imperforate anus (**IA**) and their families. Imperforate anus is a congenital disease involving a deformity of the anorectum. The early surgery and invasive follow-up treatment associated with IA may affect the child psychosocially, including the child-parent relationship. By developing and testing a questionnaire for children born with anorectal anomalies, a tool for measuring psychosocial functioning can be realized.

The Imperforate Anus Psychosocial Questionnaire (IAPSQ) contains 45 items, intended to be Likert scales. Twenty-three items were considered to represent the psychological dimension and twelve items to represent the social dimension. A total of 87 children completed the IAPSQ: 25 children with IA and two comparison groups. An IRT approach was used to evaluate the psychometric properties of the IAPSQ, where item difficulty and person ability were concurrently approximated.

The findings of the IRT analysis revealed that the psychological dimension was reasonable, and that person reliability (0.83) was moderate and item reliability (0.95) was sufficient. The social dimension showed satisfactory item reliability (0.87). The person reliability (0.52) of the social dimension was weak. Content validity seemed to be established and construct validity was recognized on the psychological dimension.

It was concluded that the IAPSQ provides a reasonably valid and reliable measure of psychosocial functioning for clinical use among children with **IA**, although some revisions are suggested for the next version of the IAPSQ. It was discovered from the analysis that specific items should be discarded and other items should be reformulated to make the questionnaire more “on target”. The “social item set” has to be expanded with further items to reasonably capture the intended social dimension.

For reasons of comparison, the questionnaire was constructed and tested so as to be appropriate even for children without **IA.** Two comparison groups with experiences of clinical care were selected for participation to enhance interpretation of the findings. Comparison Group I contained children with a chronic condition: juvenile chronic arthritis (**JCA**). Like the **IA** children, this group of children had suffered from pain and emotional stress, though of another type. The inclusion criteria were an illness debut before the age of two years and joint injections before the age of 4 years. This group of children, n=30, was recruited from the medical records at the outpatient clinic for paediatric rheumatism.

Comparison Group II consisted of children who had undergone minor surgery (e.g., for a hernia), and thus who had some experience of hospital care. The families were consecutively recruited at the day surgery clinic. The children in Comparison Group II, n=32, had no chronic condition (**NCC**) and were otherwise healthy.

The children were from 8 to 14 years old and the groups were matched for age.

The questionnaire is constructed in two parts to reveal a **social** and a **psychological** dimension. Within each dimension, the questionnaire is intended to work equally well for the three patient groups, **IA**, **JCA** and **NCC**. This means that the patients can be represented on the actual dimension by the formula:

$\theta_{n(i)} = C + G_i + \theta_n|G$, where $\theta_{n(i)}$ is the aggregated measure for patient n within group i, based on the answer profile, G is the group effect, i=1,2,3 and $\theta_n|G$ the patient's position on the latent trait with the grouping affiliation left out. In this setting, possible group differences are expected to be collected in the G factor. C is a suitable constant.
As there are three distinctly defined groups, the primary analyses have to be carried out within groups.

Investigation of the the Social questionnaire (Soc)

Step 1 Soc. The Mokken scale analysis

The observed scores from the 87 children are shown in table II.Soc.1.1.

Table II.Soc.1.1.

Frequency Table. Social dimension.
12 items. Scores from 87 subjects.

Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
SCHOO4	0	4	14	24	45	0	87
TEACH5	1	1	3	21	61	0	87
FRIEND6	0	1	3	28	54	1	87
GYMN7	3	2	4	24	53	1	87
SHOW8	10	8	12	23	34	0	87
BREAK9	3	2	2	22	58	0	87
ACTI10	1	0	10	13	62	1	87
FRIEN28	36	39	7	1	2	2	87
DECI29	26	30	25	1	2	3	87
TEAS30	2	3	13	29	38	2	87
BULLY31	0	2	2	9	72	2	87
TOGETH32	12	10	47	15	1	2	87
All Grps	94	102	142	210	482	14	1044

The answer structure from the total set of children (IA + JCA + NCC) reveals a set of easy items (SCHOO4, TEACH5, FRIEND6, GYMN7 BREAK9 andACTI10), for which most of the children scored high. FRIEN28 and DECI29 are perceived as items difficult to score high. However, this structure may be different when we look at the groups separately. There are some 'non responses'. No imputation is made.

Table II.Soc.1.2. Score frequencies within groups, IA, JCA and NCC

Frequency Table. Social dimension.
12 items. Scores from 87 subjects.

Group	Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
IA	SCHOO4	0	0	2	11	12	0	25
IA	TEACH5	0	0	2	3	20	0	25
IA	FRIEND6	0	0	2	4	19	0	25
IA	GYMN7	0	2	2	8	13	0	25
IA	SHOW8	4	2	5	8	6	0	25
IA	BREAK9	0	1	1	5	18	0	25
IA	ACTI10	0	0	3	3	18	1	25
IA	FRIEN28	7	14	1	0	2	1	25
IA	DECI29	5	8	9	0	1	2	25
IA	TEAS30	0	0	2	9	13	1	25
IA	BULLY31	0	0	0	0	24	1	25
IA	TOGETH32	2	1	17	3	1	1	25
Total		18	28	46	54	147	7	300
JCA	SCHOO4	0	2	6	4	18	0	30
JCA	TEACH5	0	0	1	8	21	0	30
JCA	FRIEND6	0	1	0	9	20	0	30
JCA	GYMN7	3	0	0	9	18	0	30
JCA	SHOW8	3	1	3	10	13	0	30
JCA	BREAK9	1	0	0	8	21	0	30
JCA	ACTI10	1	0	3	2	24	0	30
JCA	FRIEN28	17	10	2	1	0	0	30
JCA	DECI29	14	7	7	1	1	0	30
JCA	TEAS30	1	2	5	9	13	0	30
JCA	BULLY31	0	0	2	2	26	0	30
JCA	TOGETH32	2	5	19	4	0	0	30
Total		42	28	48	67	175	0	360
NCC	SCHOO4	0	2	6	9	15	0	32
NCC	TEACH5	1	1	0	10	20	0	32
NCC	FRIEND6	0	0	1	15	15	1	32
NCC	GYMN7	0	0	2	7	22	1	32
NCC	SHOW8	3	5	4	5	15	0	32
NCC	BREAK9	2	1	1	9	19	0	32
NCC	ACTI10	0	0	4	8	20	0	32
NCC	FRIEN28	12	15	4	0	0	1	32
NCC	DECI29	7	15	9	0	0	1	32
NCC	TEAS30	1	1	6	11	12	1	32
NCC	BULLY31	0	2	0	7	22	1	32
NCC	TOGETH32	8	4	11	8	0	1	32
Total		34	46	48	89	160	7	384

As is clear from table II.Soc.1.2., and also expected due to the small sample sizes, many categories are empty when we look at separate groups. In the IA group, about 50% of the answers fell in ‘score 5’, indicating a questionnaire with ‘low difficulty’. There are a few ‘non responses’. No imputation is made. In spite the many empty cells, it is worthwhile to start with the Mokken scale analysis. Some rough messages might be found.

In this first step, the analysis is performed within groups in order to keep the grouping factor outside. The analysis of the total set is also of interest as the objective is a questionnaire equally applicable for all groups.

Table II.Soc.1.3. Item scalabilities within groups.

	Item nr	All	IA	JCA	NCC
SCHOO4	1	0.244	0.293	0.365	0.149
TEACH5	**2**	**0.081**	**0.147**	**0.077**	**0.069**
FRIEND6	3	0.277	0.357	0.339	0.221
GYMN7	4	0.296	0.419	0.383	0.213
SHOW8	5	0.262	0.257	0.499	0.091
BREAK9	6	0.309	0.407	0.449	0.223
ACTI10	7	0.230	0.145	0.339	0.177
FRIEN28	8	0.216	0.389	0.303	0.046
DECI29	9	0.301	0.474	0.390	0.113
TEAS30	10	0.220	0.199	0.372	0.129
BULLY31	11	0.113	-	0.308	0.027
TOGETH32	**12**	**-0.010**	**-0.072**	**0.046**	**-0.002**

The range of the item set H_{1-12} = [-0.010,0.309
The total item set H= 0.221 n= 82

The range of the item set H_i within groups:
IA: H_i= [-0.07,0.47] H= 0.286 n= 22 BULL31 excl.
JCA: H_i= [0.05,0.50] H= 0.348 n= 30
NCC: H_i= [-0.00,0.22] H= 0.117 n= 30

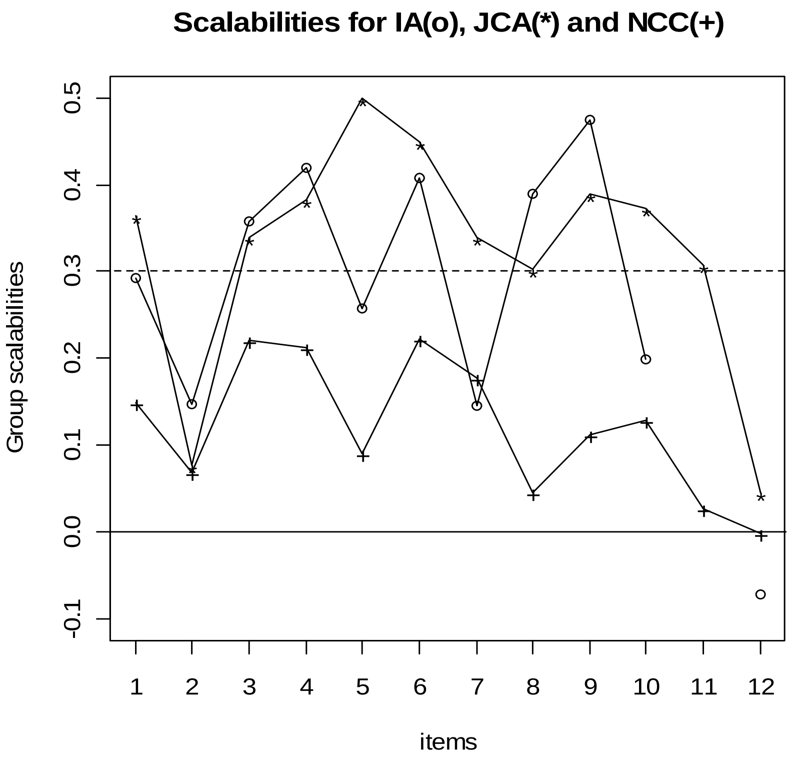
The basic analysis, shown in table II.Soc.1.3., reveals TOGETH32 as a disturbing, not contributing item. This message is the same for all three groups. Besides being not contributing, such an item often has the capability to deteriorate many of the other scalabilities in the item set.

Table II.Soc.1.4. Item scalabilities within groups. Togeth32 excluded

	Item nr	All	IA	JCA	NCC
SCHOO4	1	0.252	0.317	0.358	0.154
TEACH5	2	0.076	0.131	0.052	0.072
FRIEND6	3	0.311	0.372	0.369	0.280
GYMN7	4	0.297	0.445	0.382	0.218
SHOW8	5	0.303	0.311	0.538	0.132
BREAK9	6	0.322	0.451	0.477	0.208
ACTI10	7	0.244	0.102	0.369	0.199
FRIEN28	8	0.245	0.479	0.316	0.061
DECI29	9	0.337	0.534	0.433	0.118
TEAS30	10	0.257	0.271	0.398	0.157
BULLY31	11	0.137	-	0.334	0.040
H(11 items)			0.351	0.385	0.146
The item set		H= 0.261			

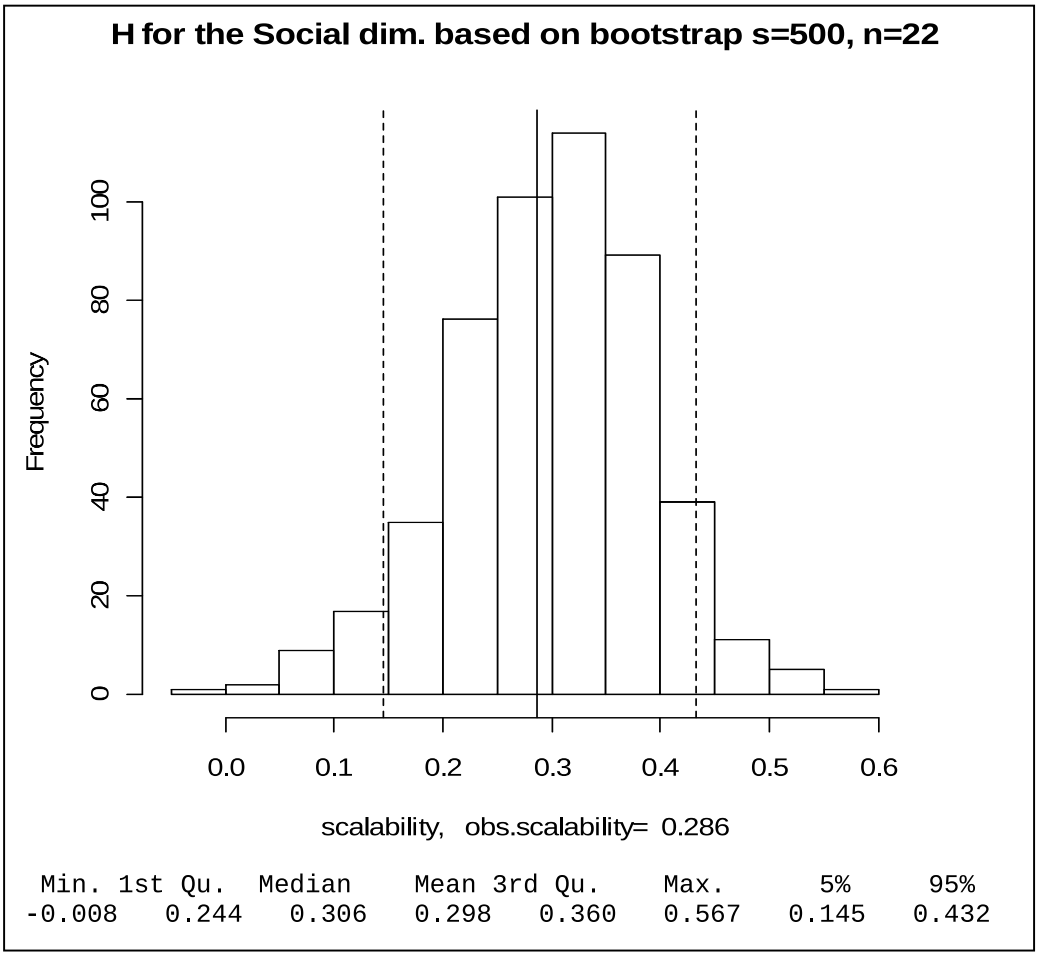
Exclusion of ‘the disturbing item’ raised the scalabilities with 23%, 11% and 25% for the three groups respectively and with 18% for the item set scalability, see table II.Soc.1.4. Based on the item scalabilities, NCC appears as a more heterogeneous group (low scalability) as compared to IA and JCA.

Fig. II.Soc.1.1. Scalabilities within groups. The dashed line indicates the recommended minimum level 0.3 for a useful item.



It is evident from table II.Soc.1.4. and Fig. II.Soc.1.1. that we are confronted with weak scales, particularly within the NCC group. TEACH5 (item 2 in fig. II.Soc.1.1.), with 4 negative item pair scalabilities and TOGETH32 (item 12 in fig. II.Soc.1.1.) with 6 negative item pair scalabilities, are recognised as non-contributing items for all three groups. No child in the group IA had been bullied at school, all responded in the extreme category, therefore no item scalability can be calculated for this group. NCC, as the healthy group with no chronic condition, is very heterogeneous with respect to the actual items, or the other way round – the items are not capable to catch this group, which make it difficult to form a sum score representing a ranking on a social scale. As a consequence, we cannot expect to find systematic differences in a comparison with the other two groups.
No strongly influential profile, as evaluated by the jacknife method, was found for the social dimension.

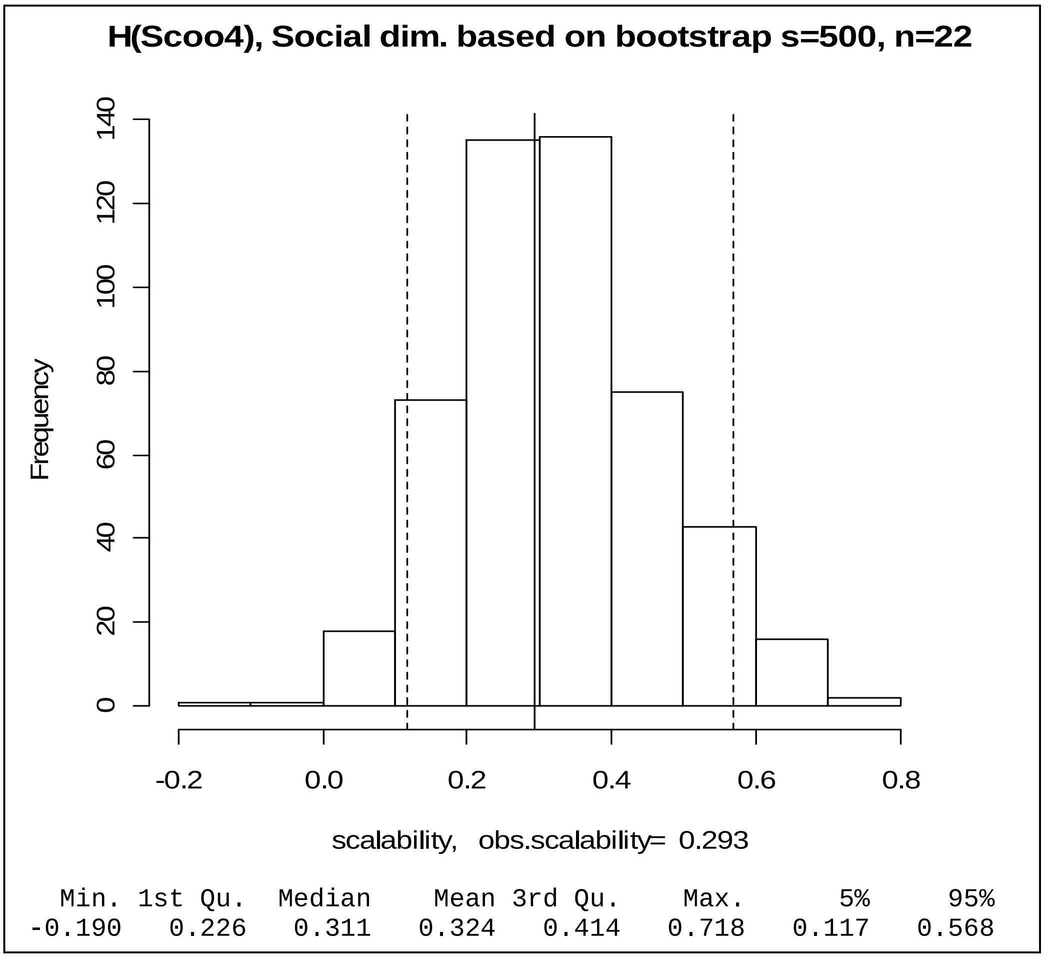
Fig II.Soc.1.2. Scalability for group IA. 12 items.



```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      5%     95% 
-0.008   0.244   0.306   0.298   0.360   0.567   0.145   0.432 
```

The 90% C.I.[0.145, 0.432], as calculated in fig II.Soc.1.2., is well on the positive side of zero and indicates that the items have at least something in common which might constitute an aggregated measure on the social dimension.

Fig. II.Soc.1.3. Scalability of the item Scoo4 for Group IA



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      5%     95% 
 -0.190   0.226   0.311   0.324   0.414   0.718   0.117   0.568 
```

Even an item at the ‘border 0.3’, such as Scoo4, has a positive lower confidence limit, 0.117. Besides a few items, the structure of the scalability profile is about the same for the three groups. Thus a test of a systematic difference in item set scalability between IA+JCA and NCC is justified. As against to what is earlier suggested, we keep the disturbing item in this comparison. Thus, the comparison between groups is based on the actual formulation of the questionnaire. This also makes the comparison conservative.

Table II.Soc.1.5.   Scalability difference H(IA+JCA) - H(NCC) by bootstrapping.

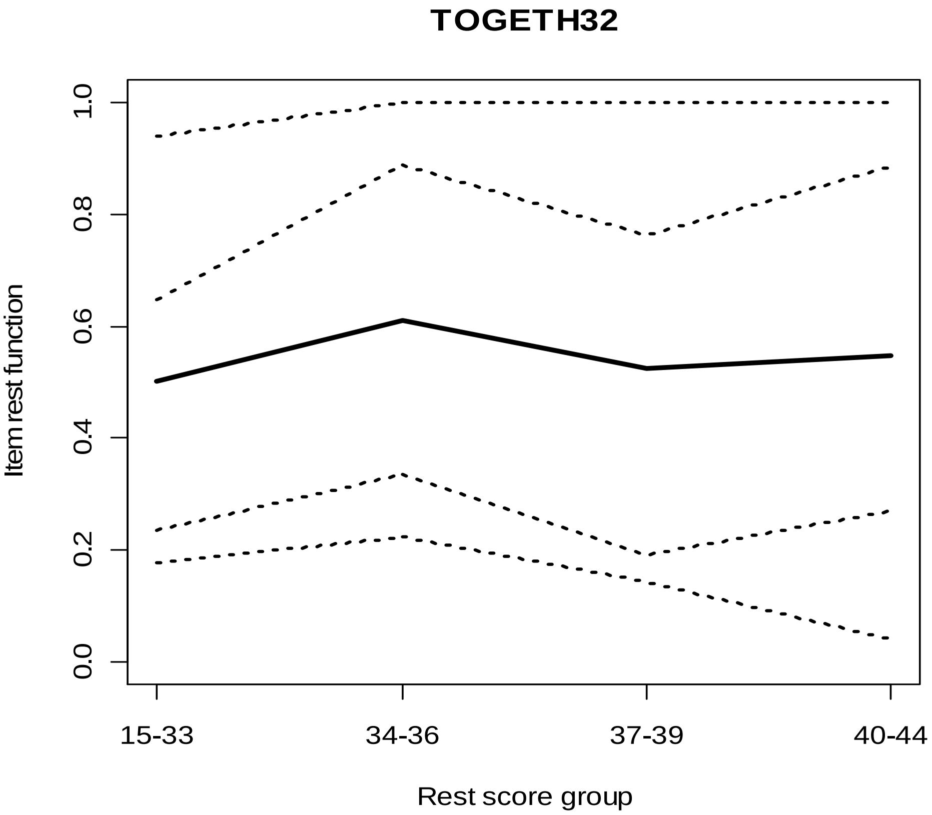
```
H(IA+JCA)= 0.306,   H(NCC)= 0.117,  H(IA+JCA) – H(NCC) = 0.189
90% C.I. for the difference by percentiles [0.028,  0.360]
A BCa corrected 90% C.I. [0.036,  0.386]
```

The C.I. indicates a one sided significant difference at the 5% level. The systematic difference between IA+JCA and NCC, says that the questionnaire is less suitable for the NCC group. The group differences are not dramatic and bearing in mind the small group sizes, the subsequent analyses are performed on the total sample. However, DIF will be investigated when feasible.

Monotonicity and non intersection. The total sample, n=82, is considered.

Investigation of item TOGETH32. The probability(of scoring higher) does not increase with the sum score for the rest score group. This illustrates TOGETH32 as a redundant item.

Fig. II.Soc.1.4. Monotonicity of Togeth32. The minimum rest score group size = 15.



Analysis of the monotonicity is mostly inefficient as there are high scores already at lower levels. Some, but not dramatic, violations against non-intersection were found.

Fig. II.Soc.1.5. Item Together32 excluded.

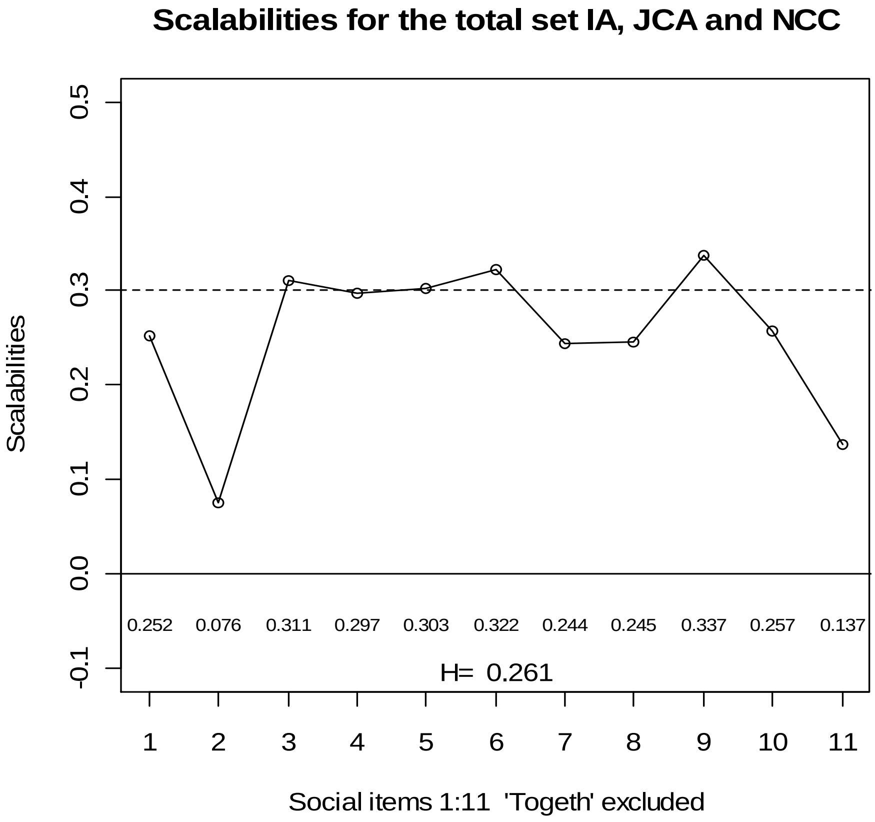


Fig. II.Soc.1.5. reveals item2(TEACH5) as the next candidate for reformulation or exclusion. Exclusion of TEACH5 yields about the same item scalabilities. The scalability for the 10 item set is raised to 0.296. A 90% C.I. for TEACH5 is [-0.020, 0.21] indicates the item as non-contributing. However, it seems slightly contributing for the IA children. A scalability plot will virtually the same as fig. II.Soc.1.5. , besides TEACH5, with slightly increased item scalabilities.

However, with the small sample size and the differences between groups in mind, there will be a risk to exclude such a variable on too loose grounds.

The raw sum score can barely be used for ranking IA and JCA children, but the NCC group perceives the questionnaire quit differently and the item set is probably not very useful.

However, a parsimonious Rasch model might reveal further characteristics, but is otherwise unlikely to constitute a reasonable model. I.e. a raw sum score might not be suitable for a transformation to a valid interval scaled variable.

<u>Step 2 Soc.</u> Analysis by a Rasch RSM

Table II.Soc.2.1. Fitting a Rasch RSM.

```
------------------------------------------------------------
|            MODEL|   INFIT  ||EXACT MATCH|ESTIM|         |
|    MEASURE  S.E. |MNSQ  ZSTD|| OBS%  EXP%|DISCR| ITEM    |
|-----------------+----------++-----------+-----+---------|
|       .11     .14| .93   -.4|| 52.9  47.7| 1.05| SCHOO4  |
|      -.60     .18|1.38   1.7|| 69.0  67.0|  .89| TEACH5  |
|      -.50     .17| .70  -1.5|| 72.1  63.5| 1.14| FRIEND6 |
|      -.16     .15|1.11    .6|| 55.8  54.5| 1.09| GYMN7   |
|       .83     .11|1.28   1.8|| 41.4  39.6|  .94| SHOW8   |
|      -.32     .16|1.18    .9|| 65.5  58.2| 1.10| BREAK9  |
|      -.50     .17|1.17    .8|| 64.0  63.4| 1.03| ACTI10  |
|       .13     .14| .99    .0|| 56.5  48.0|  .89| FRIEN28 |
|       .59     .12| .65  -2.5|| 52.4  41.9| 1.34| DECI29  |
|       .28     .13| .98    .0|| 49.4  45.5| 1.05| TEAS30  |
|     -1.20     .23|1.60   2.0|| 80.0  80.0|  .94| BULLY31 |
|      1.36     .11|1.13   1.0|| 38.8  36.4|  .28| TOGETH32|
|-----------------+----------++-----------+-----+---------|
|Mean   .00        |Person reliability = 0.61
|S.D.   .67        |  Item reliability = 0.94
--------------------
```

Standardized residuals variance (in Eigenvalue units)

```
                                                 Empirical
  Total raw variance in observations   =        19.3 100.0%
  Raw variance explained by measures   =         7.3  37.7%
    Raw variance explained by persons  =         3.5  17.9%
    Raw Variance explained by items    =         3.8  19.8%
  Raw unexplained variance (total)     =        12.0  62.3%
  Unexplained variance in 1st contrast =         2.2  11.3%
  Unexplained variance in 2nd contrast =         1.8   9.2%
```

Much of the variation is still unexplained, which could be expected from Step 1. No strong second dimension is indicated, but might well be there as we do not have any suitable first model. The low person reliability confirms the message from Step 1 about weak scalabilities. Even if the item discriminations are estimated after fitting the model, ‘Togeth32’ is indicated as a non-contributing item, a finding in agreement with the message from Step 1. Releasing the ‘fixed’ set of item thresholds does not alter the result. TEACH5 is not immediately recognised as a problematic item, but if we exclude this item along with TOGETH32, as suggested by step 1, we get a result very similar to the above stated, with person and item reliability just slightly changed to 0.60 and 0.91 respectively.

Ordering of the categories

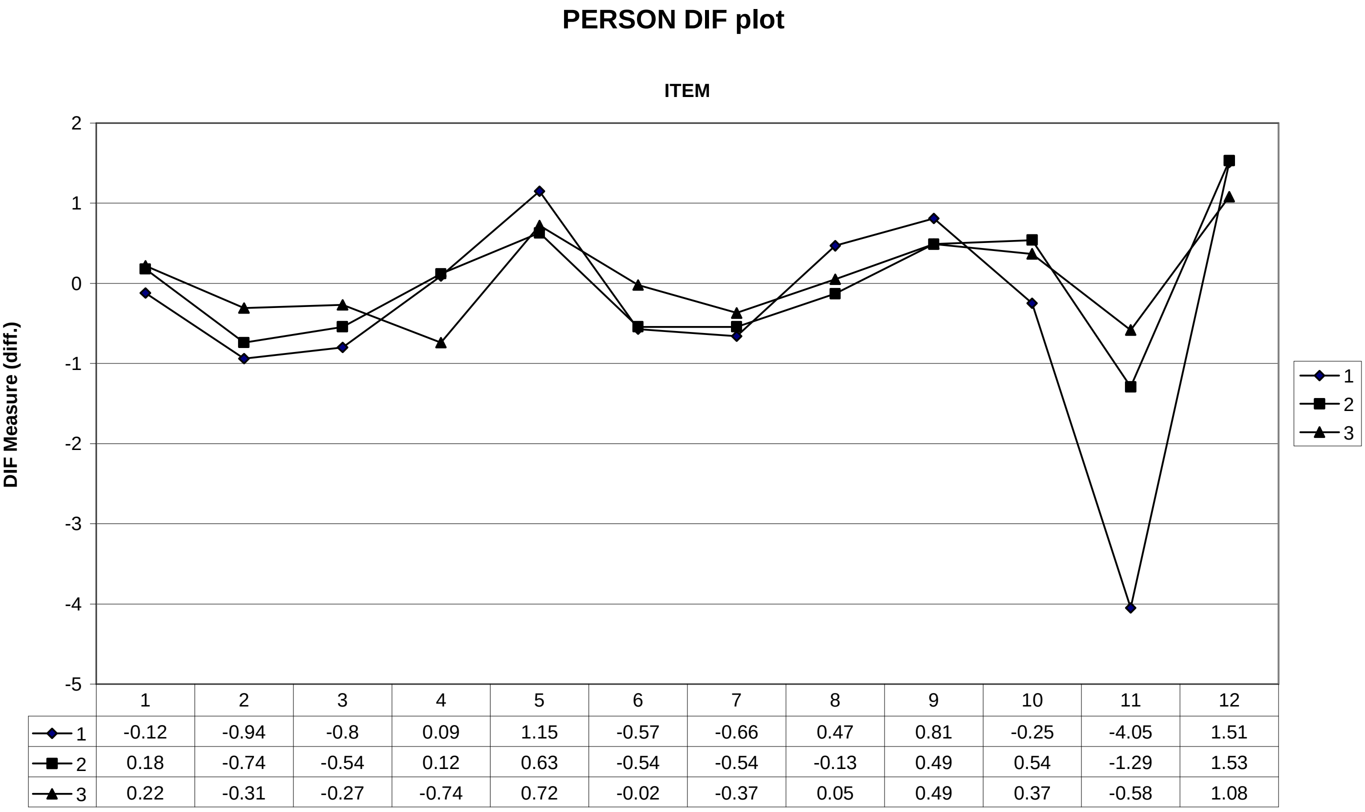
There were a number of minor violations against the ordering of the categories – mainly due to sparse data. This caused problems in the Rasch “structure calibration” – and is expected when, as in this case, we are faced with a weak scale.

Table II.Soc.2.2. Analysis of DIF between the groups IA; JCA and NCC

```
---------------------------------------------------------
| PERSON     SUMMARY DIF                ITEM            |
| CLASSES    CHI-SQUARE   D.F.  PROB.  Number Name      |
|-------------------------------------------------------|
|       3        1.0323      2  .5943       1 SCHOO4    |
|       3        2.2015      2  .3290       2 TEACH5    |
|       3        1.4554      2  .4797       3 FRIEND6   |
|       3        5.4873      2  .0630       4 GYMN7     |
|       3        3.5635      2  .1657       5 SHOW8     |
|       3        2.6850      2  .2578       6 BREAK9    |
|       3         .4436      2  .8013       7 ACTI10    |
|       3        3.1392      2  .2051       8 FRIEN28   |
|       3        1.4444      2  .4823       9 DECI29    |
|       3        4.6533      2  .0958      10 TEAS30    |
|       2*       4.3943      1  .0361      11 BULLY31   |
|       3        3.9067      2  .1394      12 TOGETH32  |
---------------------------------------------------------
```

* Group IA not included due to the same response score (no variance) for all children in the group. No systematic differences were found between IA, JCA and NCC regarding how the items are perceived by children.

Fig. II.Soc.2.1. Differences in estimated group measures, 1=IA, 2= JCA, 3= NCC



	1	2	3	4	5	6	7	8	9	10	11	12
◆ 1	-0.12	-0.94	-0.8	0.09	1.15	-0.57	-0.66	0.47	0.81	-0.25	-4.05	1.51
■ 2	0.18	-0.74	-0.54	0.12	0.63	-0.54	-0.54	-0.13	0.49	0.54	-1.29	1.53
▲ 3	0.22	-0.31	-0.27	-0.74	0.72	-0.02	-0.37	0.05	0.49	0.37	-0.58	1.08

The mean of the three Bully31 estimates in fig Fig. II.Soc.2.1. do not agree with the estimated measure in table II.Soc.2.2. due to the special estimates of the extremes in the IA group.

Even if the items seem to be adequate (item reliability), the person reliability is clearly insufficient, which to some extent can be explained by the lack of coverage, see Fig II.Soc.2.2.

Fig. II.Soc.2.2. RSM estimates. Social dimension, 87 persons, 12 items.

```
                         PERSON - MAP - ITEM
                        High score |<item difficulty>
   4                               +
                                   |
                                   |
                          2  3  3  |
                                   |
                                   |
                                   |
                                  T|
   3                   1  1  2  2  +
                                   |
                                   |
                                   |
                          2  3  3  |
                                   |
                       2  2  2  3 S|
                                   |
   2       1  1  1  1  2  2  2  3  +
                 1  1  2  2  2  3  |
                             1  3  |
           1  1  2  2  2  2  2  3  |
                    1  1  2  2  2 M|---------------- Person mean
           1  2  3  3  3  3  3  3  |T TOGETHER
     1  1  2  2  3  3  3  3  3  3  |
                 1  1  1  1  1  3  |
   1                      3  3  3  +
                       2  3  3  3  |  SHOW8
                       1  1  3  3 S|
                          2  2  3  |S DECI29
                                   |
                                2  |
                             1  2  |  TEAS30
                             1  3  |  FRIEN28   SCHOO4
   0                              T+M -------------------- Item mean
                                   |  GYMN7
                                   |
                                   |  BREAK9
                                   |
                                2  |S TEACH5
                                   |
                                   |
  -1                               +
                                   |
                                   |  BULLY31
                                   |T
                                   |
                                   |
                                   |
                                   |
  -2                               +
```

Each ‘number’ represent 1 child (1,2,3 indicate the grouping IA, JCA and NCC)
M= mean, S= 1 std, T= 2 std

Fig. II.Soc.2.2. clearly shows the discrepancy between the children, with a mean ≈ 1.5, and the questionnaire, for which the mean is set to zero. This implies poorer estimates in the upper person range. This is also clear from fig. II.Soc.2.3.,where the groups are separated.

Fig. II.Soc.2.3. Box plot of person measures in the three groups

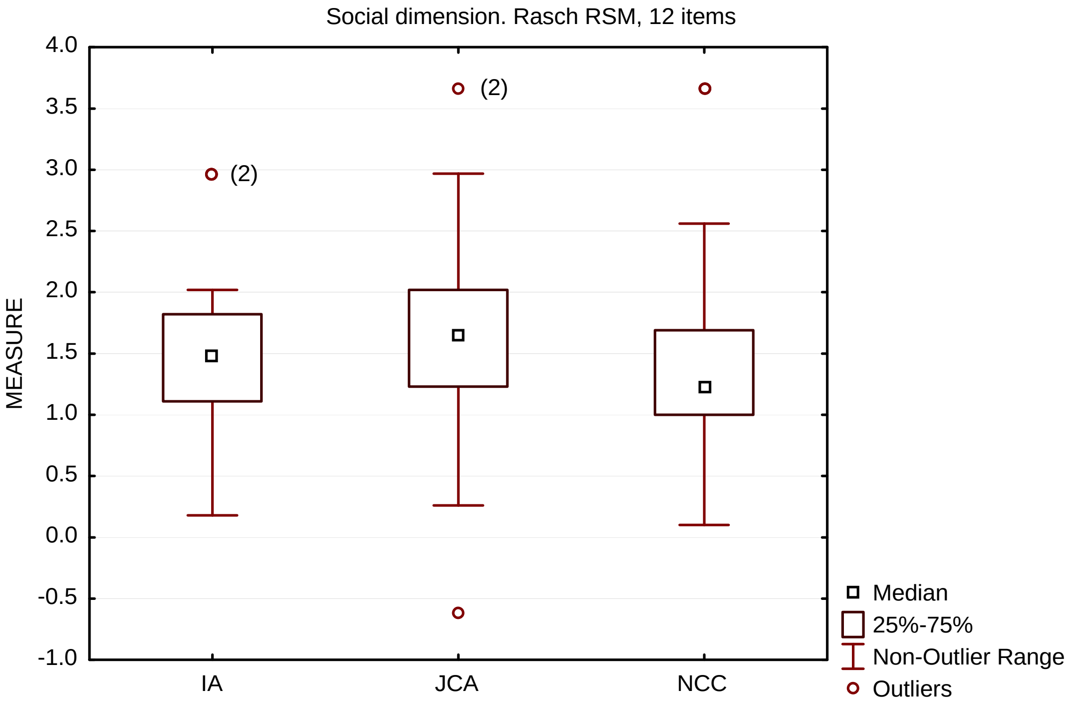
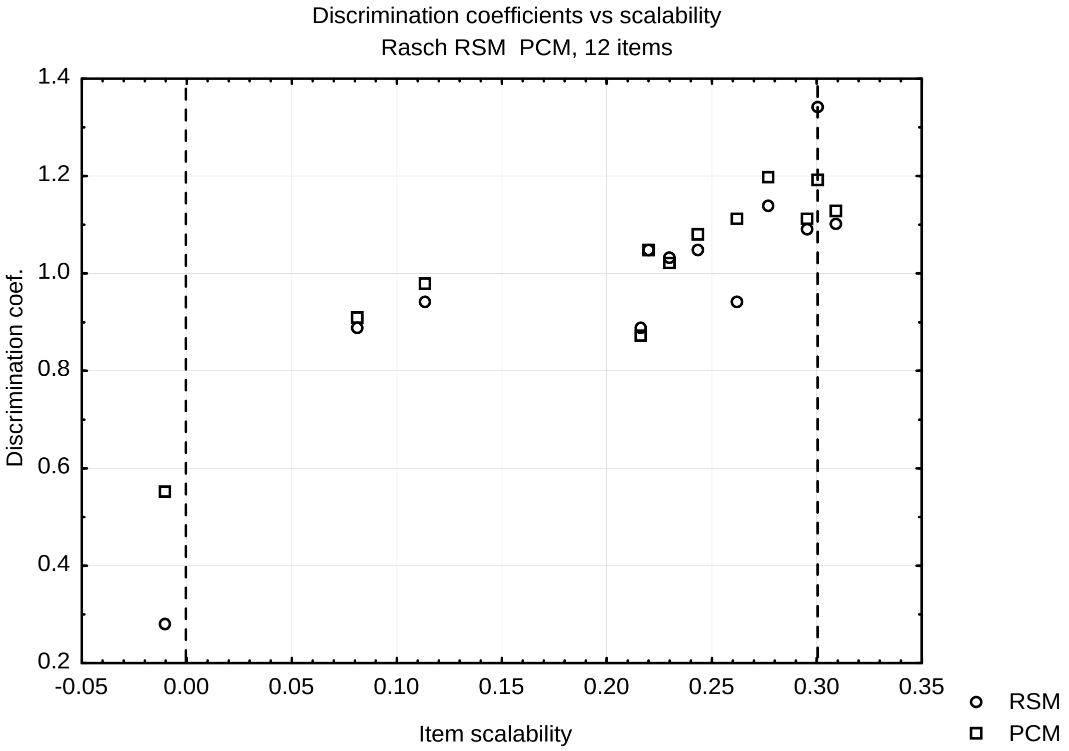


Fig. II.Soc.2.4. The Rasch RSM and the Rasch PCM vs the scalabilities, applied on the total set n=87



PCM is a model with item specific thresholds, while RSM has a common set of category thresholds. In fig. II.Soc.2.4., the scalabilities are calculated on n=82 complete answer profiles while PCM and RSM are based on the full set, n=87. Thus, the figure is a slight compromise but the effect of the different sample sizes has a negligible effect for our purpose.

Item scalabilities and item discriminations follow each other approximately. Step 1 and step2, summarized in fig. II.Soc.2.4., strongly indicate exclusion of item12 ‘Togeth32’ (to the extreme left in the figure).

An evaluation of 11 items is presented in the article. Some improvement was achieved but there were some local dependence identified.

Table II.Soc.2.3. Largest standardized residual (based on 11 items) correlations used to identify dependent item

```
-----------------------------------
|CORREL-|            |            |
|  ATION|     ITEM   |     ITEM   |
|-------+------------+------------|
|   .34 |     SCHOO4 |     TEACH5 |
|   .29 |     TEAS30 |     BULLY31|
|-------+------------+------------|
|  -.40 |     SCHOO4 |     DECI29 |
|  -.35 |     SCHOO4 |     SHOW8  |
-----------------------------------
```

There are no strikingly strong residual correlations. We have to accept moderate correlations in view of the small sample sizes.

Looking at the NCC group alone, ‘How do you like school?’ and ‘How is your relation with the teacher?’ showed higher correlation (r= 0.50), as well as ‘Have you been teased at school?’ and ‘Have you been bullied at school?’ (r=0.37). Conclusions about these correlations are hazardous but, together with the weak scalability, it might be a sign that the questionnaire does not represent one main dimension for this group. However, the locally dependent items did not share the same position on the scale.

Step 3 Soc

The weak scale, found in Step 1, with very low scalabilities and the insufficient coverage found in Step 2, indicates that there will be difficulties with an extended parametric model. No further useful gain can be achieved due to numerical instability in the estimation process.

Conclusion about the ‘Social’ questionnaire (Soc)

Step 1 indicates a weak scale and the questionnaire seems not very suitable for the NCC children. The scalability for the item set says that it will be difficult to rank the children or form a person measure of any reasonable precision. The latent dimension may be difficult to identify for the intended populations.
Taking a step further (Step2) by applying a Rasch model confirms the result from Step 1. The parsimonious RSM approves the item set as a whole (besides one variable) but the set is not sufficient to form a reliable person measure. Furthermore, the coverage is insufficient. The residual correlation indicates that certain items should be reformulated and more items should be added in order to reach a wider distribution of the item locations for a better coverage.
This is particularly needed to catch NCC children if the intention still is to create a questionnaire equally appropriate for the three populations as classified.

- The questionnaire forms a good start but has to be further developed.
- More items are needed to catch the social dimension for these children.
- TOGETH32 and TEACH5 might be redundant, or variables with negligible contribution.

- Investigation of the ‘Psychological’ questionnaire (Psych)

The psychological questionnaire consists of 23 items. n=71 out of 87 questionnaires were complete. The observed scores from the 87 children are shown in table II.Psych.1.1.

Step1 Psych. **The Mokken scale analysis.**

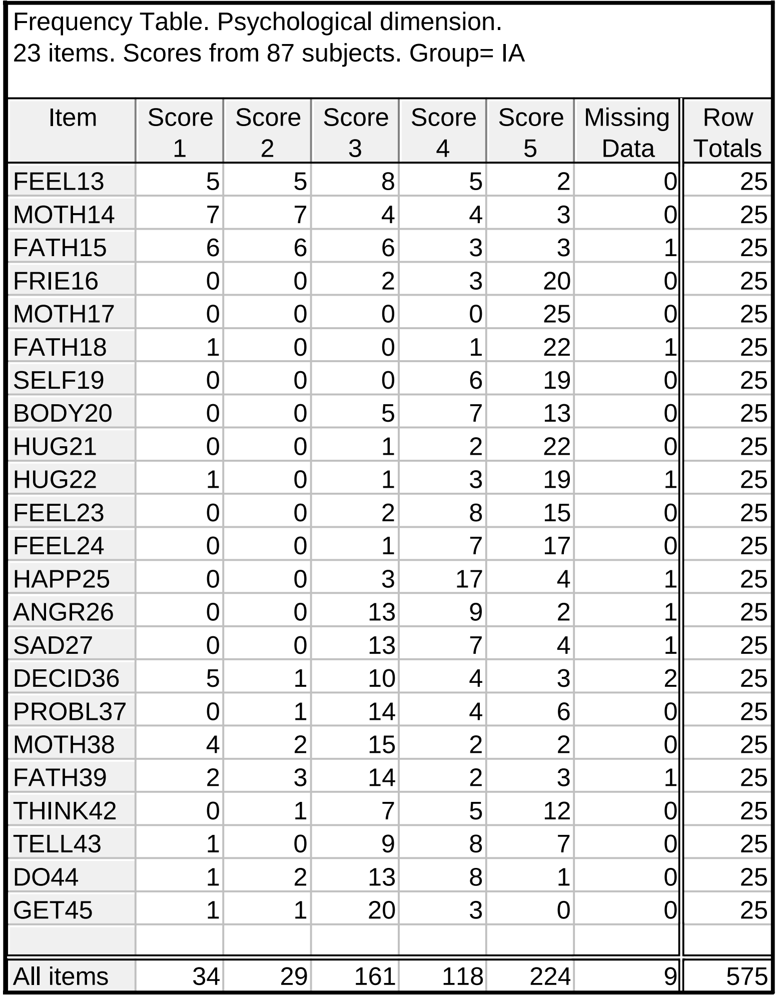
Table II.Psych.1.1. Answers from the total set, n=87

Frequency Table. The psychological dimension.
23 items. Scores from 87 subjects

Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
FEEL13	16	19	27	18	6	1	87
MOTH14	24	24	20	12	7	0	87
FATH15	19	21	25	10	8	4	87
FRIE16	0	0	4	25	58	0	87
MOTH17	0	0	1	1	85	0	87
FATH18	1	0	1	5	77	3	87
SELF19	0	0	5	19	62	1	87
BODY20	3	1	12	24	47	0	87
HUG21	2	1	6	12	66	0	87
HUG22	1	2	8	19	54	3	87
FEEL23	0	1	8	31	47	0	87
FEEL24	0	1	8	25	53	0	87
HAPP25	0	0	14	57	14	2	87
ANGR26	2	0	48	31	4	2	87
SAD27	1	1	45	30	8	2	87
DECID36	11	2	22	24	25	3	87
PROBL37	6	6	41	20	14	0	87
MOTH38	18	12	40	11	4	2	87
FATH39	13	9	37	12	8	8	87
THINK42	3	3	34	22	24	1	87
TELL43	1	2	20	38	25	1	87
DO44	1	4	51	27	3	1	87
GET45	2	5	67	12	0	1	87
All items	124	114	544	485	699	35	2001

About 85% of the answers fall in categories 3 or 5, while, for certain items, some of categories were not scored by any child. A number of ‘non responses’ is also recognised in spite of an interview directed questionnaire.
The same structure appears when we look into the groups.

Table II.Psych.1.2. Score frequencies within group IA



Frequency Table. Psychological dimension.
23 items. Scores from 87 subjects. Group= IA

Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
FEEL13	5	5	8	5	2	0	25
MOTH14	7	7	4	4	3	0	25
FATH15	6	6	6	3	3	1	25
FRIE16	0	0	2	3	20	0	25
MOTH17	0	0	0	0	25	0	25
FATH18	1	0	0	1	22	1	25
SELF19	0	0	0	6	19	0	25
BODY20	0	0	5	7	13	0	25
HUG21	0	0	1	2	22	0	25
HUG22	1	0	1	3	19	1	25
FEEL23	0	0	2	8	15	0	25
FEEL24	0	0	1	7	17	0	25
HAPP25	0	0	3	17	4	1	25
ANGR26	0	0	13	9	2	1	25
SAD27	0	0	13	7	4	1	25
DECID36	5	1	10	4	3	2	25
PROBL37	0	1	14	4	6	0	25
MOTH38	4	2	15	2	2	0	25
FATH39	2	3	14	2	3	1	25
THINK42	0	1	7	5	12	0	25
TELL43	1	0	9	8	7	0	25
DO44	1	2	13	8	1	0	25
GET45	1	1	20	3	0	0	25
All items	34	29	161	118	224	9	575

For the IA group (see table II.Psych.1.2.), the answers on items FRIE16 – FEEL24 and THINK42 are concentrated to score 5, while the rest of the items have answer more scattered or gathered around the centre.

Table II.Psych.1.3. Score frequencies within group JCA

Frequency Table. Psychological dimension.
23 items. Scores from 87 subjects. Group= JCA

Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
FEEL13	6	8	6	7	2	1	30
MOTH14	9	10	5	5	1	0	30
FATH15	9	8	8	4	1	0	30
FRIE16	0	0	1	7	22	0	30
MOTH17	0	0	1	0	29	0	30
FATH18	0	0	1	1	28	0	30
SELF19	0	0	2	5	23	0	30
BODY20	2	1	2	9	16	0	30
HUG21	0	1	0	6	23	0	30
HUG22	0	2	0	8	20	0	30
FEEL23	0	1	3	11	15	0	30
FEEL24	0	1	1	7	21	0	30
HAPP25	0	0	7	19	4	0	30
ANGR26	2	0	17	10	1	0	30
SAD27	1	0	18	9	2	0	30
DECID36	4	0	6	10	10	0	30
PROBL37	2	1	14	9	4	0	30
MOTH38	6	3	16	4	0	1	30
FATH39	7	1	14	4	1	3	30
THINK42	2	0	13	7	8	0	30
TELL43	0	2	7	13	8	0	30
DO44	0	1	15	12	2	0	30
GET45	0	3	23	4	0	0	30
All items	50	43	180	171	241	5	690

In table II.Psych.1.3. the JCA group shows approximately the same answer distribution as does group IA.

Table II.Psych.1.4. Score frequencies within group NCC

Frequency Table. Psychological dimension.
23 items. Scores from 87 subjects. Group= NCC

Item	Score 1	Score 2	Score 3	Score 4	Score 5	Missing Data	Row Totals
FEEL13	5	6	13	6	2	0	32
MOTH14	8	7	11	3	3	0	32
FATH15	4	7	11	3	4	3	32
FRIE16	0	0	1	15	16	0	32
MOTH17	0	0	0	1	31	0	32
FATH18	0	0	0	3	27	2	32
SELF19	0	0	3	8	20	1	32
BODY20	1	0	5	8	18	0	32
HUG21	2	0	5	4	21	0	32
HUG22	0	0	7	8	15	2	32
FEEL23	0	0	3	12	17	0	32
FEEL24	0	0	6	11	15	0	32
HAPP25	0	0	4	21	6	1	32
ANGR26	0	0	18	12	1	1	32
SAD27	0	1	14	14	2	1	32
DECID36	2	1	6	10	12	1	32
PROBL37	4	4	13	7	4	0	32
MOTH38	8	7	9	5	2	1	32
FATH39	4	5	9	6	4	4	32
THINK42	1	2	14	10	4	1	32
TELL43	0	0	4	17	10	1	32
DO44	0	1	23	7	0	1	32
GET45	1	1	24	5	0	1	32
All items	40	42	203	196	234	21	736

The answer distribution from the NCC group is similar to that of the IA and the JCA group.
From tables II.Psych.1.2-4. it is clear that we are confronted with many empty cells. On the whole, the questionnaire seems to be too “easy” for the purpose. All groups have 85 – 90% of the answers in category 3, 4 or 5.For some items (MOTH17, FATH18) there were virtually no variation at all. A first thought is that the small sample size is the reason, but there is a possibility that some categories would remain virtually empty irrespective of the sample size.

Before an establishment of the model $\theta_{n(i)} = C + G_i + \theta_n|G$, which says the questionnaire works equally well for the three types of children, the analyses have to be carried out within groups in order to keep the grouping factor outside. However, the total scalability (IA+JCA+NCC) is also of interest.

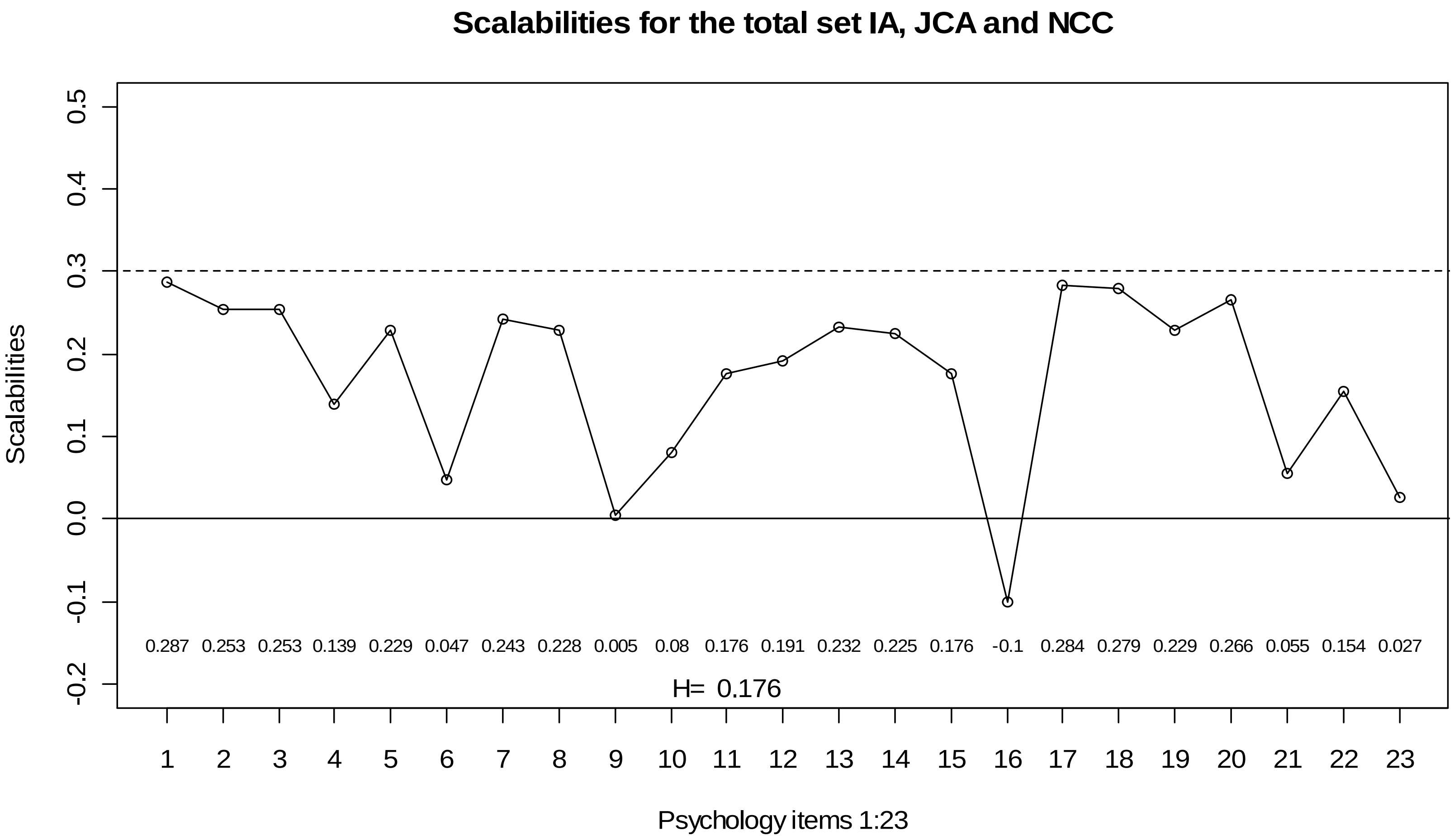
Table II.Psych.1.5. Item scalabilities within groups.

	Item_nr	all	IA	JCA	NCC
FEEL13	1	0.287	0.289	0.395	0.183
MOTH14	2	0.253	0.314	0.374	0.070
FATH15	3	0.253	0.305	0.394	0.072
FRIE16	4	0.139	0.281	0.296	-0.120
MOTH17	5	0.229	--	0.459	-0.257
FATH18	6	0.047	-0.069	0.278	0.057
SELF19	7	0.243	0.164	0.342	0.224
BODY20	8	0.228	0.313	0.219	0.225
HUG21	9	0.005	0.295	-0.162	0.019
HUG22	10	0.080	0.072	0.033	0.169
FEEL23	11	0.176	0.193	0.235	0.075
FEEL24	12	0.191	0.311	0.191	0.161
HAPP25	13	0.232	0.403	0.163	0.154
ANGR26	14	0.225	0.371	0.371	-0.185
SAD27	15	0.176	0.174	0.330	-0.064
DECID36	**16**	**-0.100**	**-0.180**	**-0.180**	**0.152**
PROBL37	17	0.284	0.274	0.403	0.228
MOTH38	18	0.279	0.284	0.411	0.187
FATH39	19	0.229	0.291	0.405	0.045
THINK42	20	0.266	0.237	0.317	0.242
TELL43	21	0.055	0.244	-0.115	0.101
DO44	22	0.154	0.287	0.161	0.029
GET45	23	0.027	0.175	-0.055	-0.036

```
The range of the item set Hi= [ -0.10,0.29], i=1,…,12
The item set scalability H= 0.176  n= 71

The range of the item set Hi within groups:

set1 IA:    Hi= [ -0.18,0.49] H= 0.223  n= 21
set2 JCA:   Hi= [ -0.18,0.46] H= 0.231  n= 26
set1 NCC:   Hi= [ -0.26,0.24] H= 0.107  n= 24
```

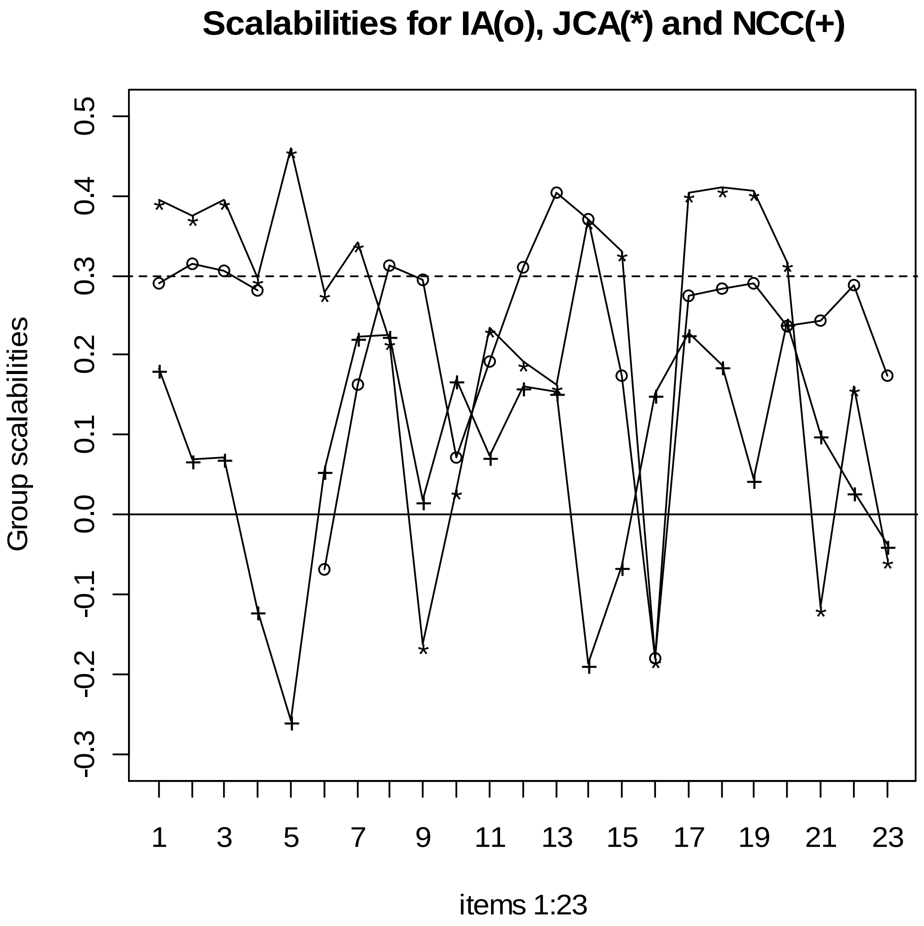
From table II.Psych.1.5 it is immediately clear that we have a weak scale with a set of items having a scalability <0.1 and with some of the group scalabilities <0. Furthermore, the split in group yields very small the sample sizes, which makes the evaluation more difficult. Like the social dimension, the NCC group seems not to perceive the questionnaire very well. The awkward situation is well illustrated in fig. II.Psych.1.1.

Fig. II.Psych.1.1.



Based on table II.Psych.1.5. and fig. II.Psych.1.2., there is a strong indication that item16 should be excluded.

Fig. II.Psych.1.2.



Once again, the questionnaire does not seem suitable for the (heterogeneous) NCC group – five negative scalabilities.

It is hard to discern any common structure for the group scalabilities showed in fig. II.Psych.1.2., but item 16 (DECID36) can be said to be the worst.

Table II.Psych.1.6. Item scalabilities within groups. Item 16 deleted

```
        Item nr   all     IA    JCA    NCC
FEEL13        1 0.313  0.326  0.455  0.156
MOTH14        2 0.276  0.363  0.422  0.045
FATH15        3 0.280  0.353  0.455  0.055
FRIE16        4 0.149  0.281  0.304 -0.122
MOTH17        5 0.244    --   0.474 -0.254
FATH18        6 0.057 -0.023  0.290  0.022
SELF19        7 0.266  0.192  0.371  0.233
BODY20        8 0.249  0.346  0.242  0.246
HUG21         9 0.025  0.326 -0.164  0.032
HUG22        10 0.104  0.121  0.063  0.153
FEEL23       11 0.210  0.211  0.293  0.084
FEEL24       12 0.222  0.341  0.220  0.179
HAPP25       13 0.257  0.437  0.203  0.154
ANGR26       14 0.262  0.445  0.410 -0.179
SAD27        15 0.198  0.212  0.357 -0.066
PROBL37      17 0.326  0.352  0.459  0.223
MOTH38       18 0.318  0.350  0.463  0.191
FATH39       19 0.260  0.327  0.460  0.049
THINK42      20 0.298  0.243  0.380  0.240
TELL43       21 0.063  0.279 -0.100  0.080
DO44         22 0.151  0.292  0.155  0.019
GET45        23 0.031  0.208 -0.060 -0.043

Item set H= 0.222
```

Child ID= 11(IA) and 81(NCC) were identified as incoherent profiles with the jackknife method.
ID= 118(NCC) was further identified (in Step 2) as a profile with a large fit statistic. This illustrates the possibility to go back and forth between steps, which might yield some advantages, although usually not necessary.

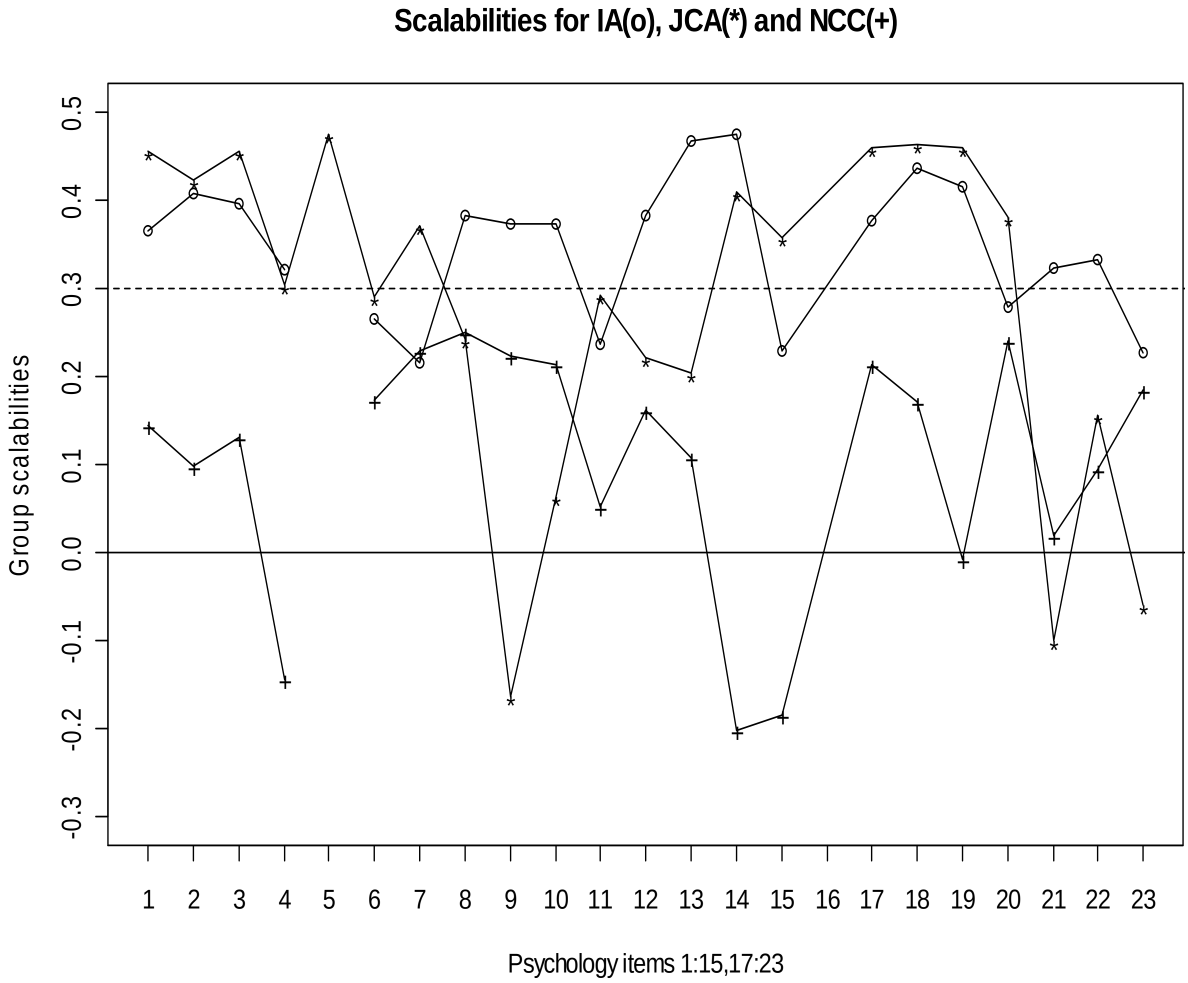
When these children were put aside the item set scalability raised to H= 0.252.
Bootstrapping a 90% C.I. yielded (0.178, 0.330).

Table II.Psych.1.7. Item scalabilities based on 68 children with complete responses.

```
        Item nr   all    IA    JCA    NCC
FEEL13        1 0.330 0.366  0.455  0.144
MOTH14        2 0.307 0.408  0.422  0.097
FATH15        3 0.313 0.396  0.455  0.131
FRIE16        4 0.166 0.320  0.304 -0.145
MOTH17        5 0.452   --   0.474    --
FATH18        6 0.253 0.266  0.290  0.172
SELF19        7 0.275 0.215  0.371  0.229
BODY20        8 0.259 0.383  0.242  0.249
HUG21         9 0.114 0.373 -0.164  0.222
HUG22        10 0.176 0.373  0.063  0.213
FEEL23       11 0.213 0.237  0.293  0.051
FEEL24       12 0.228 0.383  0.220  0.161
HAPP25       13 0.261 0.467  0.203  0.108
ANGR26       14 0.281 0.475  0.410 -0.203
SAD27        15 0.197 0.229  0.357 -0.185
PROBL37      17 0.342 0.376  0.459  0.213
MOTH38       18 0.348 0.436  0.463  0.170
FATH39       19 0.274 0.416  0.460 -0.007
THINK42      20 0.314 0.279  0.380  0.241
TELL43       21 0.050 0.323 -0.100  0.019
DO44         22 0.182 0.332  0.155  0.094
GET45        23 0.083 0.226 -0.060  0.184

H(Group)              0.356  0.301  0.118
Item set scalability H= 0.252
```

Exclusion of item16 and some influential children improve the situation (the feasibility of the questionnaire) to some extent but did not change the structure of the psychological dimension as a weak scale, particularly for the NCC group.

Fig. II.Psych.1.3. Item scalabilities for the three groups based on n= 68 and item 16 deleted



Analysis of a systematic difference between H(IA + JCA) and H(NCC) by bootstrapping

H(IA+JCA)= 0.325.   H(NCC)= 0.118. Items MOTH17 and DECID36 as well as the children 11, 81,118 are excluded.
Bootstrapping 500 comparisons yields a 90% C.I. [0.065, 0.35], BCa: [0.0596, 0.347].
(Bootstrapping 1000 comparisons yields a 90% C.I. [0.077, 0.346], BCa: [0.0562, 0.338].)

As for the social dimension, there is a strong indication that the questionnaire is not equally suitable for the three groups. It seems not very suitable for the NCC group. Further items are needed to capture the NCC children.
Due to the low scalability, it might be questioned whether the questionnaire is capable in ranking the respondents on the psychological dimension.

Monotonicity and non-intersection

As the scale is weak, we do not expect to see clear violations against monotonicity and non-intersection. Some violations were detected but not of a magnitude motivating any special measures.

Large item pair scalabilities
Items with statements about Mother and Father frequently show pair scalabilities >0.9. These are the most conspicuous. Furthermore, H(HUG21,HUG22)= 0.741 and H(ANG26,SAD27)=0.784. Much of the observed relationship is due to the concentration of answer to a few categories, but obviously there are item pairs, which for natural reasons, are strongly related.

<u>Step 2 Psych.</u> Analysis by a Rasch RSM

An introductory Rasch analysis confirmed the children ID= 11, 81 and 118 as incoherent, based on a person fit statistic >5.

There is an unfortunate mistake in the article concerning the Rasch model. It is said that the parsimonious ‘model a‘ is used. However, what is actually presented in table 1 in the article is a PCM model with item specific thresholds which should be considered as too data driven. The weak scale and the somewhat different behaviour of the three groups, found in Step 1, indicates that we probably do not get any decisive conclusion from a particular Rasch model applied in Step 2. There are three reasonable approaches:

1. A parsimonious model with a common set of category thresholds. Model 1.
2. A ‘mixed’ model with a main set of common thresholds and item specific thresholds for a small group of items. Model 2
3. A ‘free’ model, with item specific thresholds for all items. Model 3.

Table II.Psych.2.1. The Rasch model with a common scale and ID= 11, 81 and 118 excluded.

	Model 1			Model 2			Model 3		
	location	Mnsq	Zstd	location	Mnsq	Zstd	location	Mnsq	Zstd
FEEL13	1.91	0.99	0	1.68	0.93	-0.5	1.26	0.81	-1.4
MOTH14	2.22	1.16	1.2	1.98	1.1	0.8	1.33	0.9	-0.7
FATH15	2.06	1.14	1	1.82	1.07	0.6	1.2	0.9	-0.6
FRIE16	-1.44	0.99	0	-1.58	0.99	0	-0.78	1.08	0.5
MOTH17	-4.38	1.84	1.2	**<u>-1.51</u>**	0.83	0.1	-1.86	0.84	0.1
FATH18	-3.24	1.15	0.5	**<u>-1.45</u>**	0.91	0	-1.79	0.92	0
SELF19	-1.54	0.88	-0.6	-1.68	0.89	-0.5	-0.7	0.86	-0.7
BODY20	-0.55	1.49	3	-0.69	1.48	2.9	-0.31	0.98	0
HUG21	-1.55	1.88	3.7	-1.7	1.89	3.8	-0.86	1.44	1.5
HUG22	-1.15	1.4	2.2	-1.29	1.41	2.2	-0.69	1.2	1
FEEL23	-0.86	0.95	-0.2	-1	0.96	-0.2	-0.81	1	0
FEEL24	-1.03	1.03	0.3	-1.17	1.04	0.3	-0.88	0.97	-0.1
HAPP25	-0.02	0.45	-4.7	-0.16	0.45	-4.6	0.93	0.9	-0.6
ANGR26	0.91	0.6	-2.7	0.73	0.59	-2.9	0.53	0.93	-0.3
SAD27	0.82	0.7	-2	0.64	0.68	-2.1	0.17	1.02	0.2
DECID36	deleted			deleted			deleted		
PROBL37	1.08	1.04	0.3	0.89	0.96	-0.2	0.55	0.9	-0.6
MOTH38	2.03	0.85	-1.1	1.79	0.81	-1.4	1.61	0.89	-0.8
FATH39	1.74	1.12	0.8	1.51	1.05	0.4	1.14	1.03	0.2
THINK42	0.48	1.19	1.2	0.33	1.14	0.9	0.06	0.94	-0.3
TELL43	0.04	1.32	2	-0.11	1.3	1.9	-0.32	1.39	2.1
DO44	1.05	0.65	-2.3	0.87	0.63	-2.5	0.47	1.12	0.7
GET45	1.42	0.59	-2.9	**<u>0.08</u>**	1.14	0.6	-0.25	1.15	0.7
Pers.rel.	0.81			0.82			0.83		
Item rel.	0.97			0.97			0.95		

Locations for free items in model 2 are marked in bold and underlined.

Table II.Psych.2.2. Largest residual correlations

		Model 1		Model 2		Model 3
	r	items	r	items	r	items
Residual	0.88	MOTH14 FATH15	**0.87**	**MOTH14 FATH15**	0.87	MOTH14 FATH15
correlations	0.76	MOTH38 FATH39	**0.75**	**MOTH38 FATH39**	0.75	HUG21 HUG22
	0.71	HUG21 HUG22	**0.71**	**HUG21 HUG22**	0.75	MOTH38 FATH39
	0.58	ANGR26 SAD27	0.58	ANGR26 SAD27	0.58	ANGR26 SAD27

All three models yield virtually the same message. Table II.Psych.2.2. strongly indicates that duplicating the same question related to both mother and father does not add any valuable information. The same can be said about HUG21 and HUG22 as well as ANGR26 and SAD27. A Person-Item map based on the parsimonious model 1 is structurally similar to that based on model 3 (shown in the article), with FATH18 and MOTH17 pushed even further to the low extreme. Fig. II.Psych.2.2.also tells that the questionnaire is too ‘easy’.

Relying on the recommendations for Rasch models [Linacre J. 2008], items with a MNSQ in the range 0,5 – 1.5 are considered productive.

Even if the evaluation is not focused on person estimates, a person reliability = 0.81 is just moderate. This can be illustrated by looking at person C.I.. The s.e of a person estimate is appr. 0.3. Then, for a subject with estimated measure= 1, the 95% C.I.[0.4, 1.6] covers 55(65%) of the 84 subjects, which might be seen as a low precision (see fig. II.Psych.2.2.). The item reliability, 0.95, tells us that the items are relevant.

Ordering of the categories

There were some minor violations due to sparse data. This caused problems in the Rasch “structure calibration” for category 2 and 3. This is not surprising as there are a set of weak items, particular as perceived by the NCC group.

Different Item Functioning (DIF)

Analysis of DIF points to a different conception of the questionnaire regarding the three groups IA, JCC and NCC. An aggregated chi-square gives a p-value of about 0.025 for all models. This chi-square test is doubtful as the items are far from locally independent. Analysis of DIF, transformed to t-values and illustrated in fig. II.Psych.2.1., yielded just a few ’signifcant DIF:s in the range 0.01< p < 0.05. However, the NCC group (group 3 in fig. II.Psych.2.1.) behaves quite differently compared to the IA and the JCA groups. On the whole, the message from Step 1 was confirmed – the NCC group does not perceive the questionnaire as the other two groups do.

Fig. II.Psych.2.1. DIF, transformed to t-values

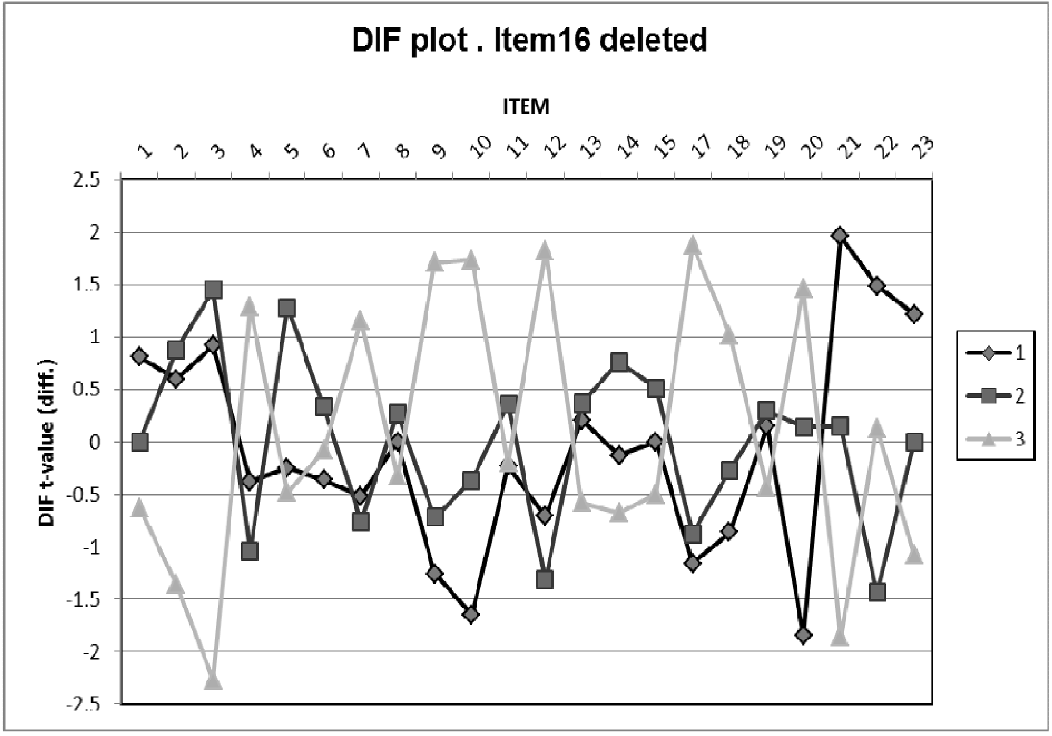


Fig. II.Psych.2.1. can be inspected visually. T-values with large differences in the vertical direction for an item indicate DIF. Item 2,3,9,10,17,20 and 21 show quiet large DIF:s. With the small sample sizes in mind it is hard to say whether there are real DIF:s or an effect of heterogeneity.

Fig. II.Psych.2.2. Person and item locations on the latent scale, based on model 1 (Table II.Psych.2.1.)

```
                            PERSONS - MAP - ITEMS
The latent scale.   Better functioning | Difficult to score high
    5                                  +
                                       |
                                       |
                                    1  |
                                       |
    4                                  +
                                       |
                                       |
                                    2  |T
                                    1  |
    3                               3 T+
                                    2  |
                                  1 3  |
                                    2  |
                              1 1 2 3 S|  MOTH14
    2                         1 1 2 2  +  FATH15   FEEL13   MOTH38
                              1 1 2 3  |S FATH39
                          2 2 2 2 3 3  |                                   -
      1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 3 M|  GET45                            |
                      1 1 2 2 2 3 3 3  |                                   |
    1   1 1 1 1 2 2 2 3 3 3 3 3 3 3 3  +  ANGR26   DO44     PROBL37         +
                      1 1 1 2 2 2 2 3  |  SAD27                            |
                              1 3 3 3 S|                                   |
                                1 2 3  |  THINK42                          -
                                    3  |                               95% C.I.
    0                               2  +M HAPP25   TELL43              measure= 1
                                      T|
                                  2 2  |
                                       |  BODY20
                                       |  FEEL23
   -1                                  +  FEEL24
                                       |  HUG22
                                       |  FRIE16
                                       |  HUG21    SELF19
                                       |S
   -2                                  +
                                       |
                                       |
                                       |
                                       |
   -3                                  +
                                       |  FATH18
                                       |T
                                       |
                                       |
   -4                                  +
                                       |
                                       |  MOTH17
                                       |
                                       |
   -5                                  +
                      Less functioning | Easy to score high
```

Step 3 Psych

The weak scale, identified in Step 1, and the analysis in step 2, does not indicate further information from a Step 3 approach. Applying an extended parametric model will cause convergence problems, due the asymmetric structure a large amount of empty categories. It might be cured by dramatically reducing the number of categories, but this will estrange us from the original pragmatic starting-point.

Conclusion about the 'Psychological' questionnaire (Psych)

The analysis strongly suggests that the item DECID36 has no relevance in the questionnaire. Furthermore, items MOTH17and FATH18 turned out to be non-productive items. They should be deleted or reformulated.
A small set of items (see table II.Psych.2.2.) have large residual correlations, which indicates a large amount of repeated information. Exclude MOTH14 orFATH15, MOTH38 or FATH39 and HUG21 or HUG22 in favour of some new productive items.
HAPP25 and GET45 might gain to be reformulated to fit in a parsimonious approach of a straightforward questionnaire, aimed for creating a relevant sum score measure of the children's position on the intended psychological dimension.
Although the items are considered relevant in Step 2, with an item reliability= 0.95, they are insufficient to reasonably capture the subjects on the intended scale, person reliability= 0.81. However, when looking in detail, the situation becomes more cumbrous.

From Step 1, as well as from Step 2, it is clear (without any decided statistical demonstration) that the NCC group does not behave, or perceive the questionnaire, as the IA and the JCA groups. A number of items (8 out of 23 items with a scalability < 0.1) does not seem to be very useful. It will probably be a laborious, but necessary, task to develop the questionnaire in such a way that it will reasonably catch the three target populations. This seems a great deal of worry for capturing the NCC children. In general, the items should, if possible, be made more 'difficult' to score high in order to improve the coverage.

Identification of the Psychological and Social dimensions

Verification the two dimensions were not possible due to the low scalabilities and the possibility that the dimensions could well be differently perceived by the three groups. An attempt with AISP, within the Mokken approach, left about half of the items outside any dimension for the IA group (16/31), a third for the CA group (16/31) and almost all for the NCC group (32/33). A few variables had to be deleted due to no variation.

Conclusion about the IAPSQ questionnaire

The questionnaire is constructed to work equally well for the three patient groups, **IA**, **JCA** and **NCC**. It can be concluded that the questionnaire is too 'easy' to capture the intended social and psychological dimensions. If the samples had been larger, an analysis to verify the two dimensions would be more informative. With a split into three specified groups, resulting in very small sample sizes, such an analysis is too hazardous. Furthermore, the generally low item scalabilities and an indication of different perception of the questionnaire regarding the three groups make the decisions rather vague. However, some advisory general decisions might be drawn.

- The questionnaire is ‘too easy’ with insufficient coverage in the upper half. A set of items have virtually no answers at the lower levels. These items should be reformulated or replaced.

- Even if the items can be considered relevant, at least to some extent, their capability to capture the intended dimension is very limited. If the authors are convinced that the Social and Psychological dimension are possible to define for the target populations, more items are needed.

- A further development of the questionnaire should give special attention to group specific (IA, JCA, NCC) characteristics in further analyses in order to investigate the possibility to create a questionnaire, able to work equally well for intended populations.

Study III

Brodin U., Gunilla M Olsson G.M., (2010). **Adolescent Adjustment Profile - revised and investigated by means of an Item Response Theory approach.** (Manuscript).

Psychosocial development in children with chronic disease is a key issue in paediatrics. A recent study investigated whether psychosocial adjustment could be reliably assessed with a 42-item Adolescent Adjustment Profile (AAP) instrument [Olson et al, 2008].. Psychosocial adjustment was measured using the AAP, which contains four domains; Attention deficit, Social competence, Externalising behaviour and Internalising behaviour. The questionnaire was developed with the sum score of ordered questions for each domain as the outcome variable. Even if the items in the questionnaire are well formulated, it seems as a proper evaluation of their characteristics is lacking. This is the reason why this study is focused on the evaluation of the measuring properties, the characteristics of the questionnaire, of three of these domains (Social competence excluded) by use of IRT analyses within the '3 step strategy' framework. A reference group of 131 healthy Swedish adolescents formed the sample for this study.
The questionnaires are outlined in Appendix A.

The manuscript presents the essential results from the analyses of the three domains. Some further considerations, which could (we could not find room for) not be presented within the space limits, would be of interest for further insight in the questionnaire.

The second index in tables and figures represents the three intended latent traits
Attention deficit = AttDef
Externalising behaviour = Ext
Internalising behaviour = Int
A third index =0 indicates a general presentation of the material before details about the dimensions.

Presentation of observed data and an introductory dimensionality analysis

The answers from the 131 adolescents, concerning the three latent dimensions, are presented in the tables Table III.AttDef.0.1, Table III.Ext.0.1 and Table III.Int.0.1.

Table III.AttDef.0.1

Frequency Table. Attention Deficit dimension
13 items scored by 131 subjects.

gender	Item	Score 1	Score 2	Score 3	Score 4	Score 5	Row Totals
Girls	AttDef1	8	30	16	9	2	65
Girls	AttDef2	6	32	23	3	1	65
Girls	AttDef3	13	36	7	9	0	65
Girls	AttDef4	15	24	18	7	1	65
Girls	AttDef5	3	30	29	3	0	65
Girls	AttDef6	6	20	29	9	1	65
Girls	AttDef7	2	29	25	8	1	65
Girls	AttDef8	9	30	18	7	1	65
Girls	AttDef9	8	28	20	6	3	65
Girls	AttDef10	12	29	16	6	2	65
Girls	AttDef11	5	31	16	8	5	65
Girls	AttDef12	18	26	14	6	1	65
Girls	AttDef13	7	31	14	8	5	65
Total		112	376	245	89	23	845
Boys	AttDef1	6	39	17	4	0	66
Boys	AttDef2	7	28	27	4	0	66
Boys	AttDef3	14	37	13	2	0	66
Boys	AttDef4	13	28	19	6	0	66
Boys	AttDef5	2	31	27	6	0	66
Boys	AttDef6	2	20	28	15	1	66
Boys	AttDef7	4	27	34	0	1	66
Boys	AttDef8	10	34	19	3	0	66
Boys	AttDef9	7	30	23	6	0	66
Boys	AttDef10	13	30	20	3	0	66
Boys	AttDef11	10	30	22	4	0	66
Boys	AttDef12	17	31	17	1	0	66
Boys	AttDef13	16	30	10	8	2	66
Total		121	395	276	62	4	858
Column Total		233	771	521	151	27	1703

The distributions of the scores, as in table III.AttDef.0.1, are fairly equal for girls and boys, with a concentration towards the middle and lower end of the category scale. Score 2, and 3 represent about 75%. Score 5 is sparsely endorsed with < 2% of the answers.

Table III.Ext.0.2

Frequency Table. Externalising behaviour dimension.
11 items scored by 131 subjects.

Gender	Item	score 1	score 2	score 3	score 4	score 5	Row Totals
Girls	Ext1	0	1	8	28	28	65
Girls	Ext2	0	2	13	28	22	65
Girls	Ext3	0	5	3	22	35	65
Girls	Ext4	5	0	7	11	42	65
Girls	Ext5	2	5	12	26	20	65
Girls	Ext6	2	3	11	32	17	65
Girls	Ext7	1	8	32	21	3	65
Girls	Ext8	1	2	9	20	33	65
Girls	Ext9	3	14	19	15	14	65
Girls	Ext10	1	8	13	16	27	65
Girls	Ext11	1	5	21	29	9	65
Total		16	53	148	248	250	715
Boys	Ext1	0	0	3	25	38	66
Boys	Ext2	0	1	9	21	35	66
Boys	Ext3	0	0	6	13	47	66
Boys	Ext4	0	3	6	12	45	66
Boys	Ext5	1	3	18	27	17	66
Boys	Ext6	1	0	2	19	44	66
Boys	Ext7	3	6	34	22	1	66
Boys	Ext8	1	3	14	29	19	66
Boys	Ext9	1	9	18	21	17	66
Boys	Ext10	0	3	10	21	32	66
Boys	Ext11	1	6	20	26	13	66
Total		8	34	140	236	308	726
Column Total		24	87	288	484	558	1441

In contrast to AttDef, the Ext questionnaire appears as more ‘easy’ to receive high scores. About 90% of the answers are placed in category 3-5, while < 2% are placed in category 1.

Table III.Int.0.3

Frequency Table. Internalising behaviour dimension.
8 items scored by 131 subjects.

Gender	Item	score 1	score 2	score 3	score 4	score 5	Row Totals
Girls	Int1	0	1	5	18	41	65
Girls	Int2	0	4	4	28	29	65
Girls	Int3	1	6	15	28	15	65
Girls	Int4	2	0	5	32	26	65
Girls	Int5	0	2	5	25	33	65
Girls	Int6	1	3	8	24	29	65
Girls	Int7	2	7	22	28	6	65
Girls	Int8	0	1	2	29	33	65
Total		6	24	66	212	212	520
Boys	Int1	0	2	14	23	27	66
Boys	Int2	1	5	14	24	22	66
Boys	Int3	1	3	18	30	14	66
Boys	Int4	0	2	15	34	15	66
Boys	Int5	0	3	8	21	34	66
Boys	Int6	0	6	20	26	14	66
Boys	Int7	1	5	27	26	7	66
Boys	Int8	0	4	11	35	16	66
Total		3	30	127	219	149	528
Column Total		9	54	193	431	361	1048

Regarding the Internalising behaviour, the scores are concentrated to the upper half of the category scale. About 95 % of the answers are placed in score 3-5.

Thus, AttDef appears as a fairly ‘centred’ questionnaire, while Ext and Int are perceived as fairly ‘easy’ questionnaires compared to the position of the sample of 131 adolescents.

Before the evaluation of the specific parts of the questionnaire, a brief investigation of the dimensionality, in terms of a Mokken scale analysis, is recommended. This is a hazardous procedure for a small sample, but as we have 131 subjects and no 'non-responses' it might well yield valuable information. Even if the intention is to identify three domains (dimensions) by dimension specific items, their possible overlap should be addressed. An item, thought to be a representative of a specific domain might well turn out to be equally or better suitable for another domain (i.e. the respondents tend to identify the item as belonging to a domain different from which the authors had in view).

The dimensionality of the total questionnaire: The combined data file =
Attention Deficit (AttDEf) + Externalising (Ext) + Internalising behaviour (Int).
The process of identifying dimensions starts with two items having the best pairwise scalability, Further items are then included until a lower limit is reached. Using the limit $H_i \geq 0.3$, the first dimension is reached where AttDef 1, 3 – 13 are identified, but AttDef2 is left outside. The process restarts and dimensions two and three are specified. On the whole, Externalising and Internalising are identified but a few items are left to 'residual' dimensions. AttDef2 did not qualify into any of these dimensions (Table III.0.0.4.)

Table III.0.0.4. Lower bound $H_{ij} = 0.3$

Dimensionality analysis of the total questionnaire									
Item	**scale**	Item	**scale**	Item	**scale**	Item	scale	Item	scale
AttDef1	**1**	Int1	**2**	Ext1	**3**	Ext5	4	Int3	5
AttDef3	**1**	Int2	**2**	Ext2	**3**	Ext9	4	Int7	5
AttDef4	**1**	Int4	**2**	Ext3	**3**				
AttDef5	**1**	Int5	**2**	Ext4	**3**				
AttDef6	**1**	Int6	**2**	Ext6	**3**				
AttDef7	**1**	Int8	**2**	Ext7	**3**				
AttDef8	**1**			Ext8	**3**				
AttDef9	**1**			Ext10	**3**				
AttDef10	**1**			Ext11	**3**				
AttDef11	**1**								
AttDef12	**1**								
AttDef13	**1**								

* AttDef 2 was not placed in any scale due to $H_i < 0.3$
(Table III.0.0.4. is slightly different to the table presented in the article due to identity problems for two persons in the data file. However, this does not affect the analyses of the specific dimensions.)

Lowering the cut off H_i to ≥0.2 yields a result which indicates that the dimensions are reasonably defined.AttDef2 is now included in scale 1.

Table III.0.0.5. Lower bound $H_i = 0.2$

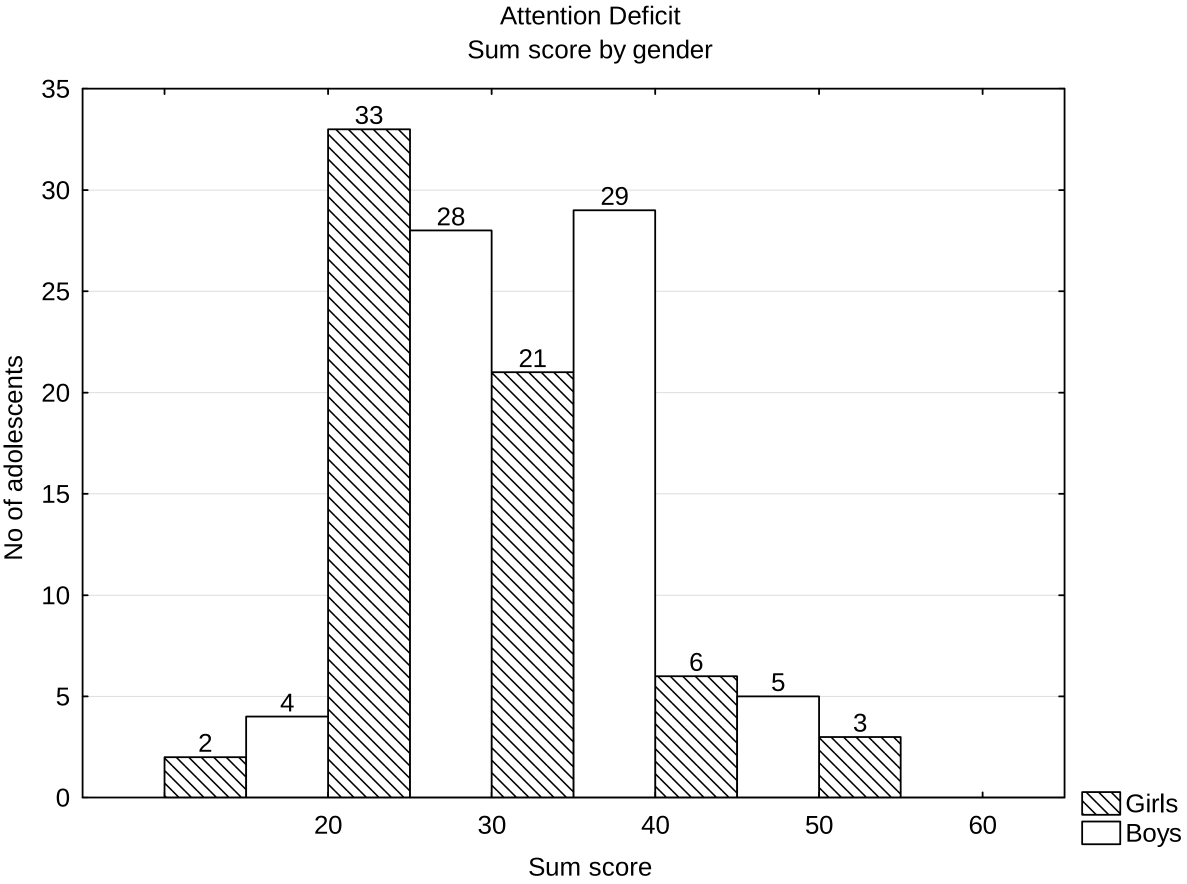
Dimensionality analysis of the total questionnaire					
Item	scale	Item	scale	Item	scale
AttDef1	1	Int1	2	Ext1	3
AttDef2	1	Int2	2	Ext2	3
AttDef3	1	Int3	2	Ext3	3
AttDef4	1	Int4	2	Ext4	3
AttDef5	1	Int5	2	EXt5	3
AttDef6	1	Int6	2	Ext6	3
AttDef7	1	Int7	2	Ext9	3
AttDef8	1	Int8	2	Ext10	3
AttDef9	1				
AttDef10	1	Ext7	2		
AttDef11	1	Ext8	2		
AttDef12	1	Ext11	2		
AttDef13	1				

From table III.0.0.4. and table III.0.0.5. we might expect problems concerning AttDef2 and a few items where Int and Ext seem to overlap, i.e. share the same information. However, the dimensionality analysis does not indicate any immediate concern about the uniqueness of the three domains. It can be argued that the strong pairwise scalability between AttDef3 and AttDEf13 (revealed in step 1), or a certain residual correlation found in step 2, will rule the process, but excluding AttDef13 did not change the structure of dimensions.

Investigation of the ‘Attention Deficit’ questionnaire (AttDef)

Step 1 AttDef. The Mokken scale analysis

Fig. III.AttDef.1.1. Gender specific distribution of AttDef sum scores



Even if the distribution of the sum scores are similar for boys and girls, it is not sufficient for saying that the questionnaire is equally perceived by the two gender groups. Gender specific scalabilities might be hidden and should be taken into account in the analyses. Table III.AttDef.1.1. presents item pair scalabilities and gender specific item scalabilities.

Table III.AttDef.1.1. Attention Deficit: item and item pairwise scalabilities.

	AttDef1	AttDef2	AttDef3	AttDef4	AttDef5	AttDef6	AttDef7	AttDef8	AttDef9	AttDef10	AttDef11	AttDef12	AttDef13
AttDef1	1.000	0.192	0.663	0.239	0.382	0.304	0.484	0.467	0.360	0.301	0.534	0.430	0.646
AttDef2	**0.192**	1.000	**0.177**	**0.244**	0.366	0.449	**0.153**	**0.185**	0.359	**0.171**	**0.209**	**0.260**	**0.029**
AttDef3	0.663	0.177	1.000	0.465	0.510	0.394	0.574	0.552	0.464	0.427	***0.807***	***0.708***	***0.822***
AttDef4	0.239	0.244	0.465	1.000	0.486	0.490	0.502	0.438	0.362	0.504	0.349	0.623	0.411
AttDef5	0.382	0.366	0.510	0.486	1.000	***0.718***	0.312	0.295	0.356	0.354	0.336	0.532	0.427
AttDef6	0.304	0.449	0.394	0.490	0.718	1.000	0.431	0.276	0.547	0.579	0.413	0.592	0.311
AttDef7	0.484	0.153	0.574	0.502	0.312	0.431	1.000	0.489	0.487	0.419	0.554	0.575	0.495
AttDef8	0.467	0.185	0.552	0.438	0.295	0.276	0.489	1.000	0.466	0.400	0.383	0.628	0.541
AttDef9	0.360	0.359	0.464	0.362	0.356	0.547	0.487	0.466	1.000	0.504	0.449	0.557	0.362
AttDef10	0.301	0.171	0.427	0.504	0.354	0.579	0.419	0.400	0.504	1.000	0.293	0.683	0.361
AttDef11	0.534	0.209	0.807	0.349	0.336	0.413	0.554	0.383	0.449	0.293	1.000	0.552	0.616
AttDef12	0.430	0.260	0.708	0.623	0.532	0.592	0.575	0.628	0.557	0.683	0.552	1.000	0.574
AttDef13	0.646	0.029	0.822	0.411	0.427	0.311	0.495	0.541	0.362	0.361	0.616	0.574	1.000

Item scalabilities				AttDef2 excluded	
Item	all	girls	boys	item	all
AttDef1	0.422	0.459	0.366	AttDef1	0.440
AttDef2	**0.228**	**0.255**	**0.208**	**----**	**---**
AttDef3	0.561	0.588	0.534	AttDef3	0.589
AttDef4	0.424	0.463	0.382	AttDef4	0.439
AttDef5	0.422	0.446	0.433	AttDef5	0.426
AttDef6	0.453	0.481	0.447	AttDef6	0.453
AttDef7	0.460	0.503	0.411	AttDef7	0.485
AttDef8	0.434	0.403	0.480	AttDef8	0.454
AttDef9	0.437	0.430	0.456	AttDef9	0.444
AttDef10	0.419	0.446	0.387	AttDef10	0.438
AttDef11	0.461	0.497	0.415	AttDef11	0.482
AttDef12	0.565	0.570	0.562	AttDef12	0.589
AttDef13	0.473	0.527	0.413	AttDef13	0.508
Item set	0.445	0.470	0.422	Item set	0.479

The matrix of pairwise item scalabilities, shown in table III.AttDef.1.1., should be screened for negative values. These might have a large negative impact on the item set and the items' ability to co-operate. A few large scalabilities (> 0.7) are also observed. The signification of these will hopefully become apparent in Step 2.
All item scalabilities are positive, just one item scalability is <0.3. This implies the sum score to be a reasonable measure for ranking the adolescents on the Attention Deficit scale. Only a slight, but general, increase in item scalabilities was noticed when AttDef2 was excluded.
No systematic difference between girls and boys could be statistically demonstrated, 90% bootstrap C.I. for the difference was [-0.086 0.216].

Fig. III.AttDef.1.2. Gender specific scalabilities

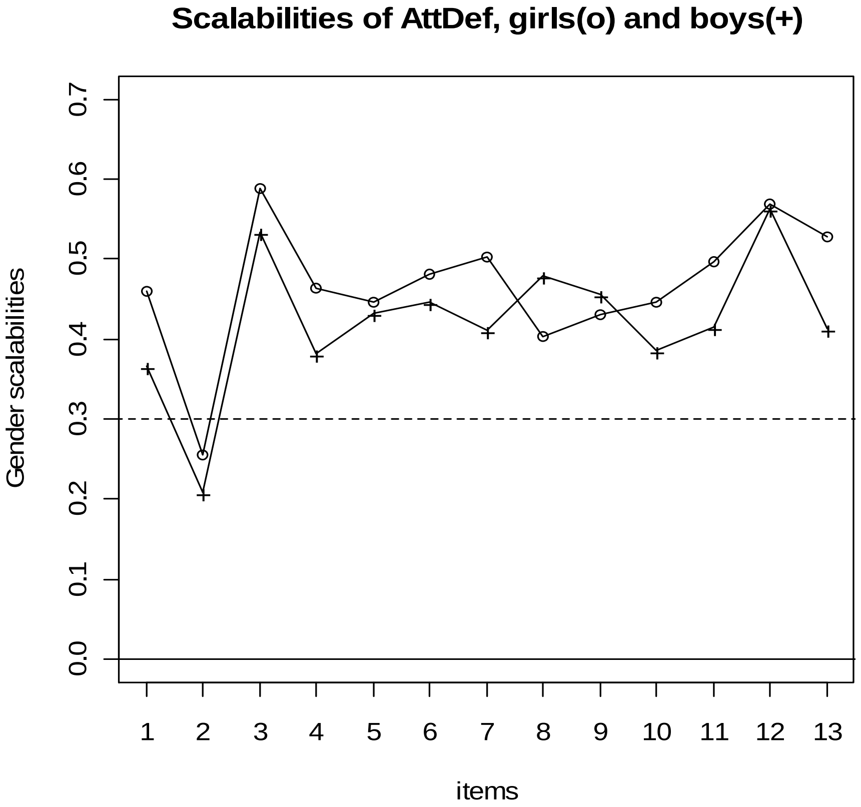
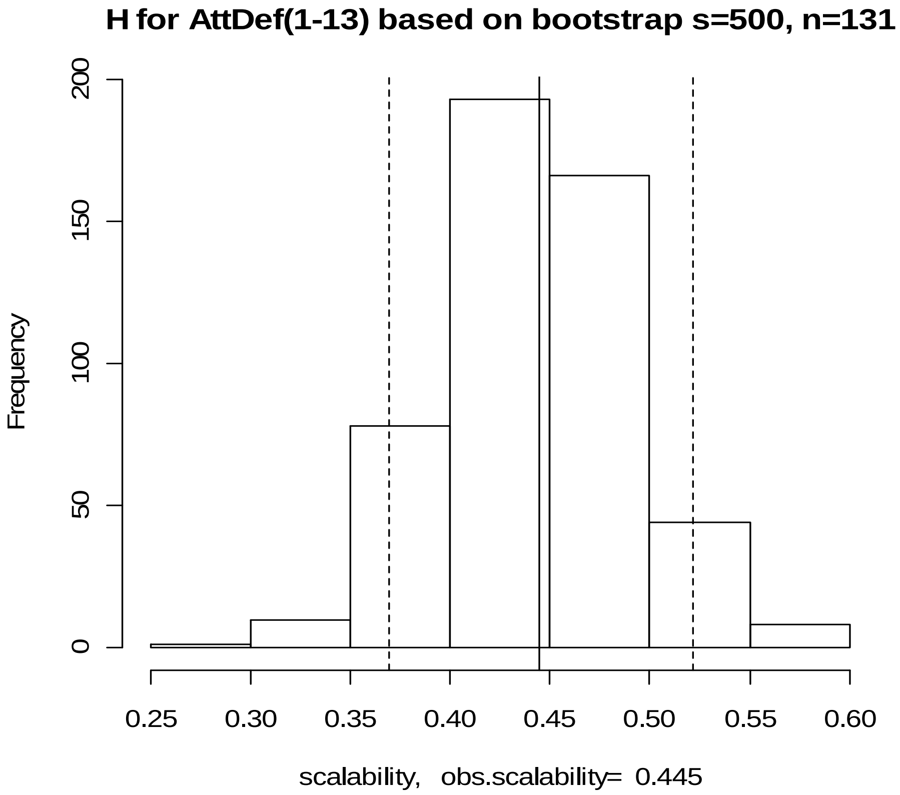


Fig. III.AttDef.1.2. shows a similar item structure for boys and girls. AttDef2 is identified as a weak term and its role has to be further investigated. However, we start with an analysis of the variability of the item set scalability.

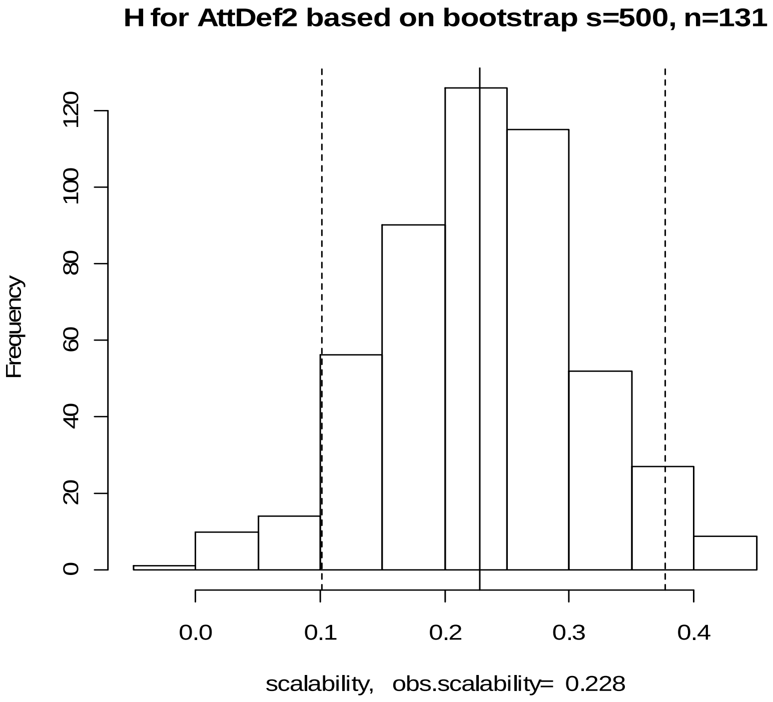
Fig. III.AttDef.1.3. Analysis of the item set variability



```
Min.    1st Qu.  Median  Mean    3rd Qu.   Max.     5%      95%
0.265   0.410   0.444   0.443   0.475   0.575   0.370   0.521
```

The C.I. for the item set scalability is well above 0.3 and can be classified as good (H> 0.4). AttDf2 can be questioned with its low scalability, so an analysis of its variability is appropriate.

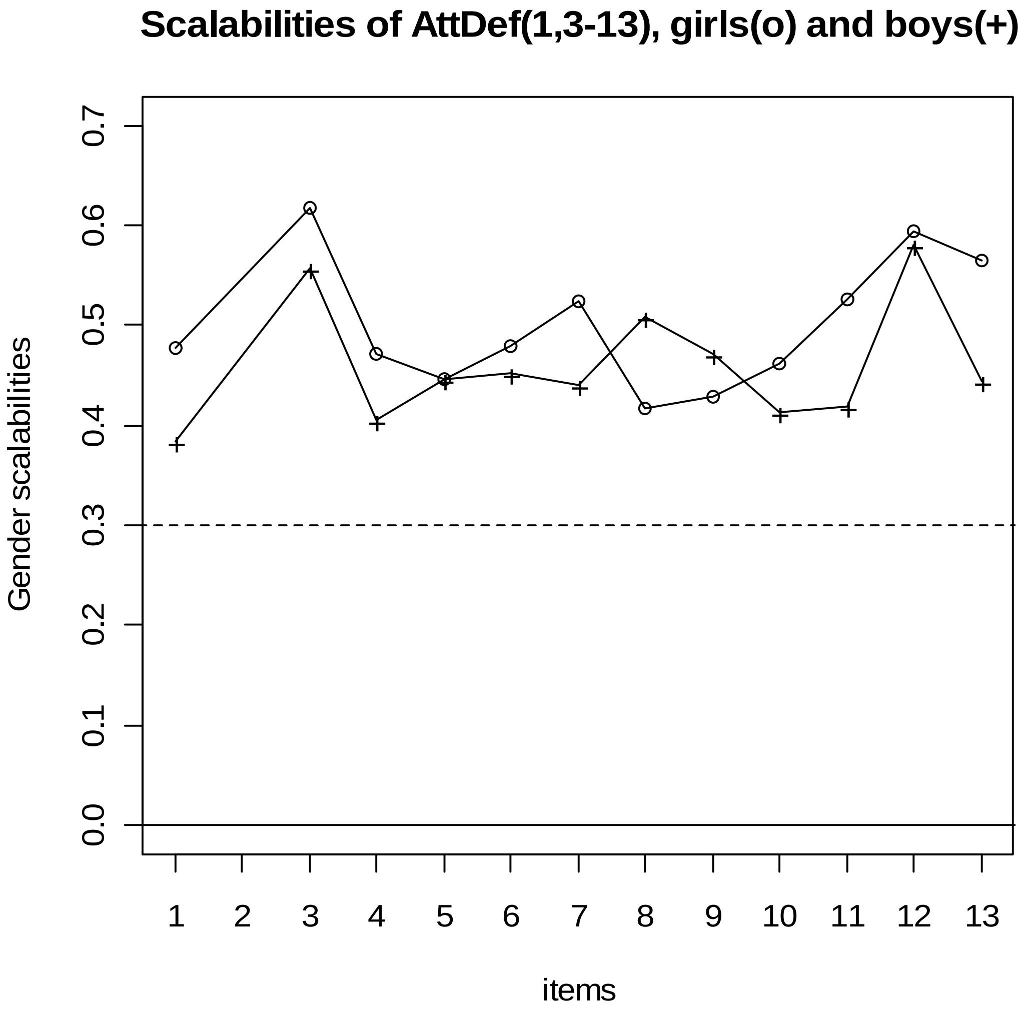
Fig. III.AttDef.1.4. Analysis of the variability of AttDef2



```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.      5%     95%
 -0.034   0.176   0.235   0.229   0.279   0.444   0.101   0.376
```

Even if AttDef2 is a weak item, its C.I. is well on the positive side. Only an negligible increase of the item scalabilities is achieved by exclusion of AttDef2 (Table III.AttDef.1.1.1.). Compare fig.III.AttDef.1.2. with fig. III.AttDef.1.5. There is no immediate concern about AttDef2 as a ‘disturbing’ item, but it might well be redundant.

Fig. III.AttDef.1.1.5. Gender specific item scalabilities with AttDef2 excluded



Exclusion of AttDef2 yields a more attractive structure of the item scalabilities, with a slight increase of the other 12 scalabilities.

Monotonicity and non-intersection

The monotonicity check caused no problems. The check of intersection, however, revealed an unstable structure for the whole AttDef item set. Varying the minimum rest score group size, with n=20, 25 30, 35 sifted out AttDef2 and AttDef12 as potentially problematic items. As revealed by Table III.AttDef.1.2., AttDef2 and AttDef12 have difficulties to find their role in the item hierarchy.

Table III.AttDef.1.2. Items with significant violations of non-intersection

Maxvi= max violation, zmax= max z score,
#zsig= No of significant violations.

Minsize=20

	ItemH	maxvi	zmax	#zsig
AttDef2	0.23	0.35	2.93	4
AttDef9	0.44	0.25	1.81	1
AttDef10	0.42	0.25	2.3	1
AttDef11	0.46	0.3	2.3	1
AttDef12	0.56	0.35	2.93	2

Minsize=25

	ItemH	maxvi	zmax	#zsig
AttDef2	0.23	0.2	1.77	1
AttDef10	0.42	0.22	2.3	1
AttDef12	0.56	0.22	2.3	3

Minsize=30

	ItemH	maxvi	zmax	#zsig
AttDef2	0.23	0.3	2.3	1
AttDef11	0.46	0.3	2.3	1

Minsize=35 (only 3 rest score groups)

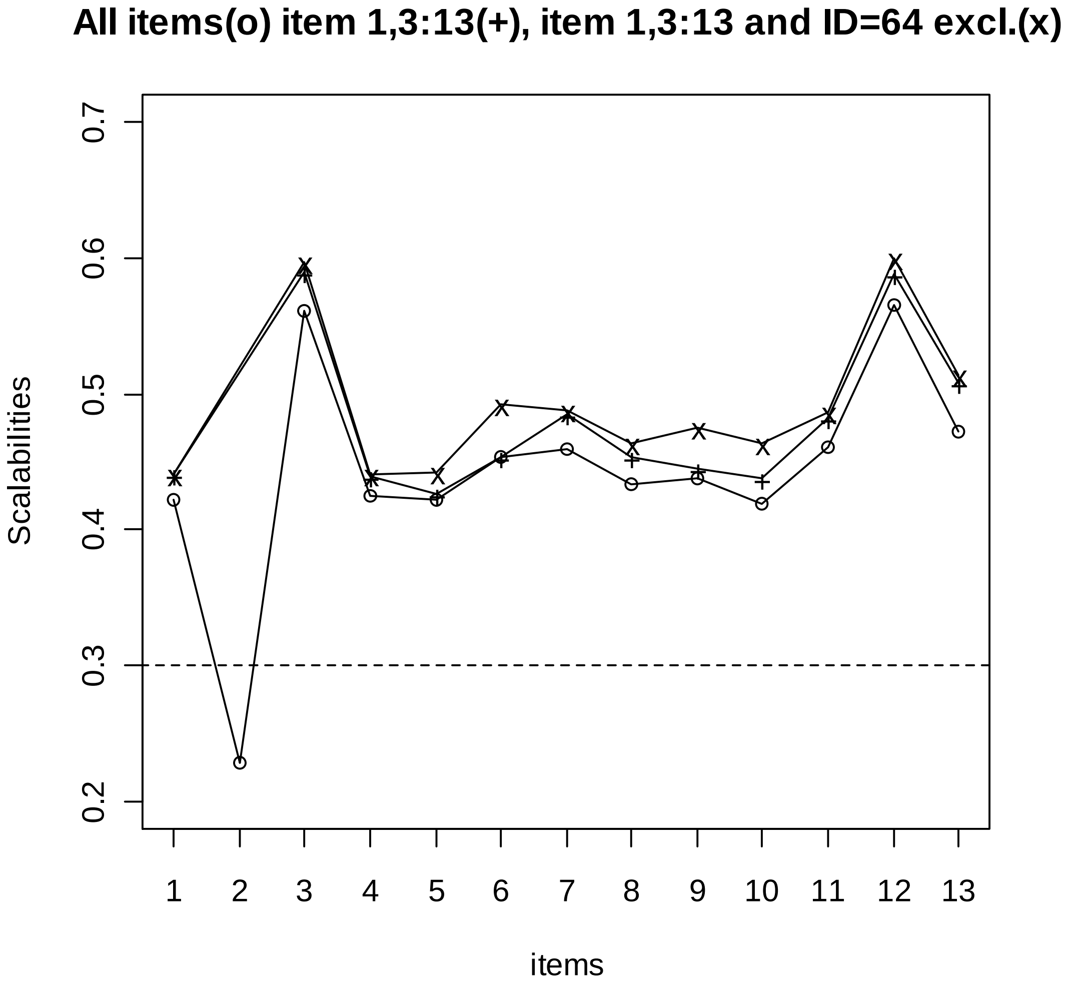
	ItemH	maxvi	zmax	#zsig
AttDef2	0.23	0.24	2.28	1
AttDef9	0.44	0.24	2.28	1

Influential persons

Potentially influential persons were looked for by screening the jackknifed item set scalabilities: Excluding ID= 64 caused a somewhat better scalability. This person may cause problems in the following steps.

Another two persons were identified as potentially influential, however at the lower end of the jackknifed H distribution. They are not likely to do any harm.

Fig. III.AttDef.1.1.6. Illustration of the effect of exclusion of AttDef2 and one influential person



The vertical axis in Fig. III.AttDef.1.1.6. is shortened to better visualise the difference between the scalabilities. The impact by excluding a weak item is as illustrated earlier. The figure also shows that the exclusion of just one subject may yield about the same impact on the scalabilities as does exclusion of an item, in spite of a fairly large (under the present circumstances) sample size.
The corresponding change of the item set scalability, H, is 0.445 → 0.479 → 0.492.

Table III.AttDef.1.1. , fig. III.AttDef.1.2.and the subsequent analyses indicate that there is a reasonable basis for proceeding to a parametric model (Step 2), bearing in mind that AttDef2 and ID=64 may cause problems.

<u>Step 2 AttDef.</u> Analysis by a Rasch model

In a first run with the Rasch RSM, the result revealed miscellaneous signs to be taken into account, see Table III.AttDef.2.1. A set of items shows |ZSTD|> 2.0, a conventional sign of a not well behaved model. However, the MNSQ:s are within an acceptable range [Linacre J. Winsteps 3.66, 2008]. Some negative ZSTD:s indicate dependence between items and AttDef2 is again, as was seen in Step 1, identified a weak, not very contributing item (est.discr. = 0.63). The same can be said about AttDef13. Furthermore, one person, ID= 64, was identified as an outlier with an incoherent answer profile from a Rasch model perspective (person infit MNSQ= 5.5). This person was identified already in Step 1. The other two potentially problematic persons, identified in Step 1, did not, however, cause any problems in this step.

Table III.AttDef.2.1.   The RSM and an analysis of the residuals.ID= 64 deleted

```
-------------------------------------------------------
|          MODEL|   INFIT  |EXACT MATCH|ESTIM|         |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM    |
|---------------+----------+-----------+-----+---------|
|    .02     .14|1.05    .4| 61.2  56.2|  .96| AttDef1 |
|   -.09     .14|1.35   2.6| 51.9  56.2|  .63| AttDef2 |
|    .72     .14| .69  -2.7| 67.4  58.2| 1.37| AttDef3 |
|    .27     .14|1.25   1.9| 46.5  57.0|  .71| AttDef4 |
|   -.35     .13| .76  -2.0| 64.3  55.8| 1.24| AttDef5 |
|   -.97     .13| .88  -1.0| 56.6  54.2| 1.10| AttDef6 |
|   -.42     .13| .77  -2.0| 63.6  55.7| 1.23| AttDef7 |
|    .23     .14| .92   -.6| 62.0  56.7| 1.09| AttDef8 |
|   -.22     .14| .99   -.1| 63.6  55.9| 1.04| AttDef9 |
|    .27     .14|1.12   1.0| 53.5  57.0|  .87| AttDef10|
|   -.17     .14|1.05    .4| 59.7  55.9|  .97| AttDef11|
|    .72     .14| .81  -1.6| 59.7  58.2| 1.23| AttDef12|
|   -.02     .14|1.38   2.8| 45.7  56.3|  .66| AttDef13|
|---------------+----------+-----------+-----+---------|
|Mean .00       | Person reliability = 0.86
| Std .44       |   Item reliability = 0.90
-----------------
```

Analysis of residuals from the RSM

```
     Standardized residual variance (in Eigenvalue units)
                                                  Empirical
Total raw variance in observations     =         23.9 100.0%
  Raw variance explained by measures   =         10.9  45.7%
    Raw variance explained by persons  =          9.8  41.0%
    Raw Variance explained by items    =          1.1   4.7%
  Raw unexplained variance (total)     =         13.0  54.3%
  Unexplained variance in 1st contrast =          2.5  10.4%
```

It is clear from fig 2 (in the manuscript) that the item locations are estimated within a too narrow range (insufficient coverage), which explains the very small amount of the variance explained by items, see Table III.AttDef.2.1.

Table III.AttDef.2.2. Largest standardized residual correlations used to identify dependent items

```
----------------------------
|CORREL-|         |         |
|  ATION| ITEM    |ITEM     |
|-------+---------+---------|
|   .43 | AttDef3 |AttDef13 |
|   .31 | AttDef5 |AttDef6  |
|-------+---------+---------|
|  -.42 | AttDef2 |AttDef13 |
|  -.38 | AttDef1 |AttDef4  |
|  -.35 | AttDef3 |AttDef6  |
|  -.34 | AttDef6 |AttDef13 |
----------------------------
```

Substantial correlations were found between some item residuals. The Rasch RSM did not sufficiently fit the data. The correlation between residuals, some significant ZSTD:s and the small amount of explained variance by items indicate that the questionnaire should be further investigated.
Releasing the constraints by applying a Rasch PCM (i.e. item specific category thresholds) and excluding AttDef2 changed the result only to some extent, see table III.AttDef.2.3.

Table III.AttDef.2.3. A model with item specific category thresholds and an analysis of the residuals. Item AttDef2 and ID= 64 deleted.

```
-----------------------------------------------------------------
|          MODEL|   INFIT  |EXACT MATCH|ESTIM|         |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM    |
|---------------+----------+-----------+-----+---------|
|    .26     .14|1.14   1.1| 63.6  60.3|  .87| AttDef1 |
|               |          |           |     | AttDef2 |
|   -.16     .15| .72  -2.3| 69.8  62.9| 1.28| AttDef3 |
|    .70     .13|1.15   1.3| 49.6  53.3|  .79| AttDef4 |
|  -1.25     .17|1.06    .5| 62.8  64.3|  .93| AttDef5 |
|   -.40     .14|1.04    .4| 51.9  57.6|  .96| AttDef6 |
|   -.10     .16|1.00    .1| 62.8  63.0| 1.00| AttDef7 |
|   -.51     .15|1.02    .2| 60.5  59.3|  .99| AttDef8 |
|    .04     .14|1.05    .5| 65.1  56.8|  .96| AttDef9 |
|    .51     .14|1.06    .6| 55.8  55.7|  .92| AttDef10|
|   -.02     .14|1.03    .3| 59.7  56.5|  .96| AttDef11|
|    .98     .14| .67  -3.1| 66.7  56.0| 1.41| AttDef12|
|   -.03     .13|1.01    .1| 58.9  54.3| 1.02| AttDef13|
|---------------+----------+-----------+-----+---------|
|Mean  .00      | Person reliability = 0.87
|Std   .56      |   Item reliability = 0.93
-----------------
```

Analysis of residuals from the PCM

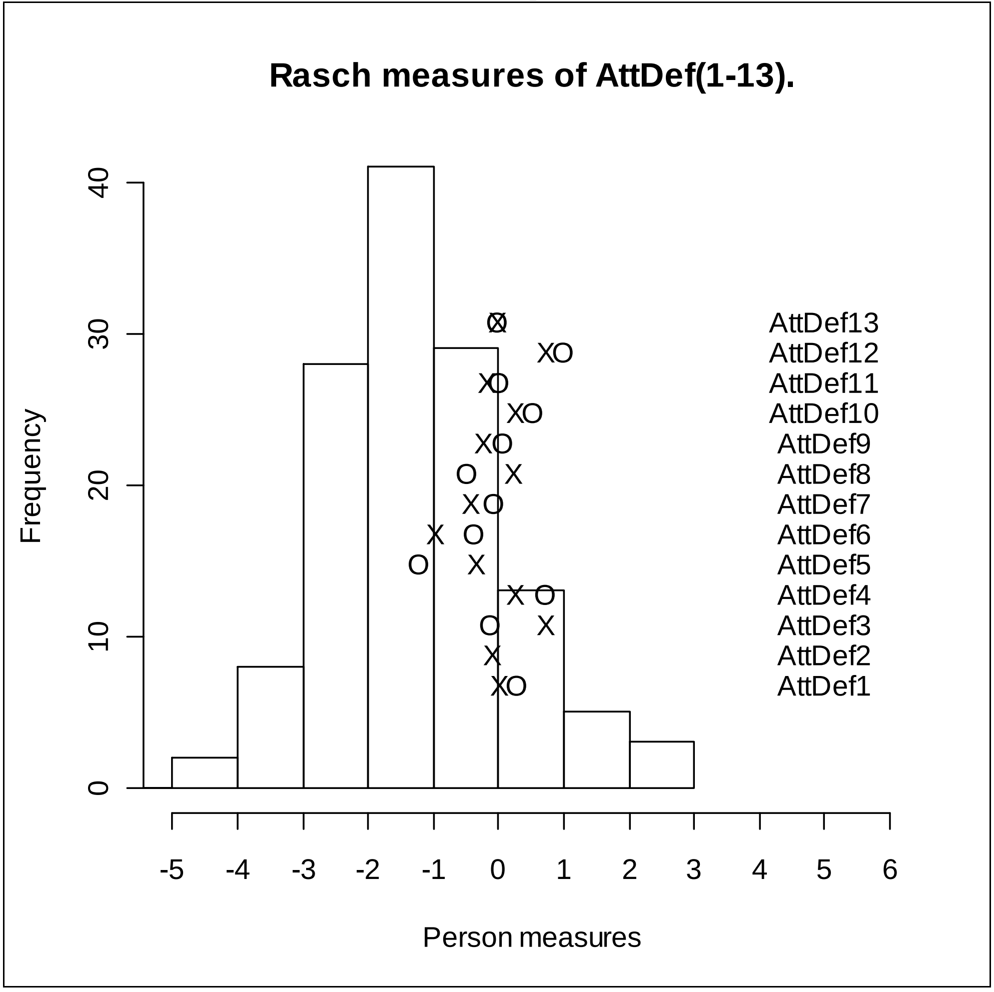
```
Standardized residual variance (in Eigenvalue units)
                                                      Empirical
Total raw variance in observations     =          24.5 100.0%
  Raw variance explained by measures   =          12.5  50.9%
    Raw variance explained by persons  =           8.4  34.2%
    Raw Variance explained by items    =           4.1  16.7%
  Raw unexplained variance (total)     =          12.0  49.1%
    Unexplained variance in 1st contrast =         2.3   9.5%
```

Table III.AttDef.2.4.

```
Largest standardized residual correlations
used to identify dependent item (AttDef2 excluded)
---------------------------------------
|CORREL-|              |              |
|  ATION|     ITEM     |      ITEM    |
|-------+--------------+--------------|
|   .33 |    AttDef3   |    AttDef13  |
|   .33 |    AttDef5   |    AttDef6   |
|-------+--------------+--------------|
|  -.36 |    AttDef1   |    AttDef4   |
|  -.35 |    AttDef6   |    AttDef13  |
|  -.34 |    AttDef3   |    AttDef6   |
---------------------------------------
```

Some improvement is achieved concerning variance explained by items and somewhat lower correlations between item residuals. There is still a problem with AttDef12, MNSQ= 0.67, Infit ZSTD=-3.2, and estim.discr.=1.41. OBS% > EXP% indicates a redundant item. This 'data driven' fit is what can be achieved within the Rasch approach but should not be considered a solution.
There were negligible violations against the ordering of category 5.
As can be seen from fig III.AttDef.2.1., there are only minor changes in item locations by the introduction of a more complicated model, even if the order of item locations is changed to some extent. Even the person estimates are very similar for the two models (not shown)

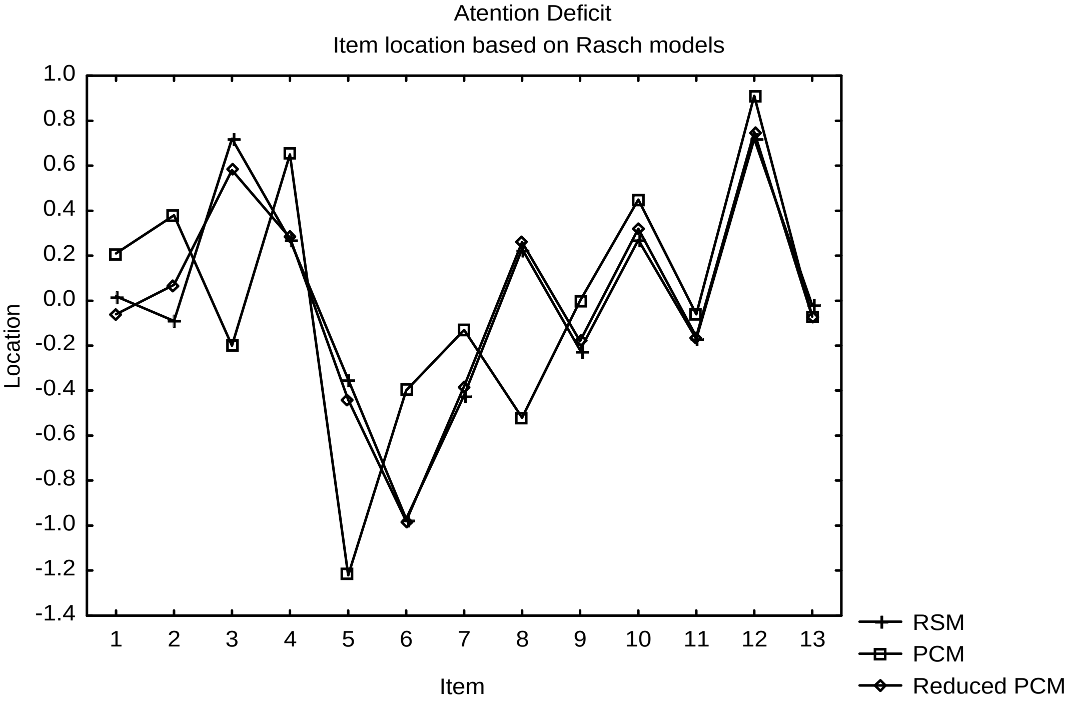
Fig. III.AttDef.2.1. Item locations from table AttDef.2.1 and AttDef.2.3
X= Common scale ID=64 deleted; **O**= Sep. scales AttDef2 deleted, ID=64 deleted



Analysis of a reduced scale

As the category 5 is sparsely endorsed, a reduction of the scale might be considered. This should generally be avoided, as it changes the idea of the intended scale. The problem is certainly an effect of the limited sample size, but in a small sample study it might be necessary for analytical and mathematical reasons. However, an analysis on a reduced scale (1,2,3,4,5)→ (1,2,3,4,4) did not change the result even if it changed the estimates to some extent, see fig. III.AttDef.2.2. As the Rasch model did not seem to be sufficient, an extended approach appeared worthwhile to investigate.

Fig. III.AttDef.2.2.



Step 3 AttDef

The model is extended with item specific discriminations (slopes) and changed to Graded Response Model, GRM.
In a first run, a common discrimination parameter is maintained, although not fixed to be equal to one, as well as a common set of category thresholds. The scale was forced to be reduced (1,2,3,4,5)→ (1,2,3,4,4)

Table III.AttDef.3.1. Parameter estimates (common slope) and item fit statistics

```
+---------+---------+---------+---------+---------+
| ITEM    |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+=========+=========+=========+=========+=========+
| AttDef1 |   0.979 |   0.020 |   0.266 |   0.157 |
| AttDef2 |   0.979 |   0.020 |   0.185 |   0.170 |
| AttDef3 |   0.979 |   0.020 |   0.774 |   0.177 |
| AttDef4 |   0.979 |   0.020 |   0.436 |   0.141 |
| AttDef5 |   0.979 |   0.020 |   0.021 |   0.181 |
| AttDef6 |   0.979 |   0.020 |  -0.563 |   0.148 |
| AttDef7 |   0.979 |   0.020 |  -0.038 |   0.170 |
| AttDef8 |   0.979 |   0.020 |   0.442 |   0.161 |
| AttDef9 |   0.979 |   0.020 |   0.086 |   0.145 |
|AttDef10 |   0.979 |   0.020 |   0.505 |   0.157 |
|AttDef11 |   0.979 |   0.020 |   0.200 |   0.162 |
|AttDef12 |   0.979 |   0.020 |   0.759 |   0.146 |
|AttDef13 |   0.979 |   0.020 |   0.345 |   0.151 |
+------+---------+---------+---------+---------+--+
```

```
            ITEM FIT STATISTICS
 ----------------------------------------
 | ITEM    | CHI-SQUARE |  D.F. | PROB. |
 ----------------------------------------
 | AttDef1 |   14.22376 |    9. | 0.114 |
 | AttDef2 |   19.82799 |    9. | 0.019 <-
 | AttDef3 |   15.03409 |    8. | 0.058 |
 | AttDef4 |   12.81565 |    9. | 0.170 |
 | AttDef5 |    7.35908 |    9. | 0.601 |
 | AttDef6 |    2.91944 |    8. | 0.939 |
 | AttDef7 |   12.02372 |    9. | 0.211 |
 | AttDef8 |    4.77234 |    9. | 0.854 |
 | AttDef9 |   12.39719 |    9. | 0.191 |
 |AttDef10 |   12.97880 |    9. | 0.163 |
 |AttDef11 |    6.74332 |    9. | 0.665 |
 |AttDef12 |   28.98368 |    8. | 0.000 <-
 |AttDef13 |   14.06791 |    9. | 0.119 |
 ----------------------------------------
 | Total   |  164.14696 |  114. | 0.001 |
 ----------------------------------------
```

The result is similar to the Rasch RSM and the fit statistics point out AttDef2 and AttDef12 as problematic items (large chi-square, p-values close to zero). Introducing item specific slopes yields the result as presented in table III.AttDef.3.2.

Table III.AttDef.3.2. Parameter estimates (item specific slopes) and item fit statistics

```
+----------+---------+---------+---------+---------+
| ITEM     |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+==========+=========+=========+=========+=========+
| AttDef1  |   0.996 |   0.098 |   0.408 |   0.166 |
| AttDef2  |   0.824 |   0.081 |   0.379 |   0.196 |
| AttDef3  |   1.245 |   0.157 |   0.949 |   0.154 |
| AttDef4  |   0.892 |   0.089 |   0.619 |   0.190 |
| AttDef5  |   1.149 |   0.141 |   0.159 |   0.165 |
| AttDef6  |   0.921 |   0.121 |  -0.475 |   0.150 |
| AttDef7  |   1.278 |   0.156 |   0.143 |   0.157 |
| AttDef8  |   1.144 |   0.106 |   0.602 |   0.177 |
| AttDef9  |   1.012 |   0.117 |   0.119 |   0.141 |
| AttDef10 |   0.831 |   0.114 |   0.727 |   0.166 |
| AttDef11 |   1.018 |   0.099 |   0.373 |   0.167 |
| AttDef12 |   0.957 |   0.167 |   0.847 |   0.128 |
| AttDef13 |   0.798 |   0.089 |   0.273 |   0.188 |
+------+---------+---------+---------+---------+---+
```

```
              ITEM FIT STATISTICS
 ----------------------------------------
 | ITEM     | CHI-SQUARE |  D.F. | PROB. |
 ----------------------------------------
 | AttDef1  |   11.64363 |   11. | 0.391 |
 | AttDef2  |   26.71606 |   12. | 0.009 <-
 | AttDef3  |   17.84268 |    9. | 0.037 |
 | AttDef4  |    7.05991 |   10. | 0.721 |
 | AttDef5  |   10.32607 |    8. | 0.242 |
 | AttDef6  |    1.78093 |    8. | 0.986 |
 | AttDef7  |   13.62694 |    8. | 0.091 |
 | AttDef8  |    9.92578 |    9. | 0.356 |
 | AttDef9  |    7.15533 |    8. | 0.521 |
 | AttDef10 |   14.01998 |   10. | 0.171 |
 | AttDef11 |   15.58011 |   11. | 0.157 |
 | AttDef12 |   24.86434 |   10. | 0.006 <-
 | AttDef13 |   18.14518 |   12. | 0.111 |
 ----------------------------------------
 |  Total   |  178.68697 |  126. | 0.001 |
 ----------------------------------------
```

AttDef2 and AttDef12 are still bad fitting items. A (weak) conditional independence is reasonably fulfilled with just one correlation > 0.3, corr(AttDef3, AttDef13)= 0.46.
The common set of categories forces AttDef2 and AttDef12 to be more in agreement with the other items' slopes than would be reasonable. There is a possibility to keep the 'common set of thresholds' approach for a majority of the items while letting a few items act on their own with respect to their category thresholds. With AttDef2 and AttDef12 released from the common thresholds' approach we get a result as presented in table III.AttDef.3.3.

Table III.AttDef.3.3. Graded response model with three blocks(sets) of items and with item specific slopes.

```
Block 1   CATEGORY PARAMETER  :     1.801    -0.112    -1.689  AttDef1,3-11,13
                  S.E.        :     0.051     0.039     0.063
Block 2   CATEGORY PARAMETER  :     3.737     0.004    -3.741  AttDef2
          S.E.                :     0.420     0.257     0.516
Block 3   CATEGORY PARAMETER  :     1.231     0.048    -1.280  AttDef12
          S.E.                :     0.080     0.087     0.170
+----------+-----+---------+---------+---------+---------+
| ITEM     |BLOCK|  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+======+========+=========+=========+=========+=========+
| AttDef1  |   1 |   0.935 |   0.096 |   0.325 |   0.168 |
| AttDef3  |   1 |   1.226 |   0.154 |   0.831 |   0.147 |
| AttDef4  |   1 |   0.839 |   0.092 |   0.443 |   0.176 |
| AttDef5  |   1 |   1.135 |   0.138 |   0.001 |   0.170 |
| AttDef6  |   1 |   0.947 |   0.115 |  -0.548 |   0.153 |
| AttDef7  |   1 |   1.182 |   0.155 |  -0.025 |   0.153 |
| AttDef8  |   1 |   1.072 |   0.112 |   0.428 |   0.160 |
| AttDef9  |   1 |   1.042 |   0.112 |   0.070 |   0.147 |
| AttDef10 |   1 |   0.898 |   0.112 |   0.479 |   0.163 |
| AttDef11 |   1 |   0.968 |   0.097 |   0.165 |   0.164 |
| AttDef13 |   1 |   0.784 |   0.091 |   0.235 |   0.176 |
+------+---------+---------+---------+---------+---------+
| AttDef2  |   2 |   0.424 |   0.043 |   0.325 |   0.275 |
+------+---------+---------+---------+---------+---------+
| AttDef12 |   3 |   2.001 |   0.320 |   0.597 |   0.122 |
+------+---------+---------+---------+---------+---------+
```

```
               ITEM FIT STATISTICS
 ---------------------------------------------------
 |  BLOCK   | ITEM     | CHI-SQUARE |  D.F. | PROB. |
 ---------------------------------------------------
 |  BLOCK 1 | AttDef1  |   17.31128 |    9. | 0.044 |
 |          | AttDef3  |   10.50215 |   10. | 0.398 |
 |          | AttDef4  |    7.73324 |    9. | 0.562 |
 |          | AttDef5  |   13.80689 |    7. | 0.054 |
 |          | AttDef6  |    7.35453 |    8. | 0.500 |
 |          | AttDef7  |   11.70203 |    7. | 0.110 |
 |          | AttDef8  |    8.09176 |    9. | 0.526 |
 |          | AttDef9  |    9.31155 |    7. | 0.230 |
 |          | AttDef10 |   14.10319 |   10. | 0.168 |
 |          | AttDef11 |    1.93278 |    7. | 0.962 |
 |          | AttDef13 |    4.81163 |    9. | 0.851 |
 | BLOCK 2  | AttDef2  |    7.21880 |   10. | 0.706 |
 | BLOCK 3  | AttDef12 |    3.55311 |    8. | 0.895 |
 ---------------------------------------------------
 |  TOTAL   |          |  117.43295 |  110. | 0.296 |
 ---------------------------------------------------
```

The result as presented in table III.AttDef.3.3. indicates that a common set of category thresholds does not seem reasonable. AttDef2 and AttDef12 should be treated on their own to achieve a reasonable fit. However, failure to reject the fit does not imply that the model is correct. AttDef2 turns out to be an ’outlying’ item with large distances between the item thresholds and a very low discrimination(slope).

Relative Item information

An unconstrained model with both item specific slopes and item specific thresholds might give further information about the items' informative value. 'unconstrained' means that there are no restrictions regarding the estimates of item discriminations and item specific thresholds. This approach yields more likely estimates of the informative value of each item. The result, based on an unconstrained model, is presented in table III.AttDef.3.4.

Table III.AttDef.3.4. Relative Item information(%) based on an unconstrained model.

Item	relative information
AttDef1	6.5
AttDef2	**2.5**
AttDef3	13.5
AttDef4	5.7
AttDef5	5.7
AttDef6	5.8
AttDef7	7.2
AttDef8	7.8
AttDef9	7.4
AttDef10	6.6
AttDef11	7.5
AttDef12	**15.8**
AttDef13	8.1

Once again, a data driven 'over-parameterized' model. However, the message is there. This approach reveals, as expected from the item slopes, that there is little information from AttDef2, but relatively much information from AttDef12 – a further indication that the Rasch approach, i.e. equal discrimination, is not suitable.

Conclusion about the 'Attention Deficit' questionnaire

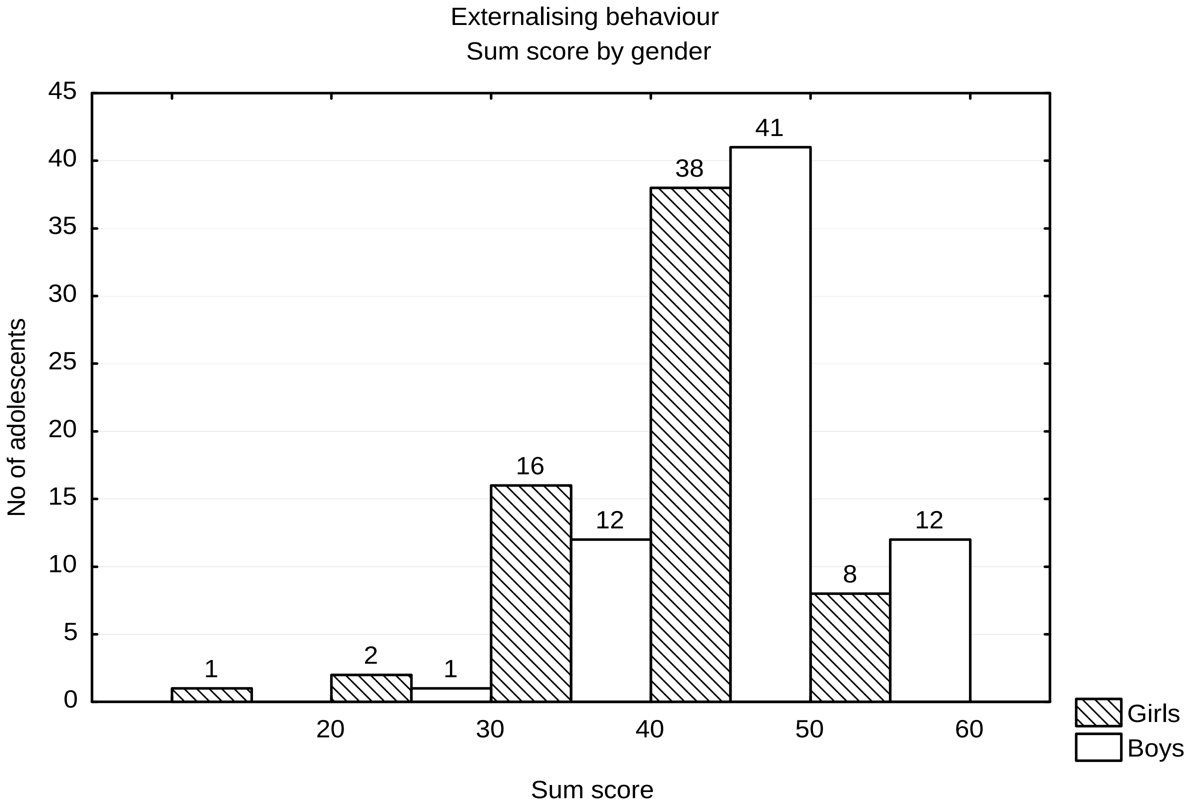
- The straightforward raw sum score may be used to rank the respondents.
- Transformation to an interval scaled variable, based on raw sum score, is difficult to achieve. Item specific weights seem to be unavoidable.
- Even if AttDef2 seems reasonable and with a C.I. well on the positive side, the questionnaire would probably improve by a reformulation or exclusion/replacement of AttDef2.

Investigation of the ‘Externalising Behaviour’ questionnaire (Ext)

Step I Ext. The Mokken scale analysis

The distribution of the sum scores, concerning dimension Ext, is about the same for boys and girls. However, it is clearly revealed (Step 1 in the manuscript) that there is a systematic difference between genders with respect to the item scalabilities.

Fig III.Ext.1.1. The sum score from 131 adolescents.



In a first run, item pair scalabilities are screened for weak and negative values as well as gender specific item scalabilities, see table III.Ext.1.1.

Table III.Ext.1.1. Item pair scalabilities (H_{ij} <0.2 in bold)

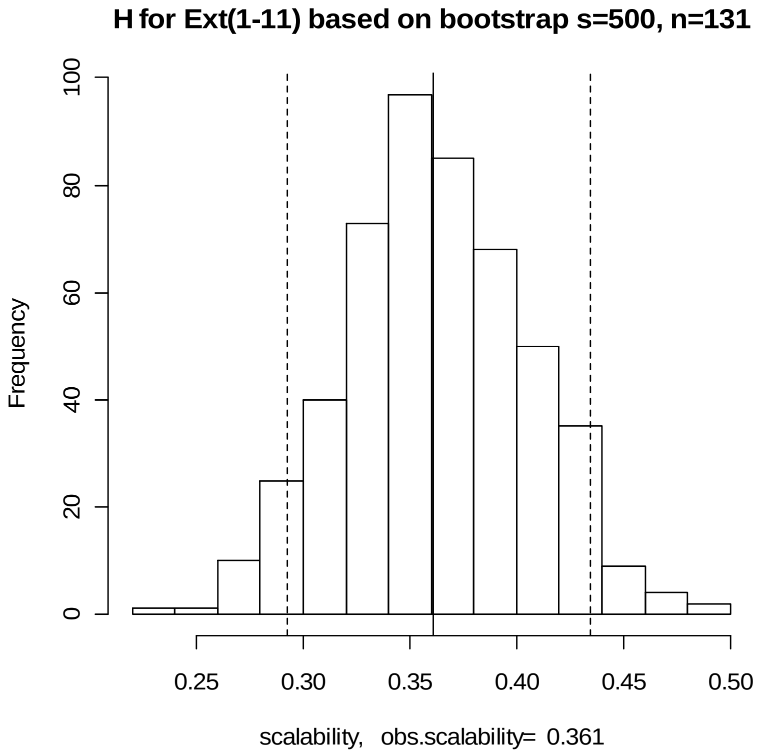
```
       Ext1  Ext2  Ext3  Ext4  EXt5  Ext6  Ext7  Ext8  Ext9 Ext10 Ext11
Ext1  1.000 0.486 0.429 0.352 0.168 0.512 0.424 0.293 0.238 0.443 0.372
Ext2  0.486 1.000 0.589 0.513 0.382 0.423 0.380 0.199 0.362 0.588 0.378
Ext3  0.429 0.589 1.000 0.371 0.273 0.593 0.297 0.282 0.192 0.542 0.380
Ext4  0.352 0.513 0.371 1.000 0.263 0.328 0.404 0.149 0.239 0.537 0.206
EXt5  0.168 0.382 0.273 0.263 1.000 0.330 0.216 0.263 0.358 0.313 0.237
Ext6  0.512 0.423 0.593 0.328 0.330 1.000 0.474 0.382 0.334 0.455 0.410
Ext7  0.424 0.380 0.297 0.404 0.216 0.474 1.000 0.448 0.427 0.458 0.597
Ext8  0.293 0.199 0.282 0.149 0.263 0.382 0.448 1.000 0.273 0.263 0.359
Ext9  0.238 0.362 0.192 0.239 0.358 0.334 0.427 0.273 1.000 0.303 0.284
Ext10 0.443 0.588 0.542 0.537 0.313 0.455 0.458 0.263 0.303 1.000 0.447
Ext11 0.372 0.378 0.380 0.206 0.237 0.410 0.597 0.359 0.284 0.447 1.000
```

```
Item scalabilities
      item   all girls  boys
Ext1     1 0.366 0.431 0.258
Ext2     2 0.426 0.544 0.290
Ext3     3 0.390 0.426 0.334
Ext4     4 0.332 0.387 0.247
EXt5     5 0.285 0.380 0.175
Ext6     6 0.416 0.486 0.367
Ext7     7 0.414 0.476 0.387
Ext8     8 0.287 0.409 0.235
Ext9     9 0.303 0.341 0.242
Ext10   10 0.427 0.541 0.268
Ext11   11 0.359 0.411 0.318
Total      0.361 0.436 0.279
```

A 90% confidence interval for the gender difference, 0.436 -0.279 = 0.157 is [0.029, 0.315] as estimated by bootstrapping.

The generally lower scalabilities (table III.Ext.1.1.) for boys is an indication that the questionnaire is more suitable for girls than for boys, or that boys are more heterogeneous than are the girls. However, all item pair scalabilities are positive, which means that no items seem to be counterproductive.

Fig. III.Ext.1.2. Variability of the tem set scalability



```
   Min. 1st Qu.  Median  Mean   3rd Qu.    Max.    5%      95%
 0.234   0.335   0.361   0.363   0.390   0.497   0.292   0.434
```

The bootstrapped estimation of the 90% C.I. for the total set is done without any account for a gender difference. A stratified estimation, which means a gender specific bootstrapping, yields approximately the same result.

Fig. III.Ext.1.3. Gender specific item scalabilities

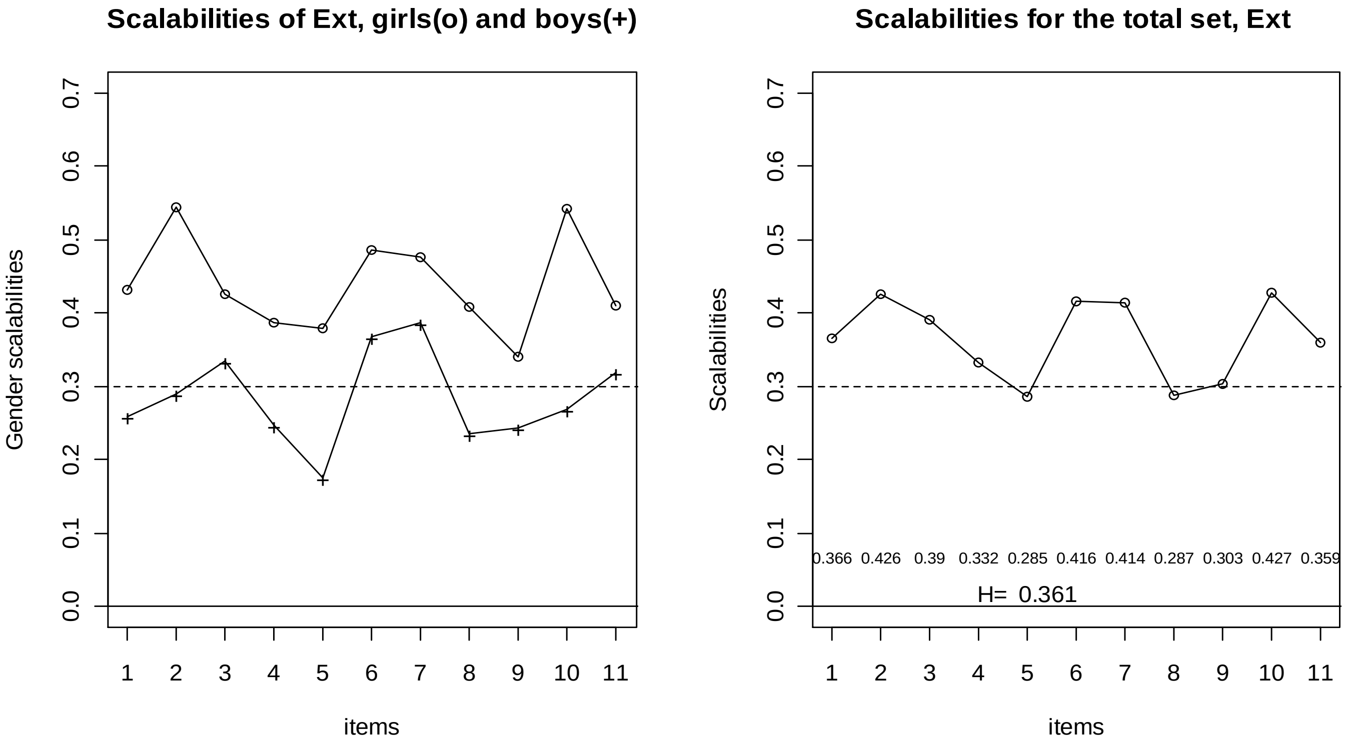


Fig. III.Ext.1.3. reveals a similar structure of the gender specific scalability profiles, with fairly similar differences for all items. This indicates a common perception of the questionnaire by the two genders, although boys seem to constitute a more heterogeneous population.

Fig. III.Ext.1.2. says that we have, over all, a weak scale (H= 0.361), which allows us to at least rank the persons on an externalizing behaviour scale. Fig. III.Ext.1.3. on the other hand, indicates that ranking boys is more hazardous. With a gender difference in mind, a parametric Rasch model (Step 2) should be considered.

Monotonicity and non-intersection

The monotonicity and the non-intersection were reasonable with just a few, non-significant violations.

<u>Step 2 Ext.</u> Analysis by a Rasch model

The Rasch RSM, presented in table III.Ext.2.1., yields a somewhat different message. Exr4 and Ext7 are pointed out as problematic items.

Table III.Ext.2.1. Estimated from the Rasch RSM

```
------------------------------------------------------
|          MODEL|   INFIT  |EXACT MATCH|ESTIM|      |
|  MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|---------------+----------+-----------+-----+------|
|    -.76     .14| .78  -1.7| 62.8  59.4| 1.17| Ext1 |
|    -.34     .13| .72  -2.3| 62.0  53.4| 1.28| Ext2 |
|    -.97     .15|1.04    .3| 63.6  61.9| 1.06| Ext3 |
|    -.76     .14|1.81   4.8| 54.3  59.4|  .61| Ext4 |
|     .35     .12|1.21   1.6| 44.2  47.9|  .77| Ext5 |
|    -.40     .13| .92   -.5| 58.9  55.0| 1.12| Ext6 |
|    1.37     .11| .61  -3.8| 50.4  44.8| 1.40| Ext7 |
|    -.09     .12|1.22   1.7| 48.8  51.4|  .76| Ext8 |
|     .93     .11|1.32   2.5| 42.6  44.9|  .63| Ext9 |
|    -.06     .12|1.01    .1| 55.0  51.3| 1.09| Ext10|
|     .72     .11| .84  -1.3| 54.3  46.3| 1.18| Ext11|
|---------------+----------+-----------+-----+------|
|Mean .00       | Person reliability = 0.78
|S.D. .73       |  Item reliability = 0.97
---------------
```

Person reliability= 0.78 is considered weak while item reliability= 0.97 is sufficient. No strongly incoherent answer profile was seen.

Table III.Ext.2.2. Largest standardized residual correlations
        used to identify dependent items

```
----------------------------------
|CORREL-| ENTRY      | ENTRY     |
|  ATION|NUMBER ITEM |NUMBER ITEM|
|-------+------------+-----------|
|  -.31 |      Ext3 |    Ext7    |
|  -.31 |      Ext2 |    Ext8    |
----------------------------------
```

Max. correlation ≈ 0.3 does not indicate any strong dependency between residuals, given the Rasch model.

Due to the ‘misfit’ of Ext4 and Ext7, and the gender difference seen in Step 1, a DIF analysis between the genders might yield more information.

Table III.Ext.2.3. DIF gender specification

```
-----------------------------------------------------
| PERSON     SUMMARY DIF                    ITEM    |
| CLASSES    CHI-SQUARE   D.F.  PROB.               |
|---------------------------------------------------|
|       2         .8407      1  .3592       Ext1    |
|       2         .9404      1  .3322       Ext2    |
|       2        1.7710      1  .1833       Ext3    |
|       2         .1313      1  .7171       Ext4    |
|       2        2.7463      1  .0975       Ext5    |
|       2       17.4265      1  .0000       Ext6    |
|       2        3.8652      1  .0493       Ext7    |
|       2       15.0779      1  .0001       Ext8    |
|       2         .6061      1  .4363       Ext9    |
|       2        1.2204      1  .2693       Ext10   |
|       2        1.1701      1  .2794       Ext11   |
-----------------------------------------------------
```

There is a clear DIF for Ext6 and Ext8. Separate scales (Rasch PCM) do not change this structure. The structure remains even after exclusion of Ext4, which was identified as a problematic item in table III.Ext.2.1.

There is a possibility to split items to gender specific locations. An analysis with Ext6 and Ext8 split into gender specific items (F/M) are shown in table III.Ext.2.4.

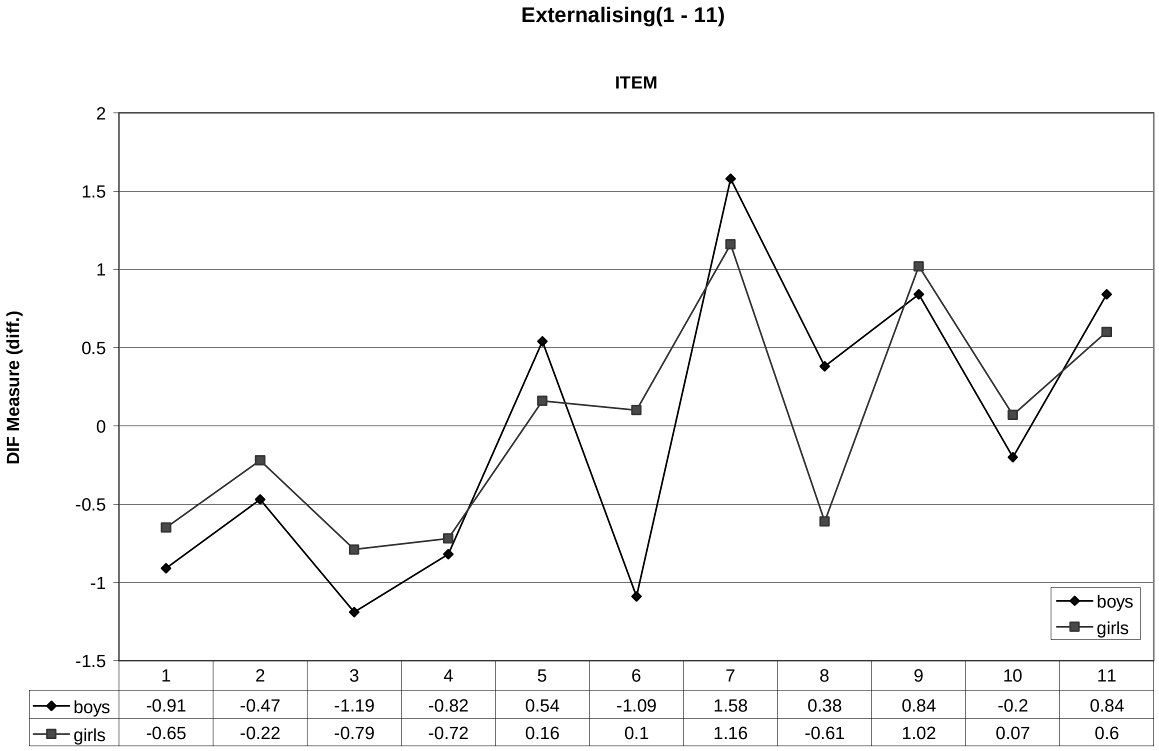
Table III.Ext.2.4. Rasch RSM with item split.

```
-----------------------------------------------------
|          MODEL|   INFIT  |EXACT MATCH|ESTIM|      |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|---------------+----------+-----------+-----+------|
|   -.72     .14| .80  -1.5| 62.8  59.8| 1.16| Ext1 |
|   -.29     .13| .74  -2.2| 62.0  53.9| 1.26| Ext2 |
|   -.93     .15|1.05    .4| 63.6  62.2| 1.05| Ext3 |
|   -.72     .14|1.84   5.0| 54.3  59.8|  .57| Ext4 |
|    .42     .12|1.24   1.8| 44.2  48.5|  .75| Ext5 |
|    .16     .17| .78  -1.3| 56.3  50.4| 1.23| Ext6F|
|   1.46     .11| .62  -3.6| 50.4  45.3| 1.39| Ext7 |
|   -.56     .19|1.16    .9| 48.4  55.8|  .92| Ext8F|
|   1.01     .11|1.35   2.7| 42.6  45.5|  .59| Ext9 |
|    .00     .12|1.03    .3| 55.0  51.8| 1.06| Ext10|
|    .80     .11| .86  -1.1| 54.3  46.9| 1.16| Ext11|
| - 1.06     .22| .96   -.1| 73.8  65.8| 1.16| Ext6M|
|    .45     .17|1.11    .7| 49.2  48.9|  .85| Ext8M|
|--------------+----------+-----------+-----+------|
|mean .00      |
| s.d .77      |
---------------
```

Ext4 remains a bad fitting item. The item measures for Ext6 and Ext8 are reversed regarding girls and boys, saying that they outweigh each other when a raw sum score is used as a basic measure. This is also clear from fig. III.Ext.2.1.

There were no violations against the ordering of categories.

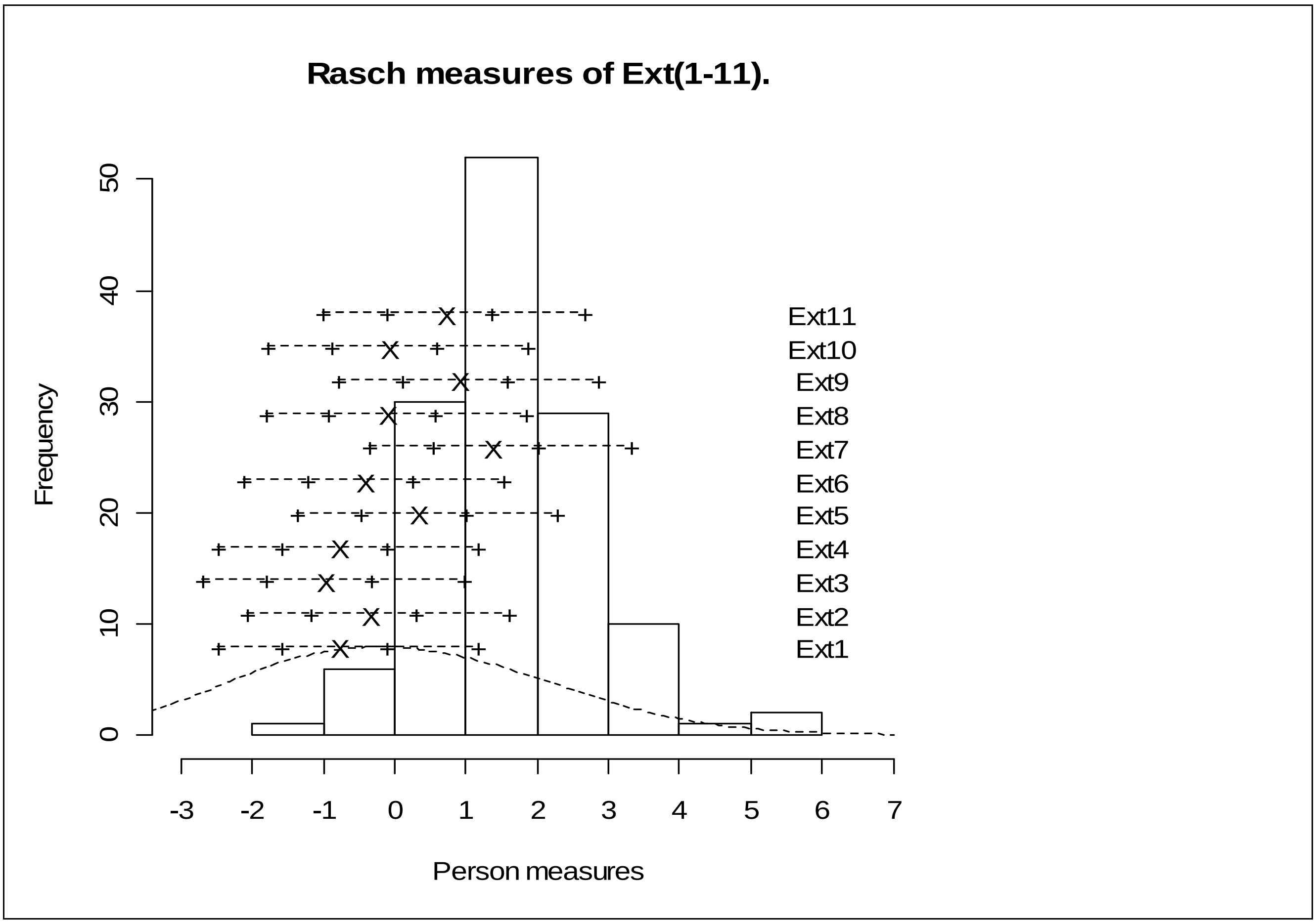
Fig. III.Ext.2.1. DIF measures based on the basic Rasch model.



	1	2	3	4	5	6	7	8	9	10	11
boys	-0.91	-0.47	-1.19	-0.82	0.54	-1.09	1.58	0.38	0.84	-0.2	0.84
girls	-0.65	-0.22	-0.79	-0.72	0.16	0.1	1.16	-0.61	1.02	0.07	0.6

Item specific thresholds (Rasch PCM) do not solve the problem.

Fig. III.Ext.2.2. Item locations and category thresholds for the Rasch RSM



The ‘Externalising questionnaire’ appears as too easy with the coverage concentrated to the lower part of the latent scale (fig. III.Ext.2.2.).

Step 3 Ext

Introducing an extended model with item specific slopes reveals Ext4, Ext6 and Ext8 as weak items (table III.Ext.3.1.). This weakness might well due to the difficulty of placing girls and boys on an equality. This was indicated already in Step 1 as well as in Step 2, where DIF between girls and boys was revealed although the constrained Rasch model was applied. Item specific slopes did not help, there is still a considerable DIF concerning Ext6 and Ext8.

Table III.Ext.3.1.

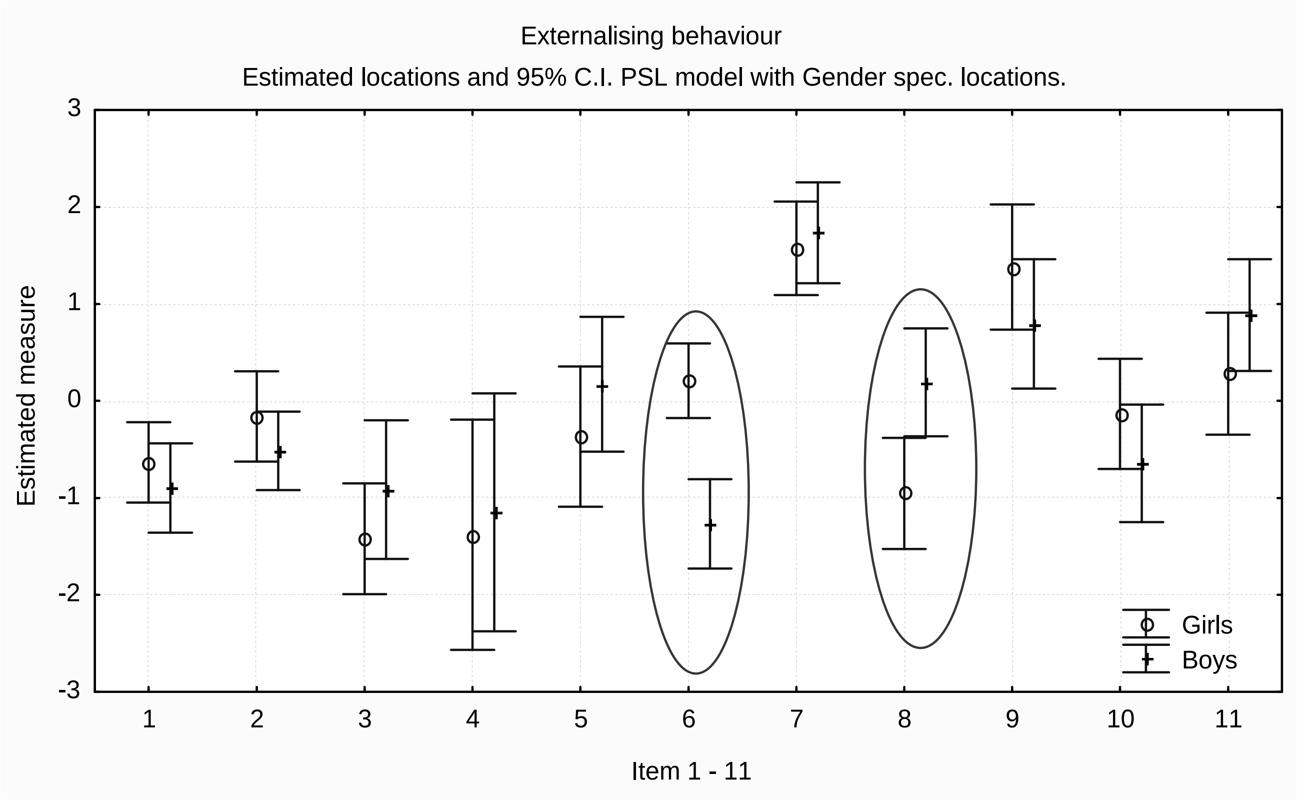
```
+-------+---------+---------+---------+---------+
| ITEM  |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+=======+=========+=========+=========+=========+
| Ext1  |   1.063 |   0.181 |  -0.898 |   0.155 |
| Ext2  |   0.957 |   0.139 |  -0.548 |   0.156 |
| Ext3  |   0.795 |   0.156 |  -1.315 |   0.203 |
| Ext4  |   0.500 |   0.100 |  -1.668 |   0.307 |
| Ext5  |   0.637 |   0.083 |   0.005 |   0.213 |
| Ext6  |   1.019 |   0.172 |  -0.697 |   0.151 |
| Ext7  |   1.298 |   0.267 |   1.331 |   0.154 |
| Ext8  |   0.600 |   0.082 |  -0.473 |   0.214 |
| Ext9  |   0.519 |   0.080 |   0.764 |   0.242 |
| Ext10 |   0.726 |   0.105 |  -0.463 |   0.184 |
| Ext11 |   0.881 |   0.151 |   0.485 |   0.161 |
+------+--------+--------+--------+---------+
```

However, gender specific item locations revealed the situation.

Table III.Ext.3.2

```
                                    ------Boys------    ------Girls ------
+-------+-----+---------+---------+---------+---------+---------+---------+
| ITEM  |BLOCK|  SLOPE  |   S.E.  |LOCATION |   S.E.  |LOCATION |   S.E.  |
+=======+=====+=========+=========+=========+=========+=========+=========+
| Ext1  |   1 |   0.855 |   0.195 |  -0.899 |   0.230 |  -0.633 |   0.207 |
| Ext2  |   1 |   0.576 |   0.142 |  -0.515 |   0.203 |  -0.160 |   0.233 |
| Ext3  |   1 |   0.787 |   0.146 |  -0.914 |   0.357 |  -1.421 |   0.285 |
| Ext4  |   1 |   0.483 |   0.078 |  -1.148 |   0.613 |  -1.380 |   0.593 |
| Ext5  |   1 |   0.693 |   0.074 |   0.172 |   0.347 |  -0.368 |   0.361 |
| Ext6  |   1 |   1.214 |   0.220 |  -1.267 |   0.230 |   0.207 |   0.192 |
| Ext7  |   1 |   1.377 |   0.244 |   1.733 |   0.260 |   1.574 |   0.241 |
| Ext8  |   1 |   0.601 |   0.092 |   0.193 |   0.278 |  -0.953 |   0.286 |
| Ext9  |   1 |   0.390 |   0.081 |   0.794 |   0.334 |   1.380 |   0.323 |
| Ext10 |   1 |   0.706 |   0.100 |  -0.643 |   0.303 |  -0.133 |   0.284 |
| Ext11 |   1 |   1.003 |   0.124 |   0.885 |   0.288 |   0.280 |   0.314 |
+------+-----+---------+---------+---------+---------+---------+---------+
```

Fig. III.Ext.3.1.



A systematic difference between boys and girls can be statistically demonstrated (Ext6 p<0.001 and Ext8 p≈ 0.002). As the difference is reversed concerning the two items, this difference might well be hidden in a raw score approach.

Relative item information

Based on a GRM approach (with gender separation), with item specific slopes and a reduced scale (1,2,3,4,5) => (3,3,3,4,5), the distribution of item information corresponds roughly to the C.I.:s in fig. III.Ext.3.1. (A wide C.I. corresponds to low relative item information).

Table III.Ext.3.3. Relative item information (%)

Item	Rel.Info
Ext1	12.8
Ext2	11.2
Ext3	8.5
Ext4	4.1
Ext5	6.5
Ext6	12.2
Ext7	16.3
Ext8	5.9
Ext9	4.7
Ext10	7.7
Ext11	10.1

Conclusion about the ‘Externalising Behaviour’ questionnaire

- The scale based on a sum score is weak to moderate and seems more suitable for girls than for boys (a somewhat larger scalability for girls). Ranking seems more reliable for girls than for boys.
- A parsimonious Rasch RSM is not sufficient due to a marked different item functioning concerning ‘Destroys things voluntarily’ and ‘Outburst of anger’. Gender split of these items improves the questionnaire to some extent. Reliable person measures, based on raw sum scores, will probably be difficult to achieve.
- Transformation to an interval scaled variable, based on raw sum score or by weighting the items, is not straightforward and difficult to achieve.
- Item discrimination and some gender specific item locations seem necessary.
- Reformulation of the items is recommended to achieve a better coverage.
- Reformulation of the items, particularly Ext6 and Ext8, is needed if the objective is to achieve a questionnaire equally valid for boys and girls.

Investigation of the ‘Internalising Behaviour’ questionnaire (Int)

Step 1 Int. The Mokken scale analysis

Fig. III.Int.1.1. The sum scores from 131 adolescents.

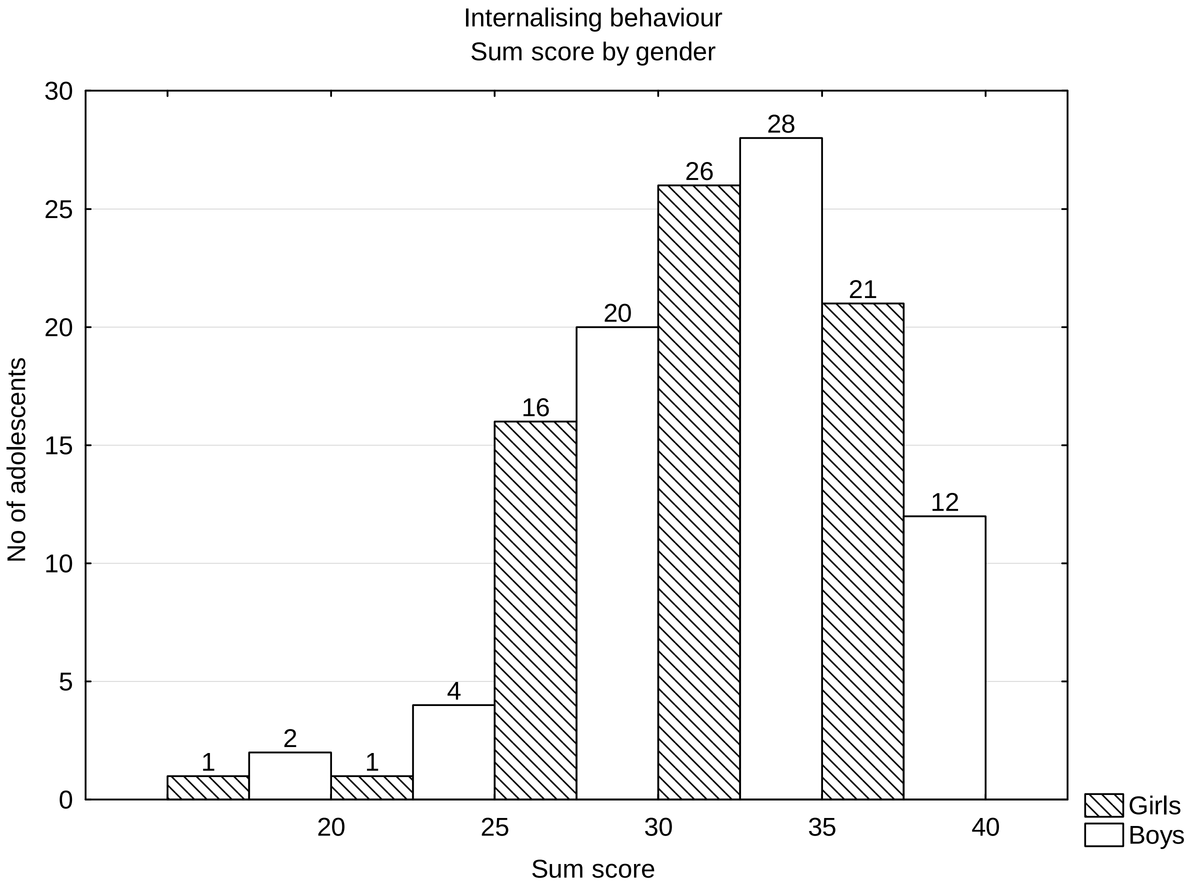


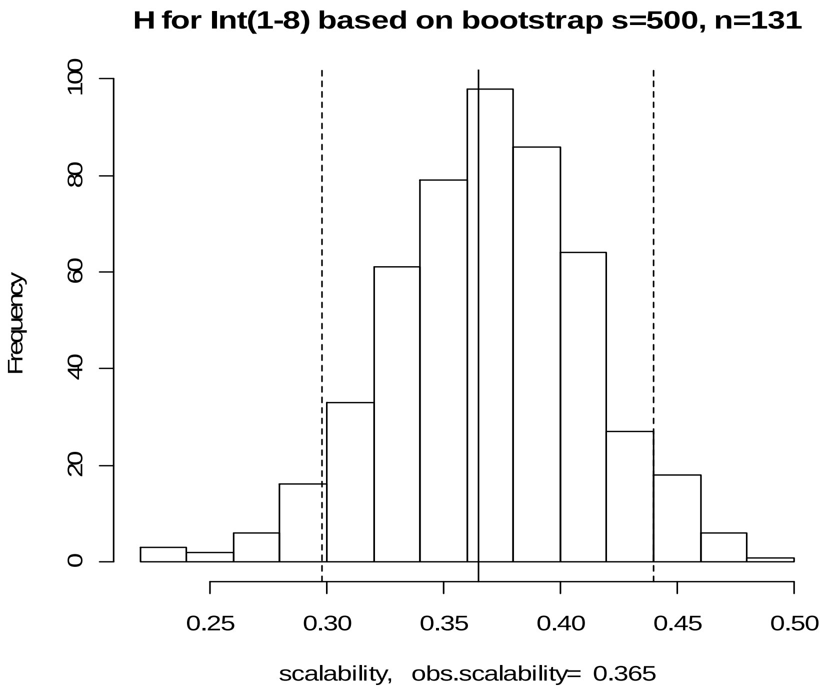
Table III.Int.1.1. Pairwise and gender specific item scalabilities.
Scalabilities <0.2 in bold.

	Int1	Int2	Int3	Int4	Int5	Int6	Int7	Int8
Int1	1.000	0.426	**0.175**	0.537	0.473	0.447	**0.151**	0.439
Int2	0.426	1.000	**0.078**	0.461	0.517	0.606	0.277	0.411
Int3	0.175	0.078	1.000	0.301	**0.091**	0.238	0.342	**0.175**
Int4	0.537	0.461	0.301	1.000	0.350	0.558	0.229	0.702
Int5	0.473	0.517	0.091	0.350	1.000	0.353	0.296	0.331
Int6	0.447	0.606	0.238	0.558	0.353	1.000	0.347	0.620
Int7	0.151	0.277	0.342	0.229	0.296	0.347	1.000	0.300
Int8	0.439	0.411	0.175	0.702	0.331	0.620	0.300	1.000

Item	all	girls	boys
Int1	0.375	0.355	0.373
Int2	0.397	0.310	0.447
Int3	**0.201**	**0.168**	**0.248**
Int4	0.448	0.437	0.439
Int5	0.346	0.396	0.311
Int6	0.451	0.430	0.446
Int7	0.279	0.212	0.380
Int8	0.427	0.386	0.445
Total	0.365	0.333	0.387

All scalabilities are positive and no difference between genders is indicated. Int3 is a potentially weak item with an item scalability = 0.201.

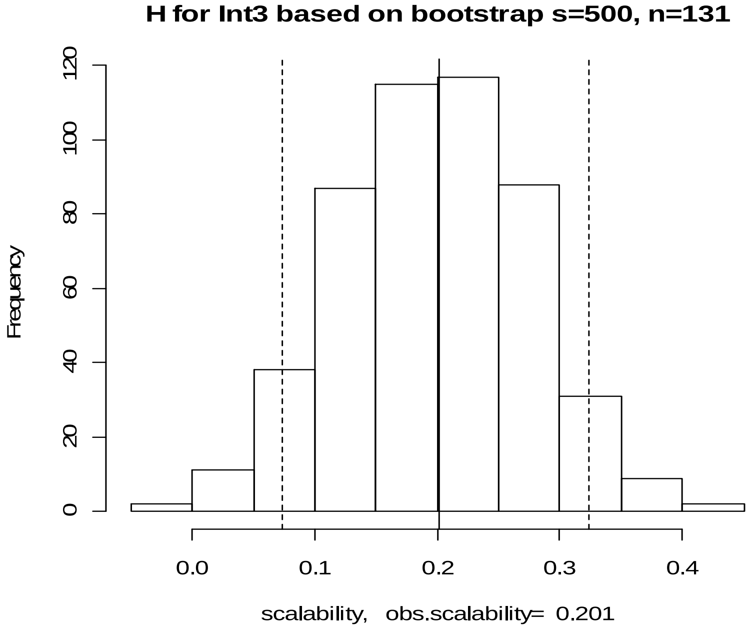
Fig. III.Int.1.1.



```
Min.   1st Qu.  Median  Mean   3rd Qu.   Max.      5%      95%
0.225   0.343   0.370   0.369   0.396   0.495   0.298   0.440
```

Based on fig. III.Int.1.1. we are convinced of an at least a weak scale to allow a reasonable ranking of persons.

Fig. III.Int.1.2. A bootstrapped C.I.for Int3.



```
  Min.   1st Qu.  Median  Mean   3rd Qu.   Max.      5%     95%
 -0.019   0.145   0.199   0.198   0.252   0.435   0.074   0.324
```

Int3 is a weak item but the analysis, show in Fig. III.Int.1.2., presents a C.I. well on the positive side, and is as such not counterproductive.

Monotonicity and non-intersection

There were no problems with the monotonicity but Int3 and Int6 showed significant violations against non-intersection for all reasonable minimum group sizes, n= 20,25,30. This can be expected regarding Int3 due its low scalability, but is more cumbersome for Int6 with its scalability= 0.451. This may cause problems in further analyses.

Potentially influential persons indicated by jackknifed item set scalabilities:
Excluding ID= 64 caused a high scalability. This person may cause problems in the following steps.

A parametric model might well be reasonable.

Step 2 Int. Analysis by a Rasch model

In a first analysis, presented in table III.Int.2.1., Int3 appears as a weak item. No extremely incoherent profile was found, even if ID=64 showed the most poorly fit statistic. There were negligible violations for Int4 against the ordering of categories, due to sparse data.

Table III.Int.2.1. Estimates from a Rasch RSM

```
-----------------------------------------------------
|          MODEL|   INFIT  |EXACT MATCH|ESTIM|      |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|---------------+----------+-----------+-----+------|
|   -.70     .15|1.07    .6| 59.4  60.1| 1.00| Int1 |
|   -.10     .13|1.07    .6| 53.9  55.7|  .99| Int2 |
|    .56     .12|1.44   3.2| 46.9  50.7|  .50| Int3 |
|   -.08     .13| .72  -2.3| 61.7  55.7| 1.30| Int4 |
|   -.72     .15|1.15   1.1| 62.5  60.2|  .93| Int5 |
|    .21     .13| .86  -1.1| 60.2  53.6| 1.18| Int6 |
|   1.16     .12|1.06    .5| 55.5  49.0|  .92| Int7 |
|   -.35     .14| .74  -2.1| 66.4  57.6| 1.25| Int8 |
|---------------+----------+-----------+-----+------|
|Mean .00      |Person reliability = 0.72
|S.D. .60      |  Item reliability = 0.95
---------------
```

Person reliability= 0.72 is weak while item reliability= 0.95 is sufficient.

Table III.Int.2.2. DIF between genders

```
-----------------------------------------------------
| PERSON     SUMMARY DIF                            |
| CLASSES    CHI-SQUARE   D.F.  PROB.       ITEM    |
|---------------------------------------------------|
|       2        3.2581      1  .0711       Int1    |
|       2         .7526      1  .3857       Int2    |
|       2        6.3289      1  .0119       Int3    |
|       2         .2073      1  .6489       Int4    |
|       2        1.5181      1  .2179       Int5    |
|       2        2.7738      1  .0958       Int6    |
|       2        7.7474      1  .0054       Int7    |
|       2        5.8921      1  .0152       Int8    |
-----------------------------------------------------
```

Int3, Int7 and Int8 show significant gender DIFs, see table III.Int.2.2., and there are substantial negative correlations between the residuals.

Item specific thresholds do not solve the problem. As Int3 shows very low item discrimination, as estimated after fitting the Rasch model, a first attempt would be to exclude the item. This did not change the situation. There were still systematic DIF:s for Int7 and Int8. Splitting Int3, Int7 and Int8 into gender specific items improved the model to some extent, see table III.Int.2.3.

Table III.Int.2.3. Rasch RSM after splitting Int3, Int7 and Int8 into gender specific items.

```
-----------------------------------------------------
|          MODEL|   INFIT   |EXACT MATCH|ESTIM|      |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|---------------+----------+-----------+-----+------|
|   -.84     .15|1.08    .6| 59.4  60.3|  .99| Int1 |
|   -.23     .14|1.09    .7| 54.7  56.4|  .96| Int2 |
|    .90     .18|1.61   2.9| 57.1  52.2|  .33| Int3F|
|   -.21     .14| .74  -2.2| 62.5  56.2| 1.28| Int4 |
|   -.86     .15|1.20   1.5| 62.5  60.9|  .88| Int5 |
|    .09     .13| .86  -1.1| 61.7  54.0| 1.18| Int6 |
|   1.53     .17|1.26   1.4| 49.2  49.5|  .75| Int7F|
|   -.79     .22| .74  -1.4| 73.0  64.3| 1.18| Int8F|
|    .05     .18|1.24   1.3| 41.5  52.1|  .74| Int3M|
|    .64     .17| .80  -1.2| 64.6  49.9| 1.21| Int7M|
|   -.28     .18| .70  -1.8| 61.5  54.3| 1.32| Int8M|
|---------------+----------+-----------+-----+------|
```

Person reliability = 0.76. Item reliability = 0.94.

However, a wider range of item locations was a positive effect of splitting items. Int3 still appears as a poor item. Excluding Int3 made Int7 a bad fitting item.

Table III.Int.2.4. Largest standardized residual correlations used to identify dependent items

```
---------------------------------
|CORREL-|           |           |
|  ATION|     ITEM  |     ITEM  |
|-------+-----------+-----------|
|  -.40 |     Int3F |     Int8F |
|  -.32 |     Int5  |     Int3M |
|  -.31 |     Int1  |     Int8M |
---------------------------------
```

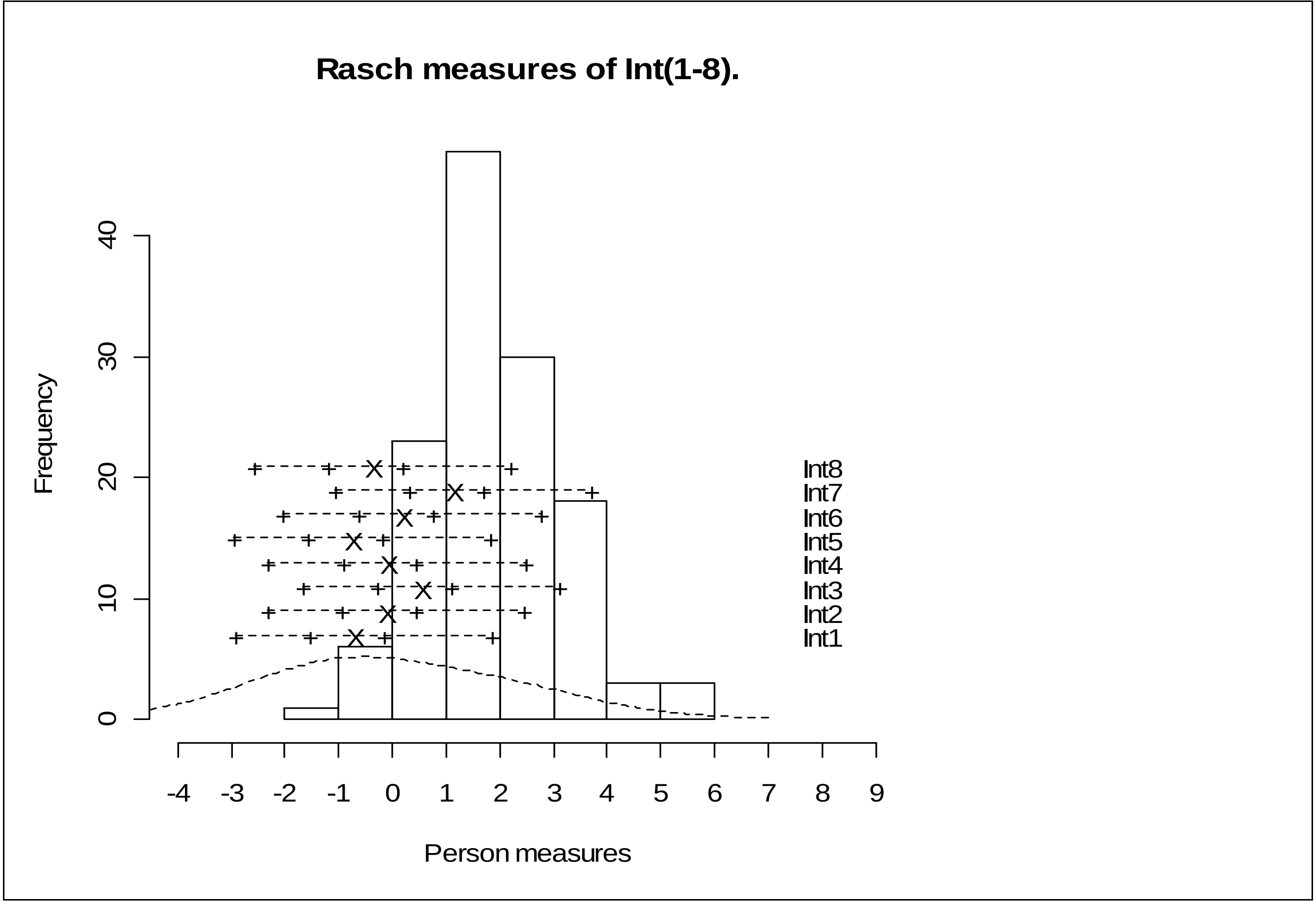
The negative correlation of pairwise residuals can to some extent be traced back to weak item pair scalabilities, i.e. contradicting items. This should not be taken too seriously but many largely negative residual correlations is a sign of a straggling item set.

Table III.Int.2.4. Analysis of residuals after fitting the Rasch RSM.

```
Total raw variance in observations     =         20.0 100.0%
  Raw variance explained by measures   =          9.0  45.1%
    Raw variance explained by persons  =          6.6  33.2%
    Raw Variance explained by items    =          2.4  11.9%
  Raw unexplained variance (total)     =         11.0  54.9%
  Unexplained variance in 1st contrast =          1.8   9.0%
```

The model is unsatisfactory even if there is no immediate concern about any second dimension. The ‘equal slopes’ constraint should be released.

Fig III.Int.2.1. Item locations and category thresholds for the Rasch RSM

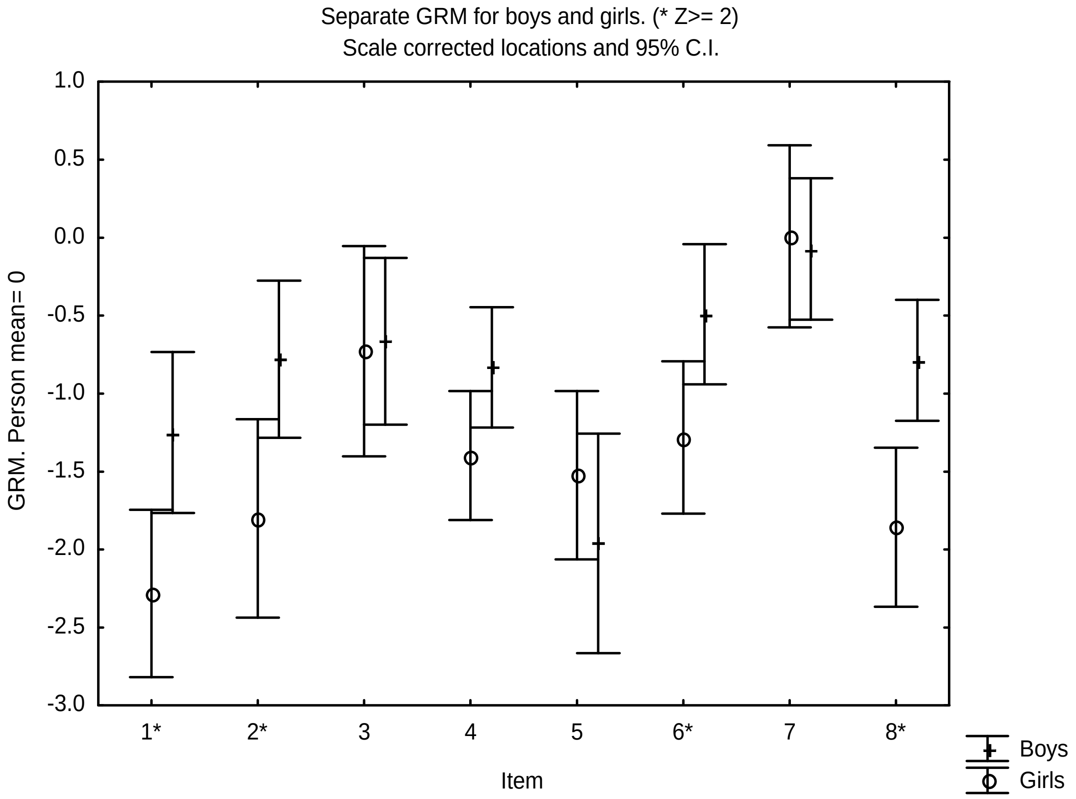


The ‘Internalising questionnaire’ appears too easy with the items locations concentrated to the lower part of the latent scale.

Step 3 Int

A GRM with item specific slopes but no split of genders revealed Int3 and Int6 as bad fitting items. A GRM with gender specific locations but a common (for genders) set of item specific slopes was tried. When splitting the genders, the scale had to be further reduced, (1,2,3,4,5)→ (3,3,3,4,5), due singularity of matrices in the estimation procedure. With the reduced scale, no systematic difference between boys and girls could be statistically demonstrated. However, there was a reasonable model fit for boys but not for girls. There were still some correlations between the item category residuals, but just a few ≥ 0.3 and none > 0.4. The heterogeneous results so far indicate a more general difference between boys and girls and separate gender investigations would be of value for further information.

Fig. III.Int.3.1.



The structure of item locations was quite different for Boys and Girls. An explorative analysis of the item information, based on an unrestricted model, is of interest before a conclusion about the characteristics of the questionnaire.

I.e., for each gender, we consider a model with item specific slopes and thresholds. Although 'over-parameterized' and too data driven, we get unrestricted estimates of the potential item information. The unconstrained model analysis is based on the term $\alpha_i(\theta_n - \delta_{ik})$. See appendix B for further information.

Table III.Int.3.1. Relative item information(%) based on an unrestricted model.

	Relative information %	
Item	Girls	Boys
Int1	18	11
Int2	11	14
Int3	3	6
Int4	17	19
Int5	16	7
Int6	14	16
Int7	5	10
Int8	15	17

Is $\alpha_i(\theta_n - \delta_{ik})$ needed as compared to the more parsimonious $\alpha(\theta_n - \delta_{ik})$, the equal slopes model?
A likelihood ratio test rejects the equal slopes model for the girls when compared to the unconstrained, p=0.002, while just indicated for boys, p= 0.094. Even if p-values cannot be directly compared, it might give a further explanation of the heterogeneous result.

Conclusion about the 'Internalising Behaviour' questionnaire

- The Internalising behaviour questionnaire constitutes a weak scale. Ranking individuals might be reasonably valid.
- No obvious straight forward transformation of the raw sum score to a reasonable interval scale variable was found.
- The item Int3 (Sleeps badly/seems tired) was found to be a weak item and should be reformulated or deleted.
- A more elaborated model did not sufficiently catch the characteristics of the questionnaire. A certain disagreement peeps out in the analyses. A transformation to a plausible interval scale seems, at this stage, difficult to achieve.
- There is a distinct indication that boys and girls perceive the questionnaire differently at certain items. Gender separate analyses/models are indicated. Reformulation of the items (or alternative items) should be considered if a uniform questionnaire for both genders is desired.

Conclusion about the 'Adolescent Adjustment Profile' questionnaire

The intention to represent three specified dimensions seems relevant, particularly for the AttDef dimension. The Ext and Int dimensions seem to be more close to each other. There might be difficulties in the formulation of certain items without a risk of crossing the border to the other dimension, or, the adolescents perceive certain items differently to what was in the author's mind.
The AttDef set of items are fairly homogeneous regarding gender, but for Ext and Int there are obvious problems of formulating a questionnaire equally applicable for boys and girls.
A general problem appears to be the coverage in terms of item locations. The AttDef item locations match with the sample, but are too concentrated. Almost all items cover most of the person measure range. The reason might be the ultimate formulation of the categories which forces the item location

towards the centre. Similar problems of concentration are seen for Ext and Int in combination with a ‘no match’ with the sample. These questionnaires are too easy. It seems a necessary task to extend the coverage of the item locations and, if possible, reformulate the category wordings to less ultimate expressions.

Furthermore, some items seem to be differently perceived by boys and girls. Even if this study is carried out on a small sample, it is strongly indicated that the formulation of items might be gender specific. Gender specific formulations might be included in the questionnaire, but will make it more complicated. If the intention is to create a questionnaire, equally suitable for boys and girls, the items pointed out as gender specific, should be reformulated.

Study IV

Adler M 1§, Hetta J, Isacsson G 1, Brodin U. **An Item Response Theory evaluation of three depression assessment instruments in a clinical sample.** BMC Med Res Methodol. 2012 Jun 21;12(1):84.

This study investigates whether an analysis, based on Item Response Theory (IRT), can be used for initial evaluations of depression assessment instruments in a limited patient sample from an affective disorder outpatient clinic, with the aim to finding major advantages and deficiencies of the instruments. Three depression assessment instruments, the depression module from the Patient Health Questionnaire (PHQ9), the depression subscale of Affective Self Rating Scale (AS-18-D) and the Montgomery-Åsberg Depression Rating Scale (MADRS) were evaluated in a sample of 61 patients with affective disorder diagnoses, mainly bipolar disorder.
Results:
In a first step, the Mokken non-parametric analysis showed that PHQ9 and AS-18-D had strong overall scalabilities, while MADRS showed a weak scalability. In a second step, a Rasch model analysis indicated large differences concerning the item discriminating capacity and was therefore considered not suitable for the data. In the third step, applying a more flexible two parameter model, all three instruments showed large differences in item information and certain items had a low capacity to reliably measure respondents at low levels of depression severity.
The study suggests that the PHQ9 and AS-18-D can be useful for measurement of depression severity in an outpatient clinic for affective disorder, while the MADRS shows weak measurement properties for this type of patients.

The questionnaires are outlined in Appendix A.
The scales:
AS-18-D has 5 categories (0,1,2,3,4). AS-18-D is shortened to ASD with items ASD1 – ASD9 in analyses and tables.
PHQ9 has 4 categories (0,1,2,3). PHQ9 is shortened to PHQ with items PHQ1 – PHQ9 in analyses and tables.
MADRS has 7 categories (0,1,2,3,4,5,6)
The observed frequencies are presented in Table 1 in the article.
In this study there are three questionnaires estimating the same phenomenon based on a common sample. Thus the person mean is set to zero, which relates the three questionnaires to each other concerning the item locations.
The first index in tables and figures, IV, indicates study IV.
The second index in tables and figures represent the questionnaire.
AS-18-D = ASD
PHQ9 = PHQ
MADRS = MADRS
The third index Step 1, 2 or 3.
The fourth index indicates table or figure no. within Step.

Investigation of the AS-18-D questionnaire (ASD).

Sixty one persons answered to the nine item ASD-questionnaire, see Appendix A, with 5 ordered answer categories for each item.

Table IV.ASD.0.1.

Frequency Table. ASD depression dimension 9 items scored by 61 subjects.							
Item	Score 0	Score 1	Score 2	Score 3	Score 4	Missing Data	Row Totals
ASD1	26	11	10	9	5	0	61
ASD2	11	12	11	15	11	1	61
ASD3	25	9	12	12	3	0	61
ASD4	14	8	14	17	8	0	61
ASD5	16	7	14	14	9	1	61
ASD6	14	10	13	15	9	0	61
ASD7	17	8	11	12	13	0	61
ASD8	22	11	15	7	5	1	61
ASD9	37	12	8	2	1	1	61
All Items	182	88	108	103	64	4	

The distribution of answer is fairly spread over the five categories with a certain predominance for score 0. There were just 4 non-responses.

Step 1 ASD. The Mokken scale analysis

Table IV.ASD.1.1.

Item pair scalabilities (H_{ij} < 0.2 in bold, H_{ij} >0.7 in italic)

	ASD1	ASD2	ASD3	ASD4	ASD5	ASD6	ASD7	ASD8	ASD9
ASD1	1.000	**0.183**	0.456	0.264	0.515	0.449	**0.137**	0.297	0.322
ASD2	**0.183**	1.000	0.418	***0.902***	***0.713***	0.406	***0.872***	0.484	***0.732***
ASD3	0.456	0.418	1.000	0.319	0.381	***0.708***	0.387	0.639	0.343
ASD4	0.264	0.902	0.319	1.000	***0.770***	0.360	***0.805***	0.434	***0.757***
ASD5	0.515	0.713	0.381	0.770	1.000	0.413	0.618	0.415	0.640
ASD6	0.449	0.406	0.708	0.360	0.413	1.000	0.393	***0.729***	0.535
ASD7	**0.137**	0.872	0.387	0.805	0.618	0.393	1.000	0.493	0.680
ASD8	0.297	0.484	0.639	0.434	0.415	0.729	0.493	1.000	0.680
ASD9	0.322	0.732	0.343	0.757	0.640	0.535	0.680	0.680	1.000

Item scalabilities

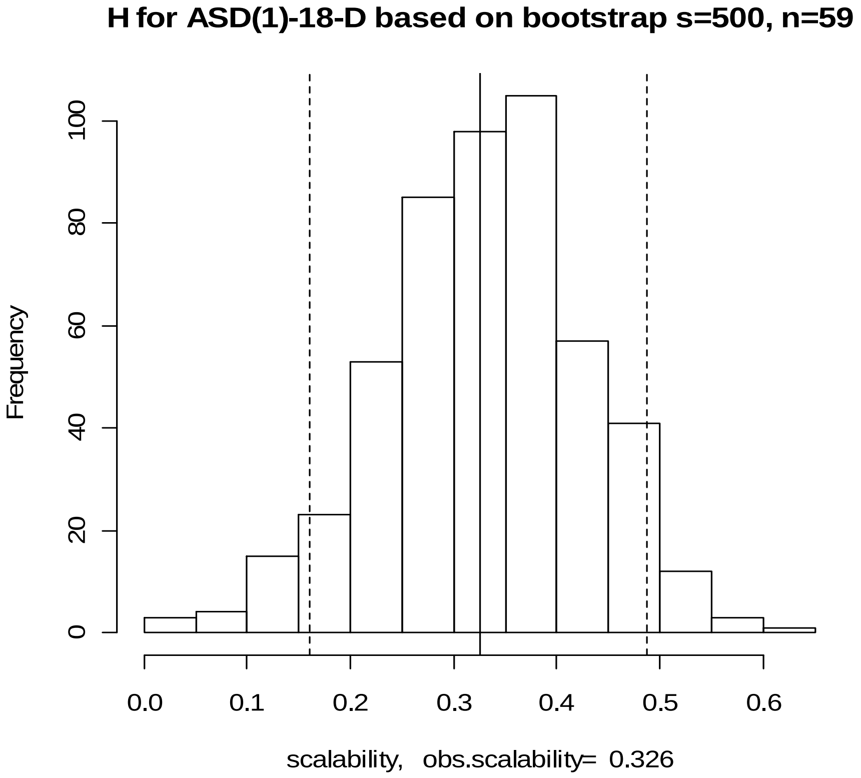
```
 ASD1  ASD2  ASD3  ASD4  ASD5  ASD6  ASD7  ASD8  ASD9
0.326 0.590 0.460 0.576 0.557 0.491 0.543 0.513 0.582
```

Scalability for the total item set=   0.513

There was no sign of multiple dimensionalities as evaluated by the Mokken dimensionality analysis. The item scalability estimates are in the range 0.4 – 0.6, well above 0.3 besides ASD1.

The item set scalability, 0.513, 90% C.I. (0.408, 0.633) implies an at least moderate scale. Just a few and negligible violations against monotonicity were seen. A set of large pair scalabilities is observed in table IV.ASD.1.1, particularly between ASD2 and ASD4,ASD7. Their signification will be further revealed in Step2 and 3.

Fig. IV.ASD.1.1. A bootstrapped C.I. for ASD1



```
  Min.  1st Qu.  Median   Mean   3rd Qu.   Max.     5%      95%
0.1120  0.2778  0.3245  0.3241  0.3732  0.4980  0.1858  0.4351
```

The C.I. for the weakest item, ASD1, is well on the positive side, with a lower limit of 0.19.

One person was identified as potentially influential, identified by the jackknife method. Excluding this person (no 24 in the person data file) yielded an item set scalability H= 0.488, which, however, is not far from H= 0.513.

However, there were a number of small violations against non-intersection, but none significant. The reason might be that the categories are too close to each other, which means a certain difficulty in ordering. But on the whole, the persons can be reasonably ranked by the raw sum score.

A parametric Rasch model for person estimates on an interval scale seems reasonable.

Step 2 ASD. Analysis by a Rasch model.

All persons and items are included in a first Rasch RSM analysis.

Table IV.ASD.2.1. Location estimates by a Rasch RSM.

```
-------------------------------------------------------
|            MODEL|   INFIT   |EXACT MATCH|ESTIM|      |
|   MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|-----------------+----------+-----------+-----+------|
|      .74     .14|1.74   3.4| 26.3  41.2|  .23| ASD1 |
|     -.13     .14| .79  -1.2| 43.9  40.2| 1.20| ASD2 |
|      .68     .14|1.12    .7| 38.6  41.5|  .95| ASD3 |
|     -.05     .14| .75  -1.5| 42.1  40.2| 1.31| ASD4 |
|      .04     .14| .87   -.7| 51.8  40.1| 1.25| ASD5 |
|     -.01     .14|1.06    .4| 49.1  40.0|  .85| ASD6 |
|     -.03     .14|1.04    .3| 35.1  40.0| 1.11| ASD7 |
|      .62     .14| .91   -.4| 44.6  40.5| 1.05| ASD8 |
|     1.72     .18| .85   -.6| 66.1  56.8| 1.05| ASD9 |
|-----------------+----------+-----------+-----+------|
|Mean  .40       | Person reliability = 0.83
|S.D.  .57       |   Item reliability = 0.93
----------------
```

From Table IV.ASD.2.1. it is noticed that ASD1 – ASD8 have locations within a narrow range while ASD9 is estimated at a much higher level. Furthermore, ASD1 appears to be a weak, not very contributing item. A ‘MNSQ>1.5’ indicates more noise than information.

Table IV.ASD.2.2. Standardised residual variance (in Eigenvalue units)

```
                                                    -- Empirical
Total raw variance in observations     =        19.9 100.0%
  Raw variance explained by measures   =        10.9  54.8%
    Raw variance explained by persons  =         7.8  39.3%
    Raw Variance explained by items    =         3.1  15.5%
  Raw unexplained variance (total)     =         9.0  45.2%
  Unexplained variance in 1st contrast =         3.2  16.2%
```

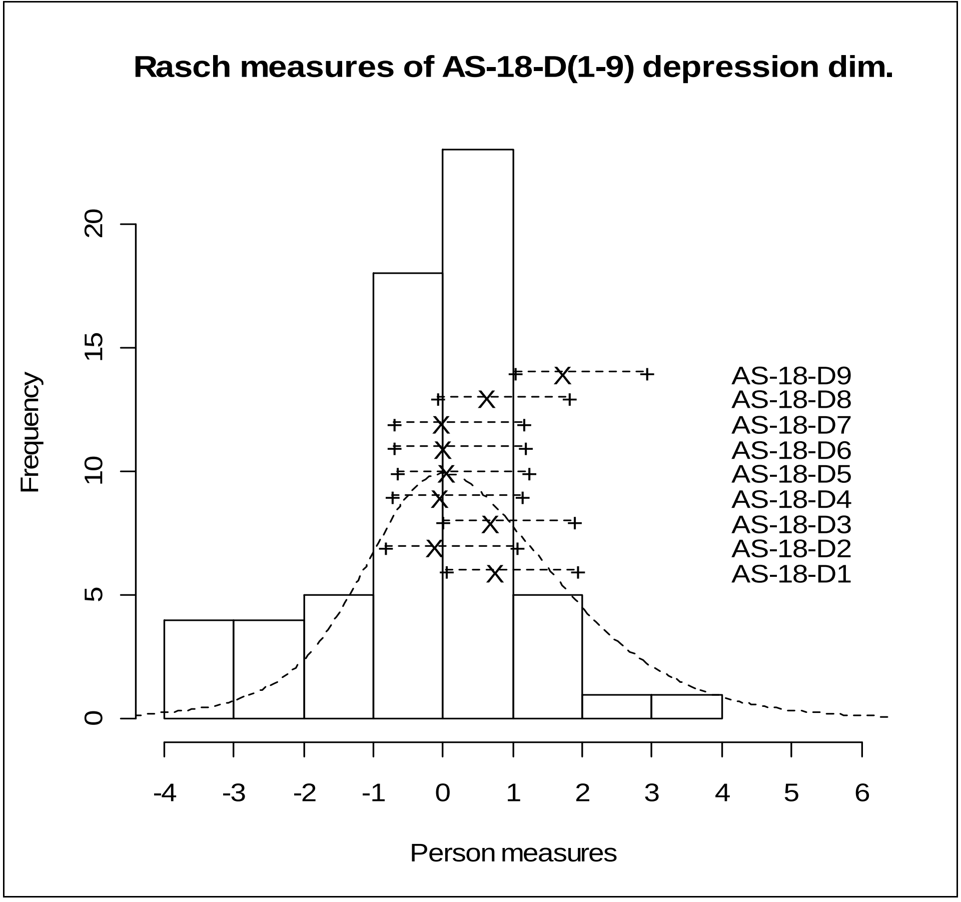
The infit MNSQ and estimated item discrimination for ASD1 as well as an unexplained variance in 1:st contrast, >3, indicate that the parsimonious Rasch model might not be sufficient. There were just a few minor violations against the ordering of categories. Exclusion of ASD1 raised the variance explained by measures to 60% but otherwise, the result was only slightly changed.

Table IV.ASD.2.3. Largest standardized residual correlations used to identify dependent items.

```
----------------------
|CORREL-|     |      |
|  ATION|ITEM | ITEM |
|-------+-----+------|
|    .52 |ASD2 | ASD7 |
|    .45 |ASD2 | ASD4 |
|-------+-----+------|
|   -.52 |ASD2 | ASD6 |
|   -.51 |ASD3 | ASD4 |
|   -.47 |ASD1 | ASD7 |
|   -.44 |ASD4 | ASD6 |
|   -.44 |ASD3 | ASD9 |
|   -.43 |ASD1 | ASD2 |
|   -.42 |ASD6 | ASD7 |
|   -.42 |ASD4 | ASD8 |
----------------------
```

Furthermore, there are too many large correlations for a reasonable Rasch model. A model with separate item thresholds does not change the result. Even if the Rasch model is not suitable, it might be of interest to see the result if we 'force' the model onto the data. Insufficient coverage and a peaked information is indicated, see fig**.** IV.ASD.2.1 .

Fig**.** IV.ASD.2.1. Estimates from a Rasch RSM



Step 3 ASD

A graded response model (GRM) with common slopes and a common set of category thresholds is, as expected, similar to the Rasch approach and reveals the insufficient coverage in terms of item locations. This constrained model could not be statistically rejected. However, the common slopes constraint might hide an underlying structure of different slopes, particularly for ASD1 (as was seen in step2).

A GRM with item specific slopes and a common set of category thresholds reveals a large variability in the slopes (table IV.ASD.3.1.), even if the S.E. for some of the item slopes are rather large.

Table IV.ASD.3.1.

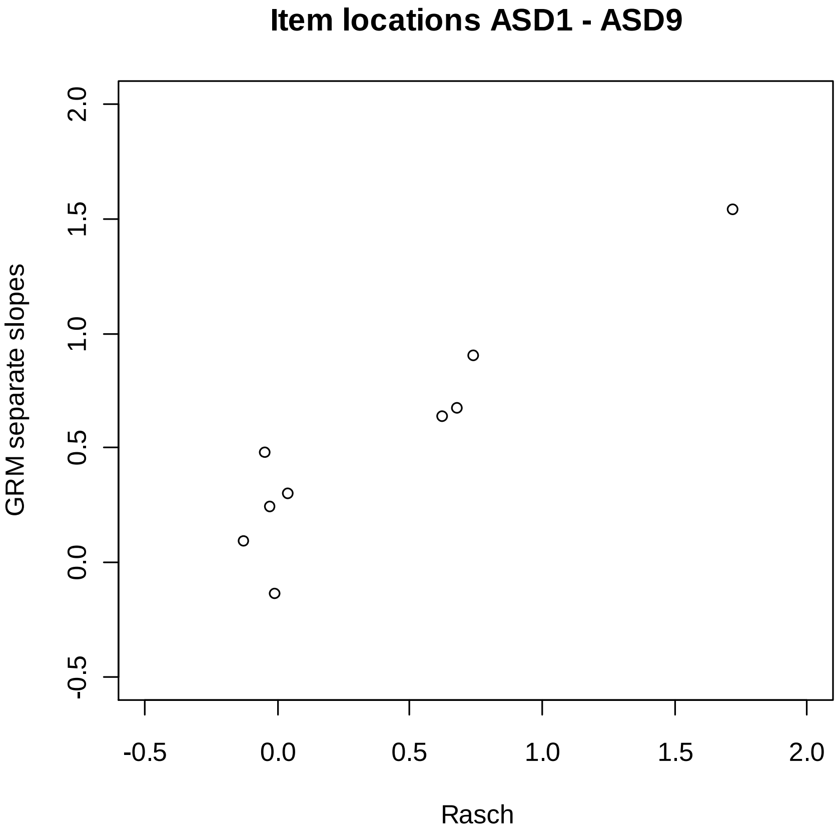
```
CATEGORY PARAMETER  :     1.021     0.439    -0.276    -1.183
              S.E.  :     0.056     0.051     0.052     0.069
+------+---------+---------+---------+---------+
| ITEM |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+======+=========+=========+=========+=========+
| ASD1 |   0.574 |   0.188 |   0.902 |   0.354 |
| ASD2 |   2.845 |   0.723 |   0.096 |   0.164 |
| ASD3 |   0.923 |   0.220 |   0.677 |   0.324 |
| ASD4 |   1.256 |   0.908 |   0.479 |   0.182 |
| ASD5 |   1.751 |   0.274 |   0.299 |   0.221 |
| ASD6 |   0.861 |   0.189 |  -0.134 |   0.303 |
| ASD7 |   1.313 |   0.331 |   0.241 |   0.187 |
| ASD8 |   1.102 |   0.253 |   0.635 |   0.276 |
| ASD9 |   1.710 |   0.416 |   1.542 |   0.239 |
+------+---------+---------+---------+---------+
```

The item fit statistics (table IV.ASD.3.2.) are not perfect but reasonable. The Total Chi-square p-value is probably too pessimistic due to a certain amount of remaining correlations between items.

Table IV.ASD.3.2. ITEM FIT STATISTICS

```
 ------------------------------------
 | ITEM | CHI-SQUARE |  D.F. | PROB. |
 ------------------------------------
 | ASD1 |    6.17553 |    5. | 0.289 |
 | ASD2 |    8.62574 |    5. | 0.124 |
 | ASD3 |    4.30854 |    5. | 0.507 |
 | ASD4 |   12.07294 |    5. | 0.034 |
 | ASD5 |    6.55808 |    6. | 0.364 |
 | ASD6 |    8.94030 |    5. | 0.110 |
 | ASD7 |   16.65259 |    6. | 0.011 |
 | ASD8 |    3.17387 |    5. | 0.676 |
 | ASD9 |    3.17016 |    5. | 0.676 |
 ------------------------------------
 |Total |   69.67776 |   47. | 0.018 |
 ------------------------------------
```

Fig. IV.ASD.3.1. Item locations based on the Rasch RSM and a GRM with item specific slopes.



The ‘blocking’ of item locations from the Rasch model is clearly illustrated in fig. IV.ASD.3.1. This blocking was broken up to some extent by the GRM approach. The order of the items, illustrated in fig. IV.ASD.3.1., is very similar for the Rasch and the GRM model. The order of the items, in terms of ASD’nr’ is: Rasch: 2, 4, 7. 6, 5, 8, 3, 1, 9   vs   GRM: 6, 2, 7, 5, 4, 8, 3, 1, 9.
The common slopes GRM model and a GRM with separate slopes yielded approximately the same results regarding the item locations.

Table IV.ASD.3.3. Correlation of residuals from a GRM with separate slopes..

	ASD1	ASD2	ASD3	ASD4	ASD5	ASD6	ASD7	ASD8	ASD9
ASD1	1.00	**-0.47**	**0.33**	-0.13	**0.3**5	0.28	**-0.31**	0.09	0.09
ASD2	-0.47	1.00	**-0.35**	**0.36**	-0.25	**-0.44**	**0.35**	**-0.43**	-0.28
ASD3	0.33	-0.35	1.00	**-0.40**	-0.14	**0.51**	-0.18	**0.42**	-0.19
ASD4	-0.13	0.36	-0.40	1.00	0.20	-0.30	0.27	**-0.35**	-0.08
ASD5	0.35	-0.25	-0.14	0.20	1.00	-0.09	-0.17	-0.24	-0.14
ASD6	0.28	-0.44	0.51	-0.30	-0.09	1.00	-0.18	**0.51**	0.05
ASD7	-0.31	0.35	-0.18	0.27	-0.17	-0.18	1.00	-0.14	-0.19
ASD8	0.09	-0.43	0.42	-0.35	-0.24	0.51	-0.14	1.00	0.22
ASD9	0.09	-0.28	-0.19	-0.08	-0.14	0.05	-0.19	0.22	1.00

There is still a large amount of correlation between item residuals. ASD6 and ASD8 have about 25% of explained variation in common, cor(ASD6, ASD8) = 0.51 in table IV.ASD.3.3.

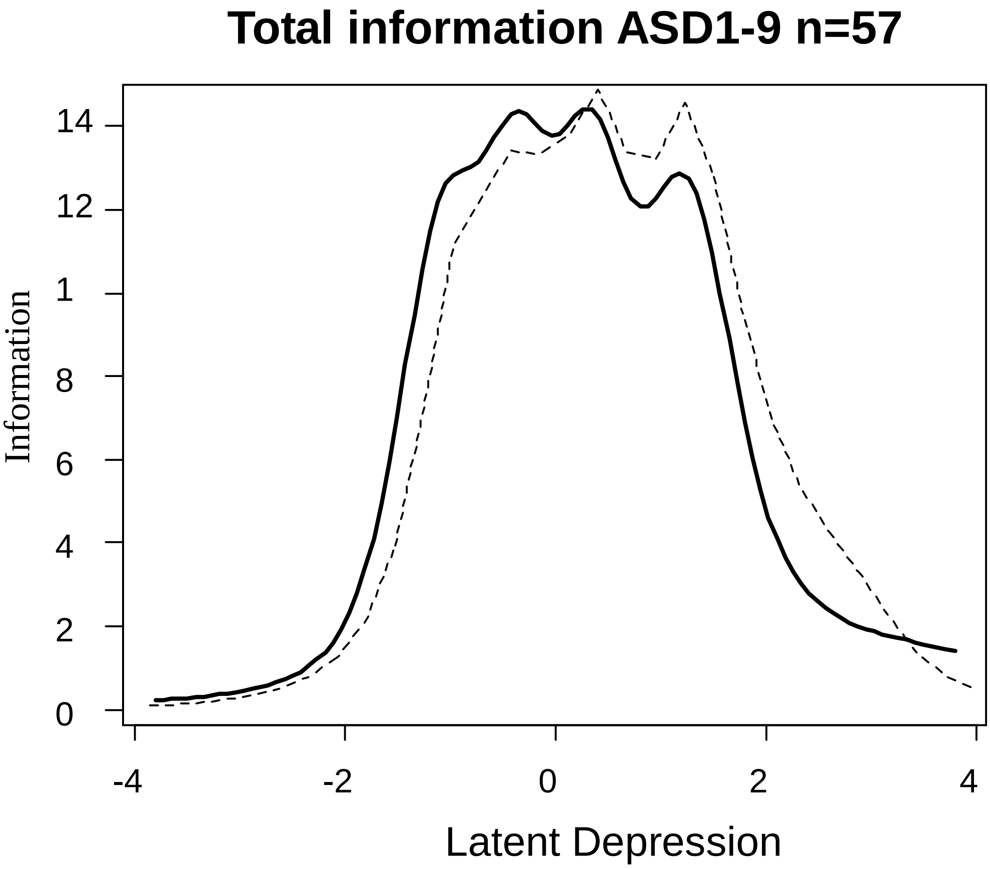
The GRM also reveals a very small amount of information from ASD1.
The information can be further investigated by an unconstrained model with separate thresholds and separate slopes, based on the expression $\alpha_i[\ \theta_n - (\delta_i + \tau_{ik})]$.
This can be interpreted as 'α_i is released from the common thresholds restriction'

Table IV.ASD.3.4. Relative item information(%). GRM with common category thresholds(constrained) vs an unconstrained model.

Item	rel.info. % GRM constr.	Rel.info. % GRM unconstr.
ASD1	**1.9**	**2.7**
ASD2	27.3	30.0
ASD3	4.2	5.8
ASD4	19.4	9.2
ASD5	11.5	15.2
ASD6	4.8	5.2
ASD7	15.3	9.9
ASD8	5.6	7.6
ASD9	9.9	14.5

Even if the unconstrained and the constrained model agree only partially, they are in agreement about ASD1, identified as a weak, not very contributing item.

Fig. IV.ASD.3.2. Test information based on GRM, the constrained and the unconstrained model.



The test information based on the constrained model (dashed line) is almost identical (when rescaled) with the information based on the unconstrained model (solid line). In essence, there is good information in the range ≈ [-1.8, 1.8]. The information (the y- scale) is not very informative; the focus is on the shape of the curve.

A Likelihood Ratio test of the a model based on $\alpha[\theta_n - (\delta_i + \tau_{ik})]$ against the unconstrained model based on $\alpha_i[\theta_n - (\delta_i + \tau_{ik})]$ showed the following result:

```
                    AIC     BIC log.Lik   LRT df p.value
constr_model   1468.73 1546.83 -697.36
unconstr_model 1451.68 1546.67 -680.84 33.05  8  <0.001
```

The unconstrained model is expected show a better fit but is too ‘data driven’. It only serves to bring light upon ‘unconstrained’ item information.

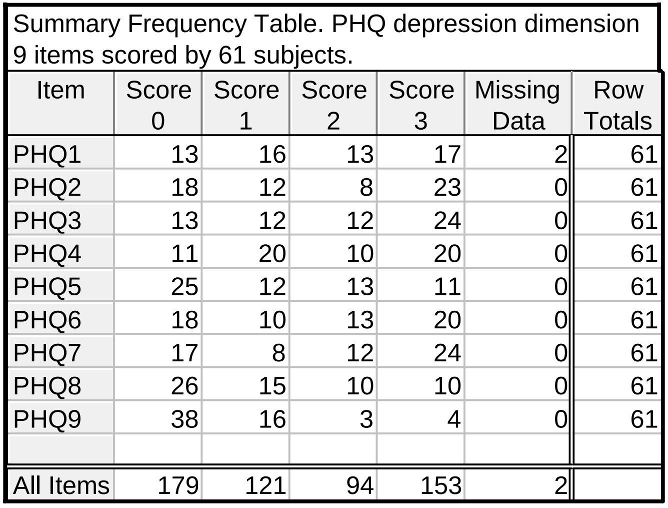
<u>Conclusion about the ASD-18-D questionnaire</u>

- A moderate scale which allows ranking of the persons based on the raw sum score.
- ASD1 is a weak item. Although not very informative, it does not deteriorate the questionnaire. This item should be reformulated.
- A GRM with item specific slopes should be considered when further patients are included, even if a simpler model could not be formally rejected at this phase.
- Enlarge the ‘coverage’ by additional items.

Investigations of the PHQ questionnaire (PHQ)

Sixty-one persons answered to the nine item PHQ-questionnaire, see Appendix I, with 4 ordered answer categories for each item. The score frequencies are presented in table IV.PHQ.0.1.

Table IV.PHQ.0.1.



Summary Frequency Table. PHQ depression dimension
9 items scored by 61 subjects.

Item	Score 0	Score 1	Score 2	Score 3	Missing Data	Row Totals
PHQ1	13	16	13	17	2	61
PHQ2	18	12	8	23	0	61
PHQ3	13	12	12	24	0	61
PHQ4	11	20	10	20	0	61
PHQ5	25	12	13	11	0	61
PHQ6	18	10	13	20	0	61
PHQ7	17	8	12	24	0	61
PHQ8	26	15	10	10	0	61
PHQ9	38	16	3	4	0	61
All Items	179	121	94	153	2	

The responses to the items are fairly distributed over the four categories, with only two non-responses for PHQ1.

Step1 PHQ. The Mokken scale analysis

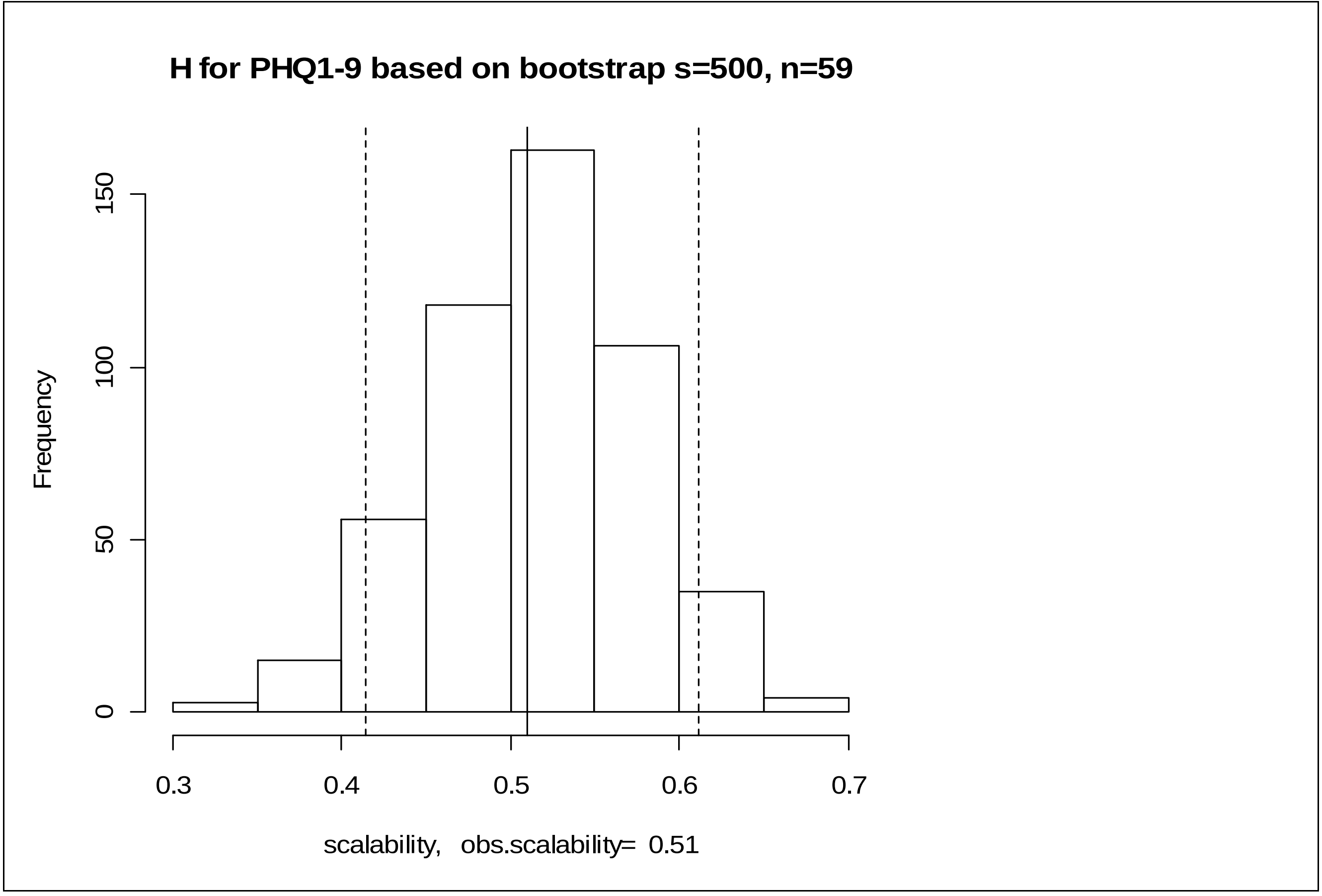
The analysis is based on 59 complete questionnaires.

Table IV.PHQ.1.1. Item pair scalabilities

```
     PHQ1  PHQ2  PHQ3  PHQ4  PHQ5  PHQ6  PHQ7  PHQ8  PHQ9
PHQ1 1.000 0.772 0.397 0.570 0.383 0.779 0.640 0.504 0.595
PHQ2 0.772 1.000 0.559 0.597 0.414 0.817 0.548 0.401 0.742
PHQ3 0.397 0.559 1.000 0.464 0.444 0.496 0.521 0.387 0.613
PHQ4 0.570 0.597 0.464 1.000 0.485 0.583 0.493 0.196 0.345
PHQ5 0.383 0.414 0.444 0.485 1.000 0.323 0.484 0.323 0.140
PHQ6 0.779 0.817 0.496 0.583 0.323 1.000 0.666 0.389 0.618
PHQ7 0.640 0.548 0.521 0.493 0.484 0.666 1.000 0.683 0.461
PHQ8 0.504 0.401 0.387 0.196 0.323 0.389 0.683 1.000 0.280
PHQ9 0.595 0.742 0.613 0.345 0.140 0.618 0.461 0.280 1.000

Item scalability
 PHQ1  PHQ2  PHQ3  PHQ4  PHQ5  PHQ6  PHQ7  PHQ8  PHQ9
0.584 0.603 0.481 0.478 0.385 0.588 0.567 0.400 0.468

Item set scalability
 H= 0.51 with a bootstrap C.I. [0.414, 0.612], see fig. IV.PHQ.1.1.
```

The PHQ scale does not exhibit any weak parts besides a few item pair scalabilities, which are <0.2 (marked with bold in table IV.PHQ.1.1.). Even the smallest item scalability (PHQ5) presents a satisfactory C.I. [0.238, 0.535]. The scale is considered as moderate to strong and indicates that the persons can be ranked reasonably according the raw sum score. A dimension analysis did not indicate any obvious second dimension.

Fig. IV.PHQ.1.1. Bootstrapped C.I.for the item set PHQ.



Monotonicity and non-intersection

Only a few, negligible violations against monotonicity was observed.
However, there are some problems with the non-intersection.

Table IV.PHQ.1.2. Analysis of violations against non intersection
Minimum groupsize =15

	ItemH	#vi	maxvi	sum	zmax	#zsig
PHQ1	0.58	12	0.18	1.37	1.17	0
PHQ2	**0.60**	**26**	**0.29**	**3.00**	**1.87**	**1**
PHQ3	0.48	18	0.20	1.61	1.21	0
PHQ4	0.48	21	0.13	1.70	0.69	0
PHQ5	0.38	18	0.22	1.66	1.34	0
PHQ6	0.59	15	0.18	1.18	1.17	0
PHQ7	0.57	14	0.18	1.24	1.17	0
PHQ8	**0.40**	**15**	**0.29**	**1.56**	**1.87**	**1**
PHQ9	0.47	9	0.15	0.77	0.86	0

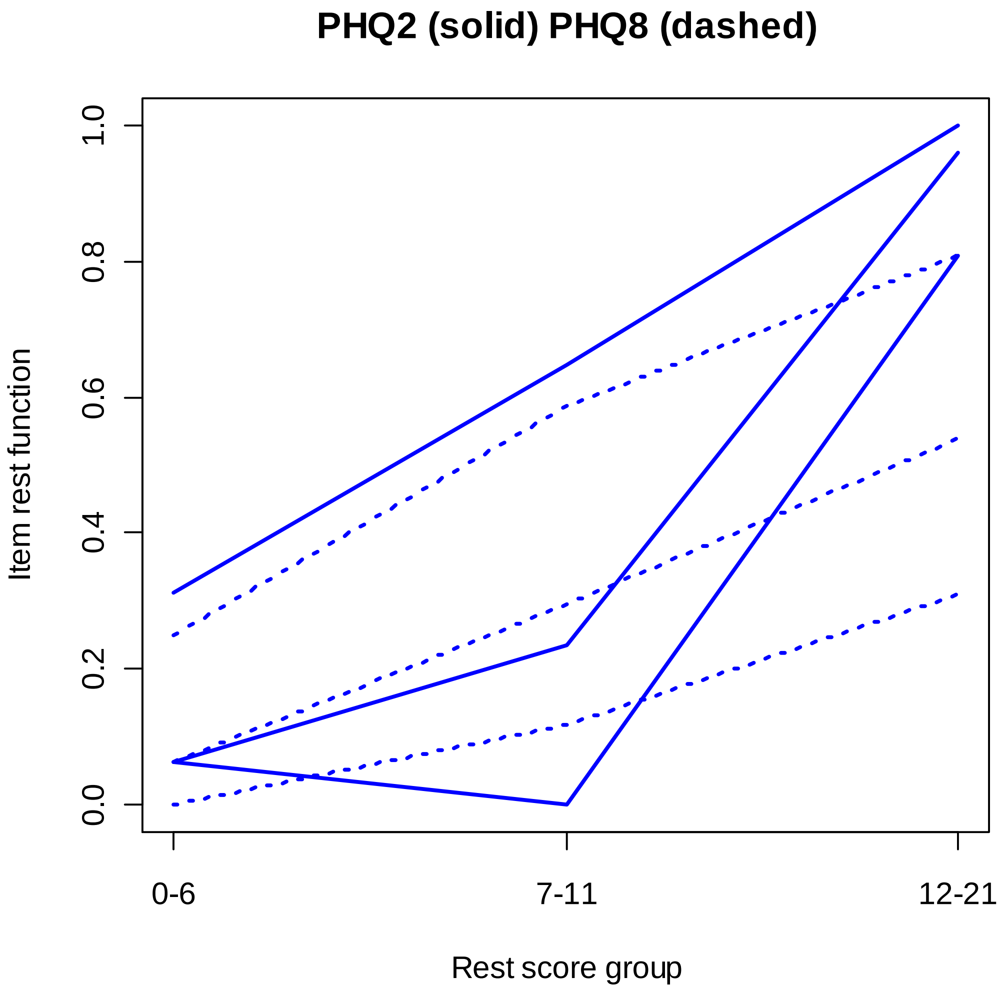
There are a number of small violations, significant for PHQ2 and PHQ8.
This might well be a sign of a too narrow set of items with respect to their item locations when a parametric model with location estimates will be tried. Table IV.PHQ.1.3. is a rough illustration of the narrow range of the item locations as expressed by the sum scores for the items.

Table IV.PHQ.1.3. The sum score per item

```
  PHQ1 PHQ2 PHQ3 PHQ4 PHQ5 PHQ6 PHQ7 PHQ8 PHQ9
   93   94  103   97   70   92  104   65   33
```

As PHQ2 and PHQ8 are located relatively far from each other, an intersection is less probable and thus more serious when it occurs. The message from the analysis and from fig. IV.PHQ.1.2. is that PHQ2 is perceived as an easier item, or approximately equal to PHQ8 by low scoring persons (sum score 0-6), while PHQ2 is perceived more difficult than PHQ8 for persons scoring high, (sum score 12-21).

Fig. IV.PHQ.1.2. Graphic (explorative) illustration of the intersection between PHQ2 and PHQ8



<u>Step 2 PHQ.</u> Analysis by a Rasch model

In a first analysis, a Rasch RSM,with common category thresholds, is applied.

Table IV.PHQ.2.1. Estimates from a Rasch RSM.

```
-----------------------------------------------------
|          MODEL|   INFIT  |EXACT MATCH|ESTIM|      |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|---------------+----------+-----------+-----+------|
|   -.14     .16| .66  -2.0| 56.4  43.2| 1.24| PHQ1 |
|   -.18     .16| .74  -1.5| 38.6  42.7| 1.46| PHQ2 |
|   -.47     .16|1.23   1.2| 47.4  46.9|  .72| PHQ3 |
|   -.26     .16|1.05    .4| 52.6  43.7|  .87| PHQ4 |
|    .47     .16|1.35   1.9| 36.8  43.5|  .53| PHQ5 |
|   -.16     .16| .78  -1.3| 52.6  42.8| 1.32| PHQ6 |
|   -.36     .16| .97   -.1| 52.6  48.8| 1.19| PHQ7 |
|    .62     .16|1.28   1.5| 47.4  46.2|  .62| PHQ8 |
|   1.57     .20|1.17    .8| 66.7  61.0|  .97| PHQ9 |
|---------------+----------+-----------+-----+------|
|    Mean    .16| Person reliability= 0.80
|    S.D.    .01|   Item reliability= 0.92
-----------------
```

The item locations are separated in three blocks, items (PHQ1 - 4,6,7), items (PHQ5,8) and item PHQ9, which correspond to the item locations (- 0.47, ..,-0.14), (0.47, 0.62) and 1.57. Six of the 9 items are gathered, close to each other, in the first block. There was no sign of any second dimension.

Table IV.PHQ.2.2. Table of standardized residual variance (in Eigenvalue units)

```
                                                 -- Empirical --
Total raw variance in observations     =        18.9 100.0%
  Raw variance explained by measures   =         9.9  52.4%
    Raw variance explained by persons  =         7.1  37.3%
    Raw Variance explained by items    =         2.9  15.1%
  Raw unexplained variance (total)     =         9.0  47.6%
  Unexplained variance in 1st contrast =         2.2  11.8%
```

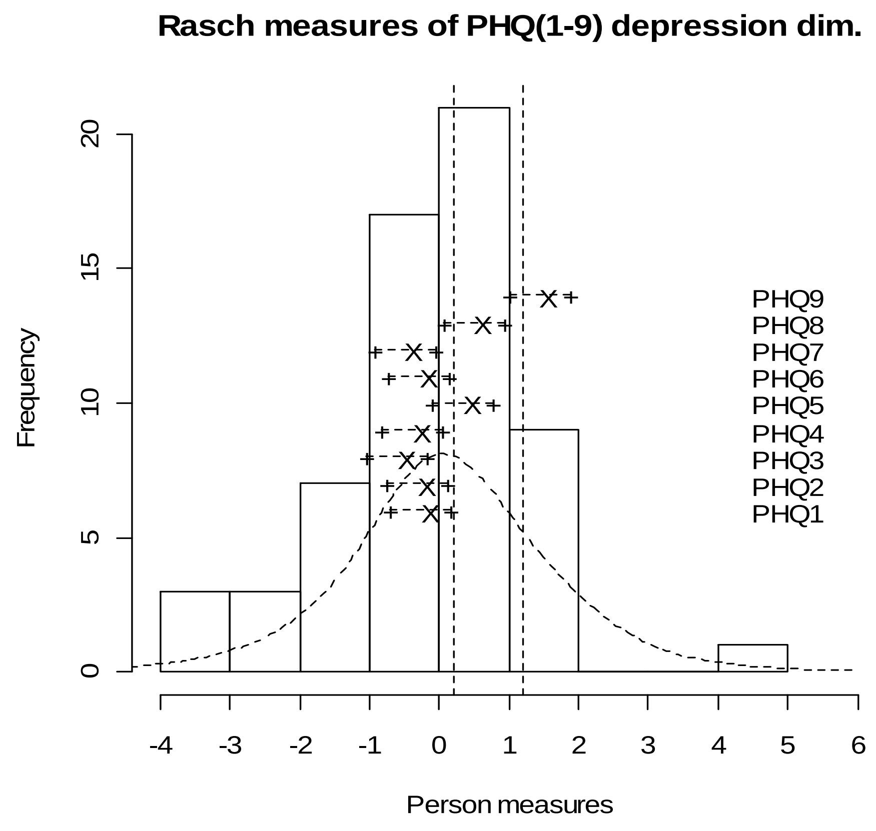
Just about 15% of the variation is explained by items, which is a consequence of the insufficient coverage, see fig. IV.PHQ.2.1.

Table IV.PHQ.2.3. Largest standardized residual correlations used to identify dependent items

```
---------------------------------
|CORREL-|           |           |
|  ATION|      ITEM |      ITEM |
|-------+-----------+-----------|
|   .30 |      PHQ2 |      PHQ6 |
|-------+-----------+-----------|
|  -.41 |      PHQ1 |      PHQ3 |
|  -.37 |      PHQ2 |      PHQ8 |
|  -.35 |      PHQ2 |      PHQ7 |
|  -.34 |      PHQ6 |      PHQ8 |
|  -.32 |      PHQ5 |      PHQ9 |
|  -.32 |      PHQ5 |      PHQ6 |
---------------------------------
```

A certain amount of residual correlation remains after fitting the Rasch model. The correlation between PHQ2 and PHQ6 is reflected by the close locations of these items, -0.14 and -0.16 resp..
There were no violations against ordering of the categories.

Fig. IV.PHQ.2.1. Dashed vertical lines between item blocks.



Discriminations, estimated after the model fit, show a rather wide range, Even if there is no strong indication against the Rasch model, this indicates that a model with item specific slopes might be a more appropriate approach.

<u>Step 3 PHQ</u>

A GRM with item specific slopes and a common set of category thresholds was applied.

Table IV.PHQ.3.1. Estimtes and fit statistics from the GRM with tem specific slopes.

```
   CATEGORY PARAMETER  :     0.774    -0.056    -0.718 thresholds
   S.E.                :     0.058     0.053     0.057
+------+---------+---------+---------+---------+
| ITEM |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+======+=========+=========+=========+=========+
| PHQ1 |   2.425 |   0.390 |   0.055 |   0.189 |
| PHQ2 |   1.266 |   0.593 |  -0.094 |   0.168 |
| PHQ3 |   0.918 |   0.183 |  -0.266 |   0.244 |
| PHQ4 |   1.209 |   0.216 |  -0.114 |   0.221 |
| PHQ5 |   0.704 |   0.209 |   0.551 |   0.284 |
| PHQ6 |   1.484 |   0.426 |   0.067 |   0.174 |
| PHQ7 |   1.245 |   0.269 |  -0.393 |   0.225 |
| PHQ8 |   0.891 |   0.173 |   0.728 |   0.273 |
| PHQ9 |   1.150 |   0.278 |   1.417 |   0.263 |
+------+---------+---------+---------+---------+
```

```
               ITEM FIT STATISTICS
-----------------------------------
| ITEM | CHI-SQUARE |  D.F. | PROB. |
------------------------------------
| PHQ1 |    7.32736 |    5. | 0.196 |
| PHQ2 |    9.17564 |    4. | 0.056 |
| PHQ3 |    0.14748 |    2. | 0.921 |
| PHQ4 |    2.70626 |    4. | 0.611 |
| PHQ5 |   10.83780 |    5. | 0.054 |
| PHQ6 |    6.21241 |    5. | 0.285 |
| PHQ7 |    0.17534 |    2. | 0.909 |
| PHQ8 |    3.33640 |    5. | 0.651 |
| PHQ9 |   12.20423 |    6. | 0.057 |
------------------------------------
|Total |   52.12292 |   38. | 0.063 |
-------------------------------------
```

The GRM reveals a wider range of item slopes than was estimated by the Rasch model. The tendency of blocking the items remains.
The fit statistics are acceptable. The total chi-square is too pessimistic due to a certain amount of local dependence.

Table IV.PHQ.3.2. Correlation of item residuals after fitting a GRM with item specific slopes..

	PHQ1	PHQ2	PHQ3	PHQ4	PHQ5	PHQ6	PHQ7	PHQ8	PHQ9
PHQ1	1.00	0.15	**-0.35**	-0.06	-0.14	0.09	-0.16	-0.08	-0.18
PHQ2	0.15	1.00	0.13	0.12	0.04	**0.41**	-0.06	-0.11	0.09
PHQ3	-0.35	0.13	1.00	0.01	0.11	-0.06	0.18	0.05	0.10
PHQ4	-0.06	0.12	0.01	1.00	0.12	0.07	-0.08	-0.27	-0.12
PHQ5	-0.14	0.04	0.11	0.12	1.00	-0.16	0.11	0.05	-0.28
PHQ6	0.09	0.41	-0.06	0.07	-0.16	1.00	0.11	-0.20	-0.12
PHQ7	-0.16	-0.06	0.18	-0.08	0.11	0.11	1.00	**0.35**	-0.17
PHQ8	-0.08	-0.11	0.05	-0.27	0.05	-0.20	0.35	1.00	-0.06
PHQ9	-0.18	0.09	0.10	-0.12	-0.28	-0.12	-0.17	-0.06	1.00

A few residual correlations remain.PHQ2 and PHQ6 share almost exactly the same location.

Fig. IV.PHQ.3.1. Item locations based on a Rasch model and a GRM with common slopes.

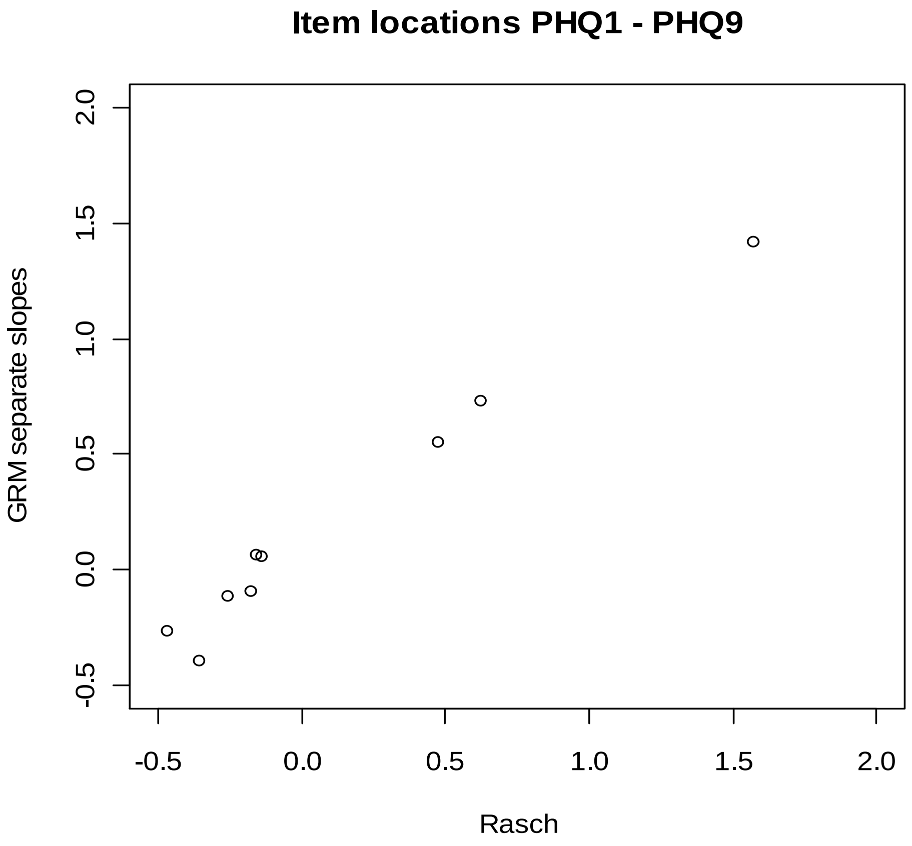
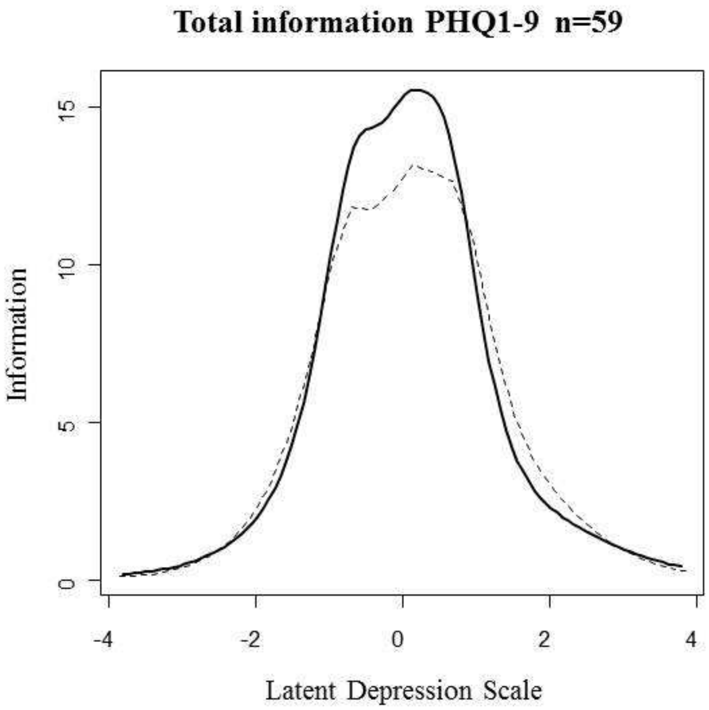


Fig. IV.PHQ.3.2. illustrates that the GRM approach did not break the blocking of items.

Fig. IV.PHQ.3.2. Relative (total) test information (TIF), based on the total item set PHQ1-9, from the unconstrained model (solid line) and TIF, as summed IIF:s, from a GRM with a common set of category thresholds (the unconstrained model)



The valuable information is concentrated in the range [-1.6, 1.6].
The GRM approach changes, but does not improve, the correlation structure.

Table IV.PHQ.3.3. Relative item information, GRM with common category thresholds(constrained) vs an unconstrained model.

Item	rel. info % GRM constr.	rel. info % GRM unconstr.
PHQ1	27.0	19.3
PHQ2	10.9	23.1
PHQ3	6.9	5.4
PHQ4	10.2	9.1
PHQ5	4.7	4.4
PHQ6	13.7	18.8
PHQ7	10.6	7.5
PHQ8	6.6	4.5
PHQ9	9.4	7.9

Even if the estimated relative item information 'moves' between items for the two models in table IV.PHQ.3.3., the structure remains, showing PHQ1, PHQ2 and PHQ6 with more information than the rest of the items.

The unconstrained GRM shows a significant better fit than does the GRM with a common set of category thresholds, This is, however , expected.

Conclusion about the PHQ questionnaire

- Insufficient coverage, too many items cover the same range. Reformulate items or introduce new items for a wider coverage. Items PHQ2 and PHQ6 seem to have very much of the information in common.
- Varying item slopes indicates that a Rasch model is too parsimonious. It might be used but is probably not the best choice. A GRM with item specific slopes but a common set of category thresholds might be more adequate even if a more parsimonious model could not be rejected.
- The violation against non-intersection, found between PHQ2 and PHQ8 should not be taken too seriously at this phase, but the phenomenon should be followed as more persons are studied.

Investigation of the MADRS questionnaire (MADRS)

The MADRS questionnaire has many categories for the 10 items, where just category 0, 2, 4 and 6 have a specified leading text. A tendency of placing the response in one of those is seen in the frequency table of answers from the 61 respondents, table IV.MADRS.0.1.

Table IV.MADRS.0.1.

Frequency Table. MADRS depression dimension
10 items scored by 61 subjects

Item	Score 0	Score 1	Score 2	Score 3	Score 4	Score 5	Score 6	Missing Data	Row Totals
MADRS1	24	12	12	2	7	3	1	0	61
MADRS2	26	8	15	6	2	0	0	4	61
MADRS3	15	7	15	17	6	1	0	0	61
MADRS4	26	3	10	7	13	0	2	0	61
MADRS5	40	4	9	3	5	0	0	0	61
MADRS6	12	6	12	9	21	0	1	0	61
MADRS7	24	6	18	3	9	1	0	0	61
MADRS8	29	6	15	7	3	0	1	0	61
MADRS9	17	3	19	13	8	0	0	1	61
MADRS10	38	6	10	3	4	0	0	0	61
All Items	251	61	135	70	78	5	5	5	

Due to non-responses for MADRS2 and MADRS9, there are 56 complete questionnaires.
As there are sparse data in higher categories, a reduced scale might be considered. It would slightly change the character of the questionnaires but might be necessary due to numerical problems in some of the analyses.
When looking at table IV.MADRS.0.1., a natural, radical reduction is grouping the categories (0,1), (2,3) and (4,5,6). Then we get the following reduced scale: (0,1,2,3,4,5,6) → (0,0,1,1,2,2,2). A more moderate reduction would be (0,1,2,3,4,5,6) → (0,1,2,3,4,4,4) or → (0,1,2,3,4,5,5).

Step 1 MADRS. The Mokken scale analysis

The analysis is based on 56 complete questionnaires.

Table IV.MADRS.1.1. Item pair scalabilities for MADRS1-10.

```
         MADRS1 MADRS2 MADRS3 MADRS4 MADRS5 MADRS6 MADRS7 MADRS8 MADRS9 MADRS10
MADRS1    1.000  0.746  0.704  0.203  0.452  0.406  0.353  0.545  0.558   0.392
MADRS2    0.746  1.000  0.533  0.233  0.430  0.318  0.412  0.455  0.481   0.377
MADRS3    0.704  0.533  1.000  0.222  0.350  0.494  0.271  0.348  0.498   0.325
MADRS4    0.203  0.233  0.222  1.000  0.202  0.289 -0.130 -0.030  0.124  -0.212
MADRS5    0.452  0.430  0.350  0.202  1.000  0.371  0.371  0.486  0.239   0.154
MADRS6    0.406  0.318  0.494  0.289  0.371  1.000  0.412  0.430  0.231   0.257
MADRS7    0.353  0.412  0.271 -0.130  0.371  0.412  1.000  0.529  0.254   0.153
MADRS8    0.545  0.455  0.348 -0.030  0.486  0.430  0.529  1.000  0.425   0.389
MADRS9    0.558  0.481  0.498  0.124  0.239  0.231  0.254  0.425  1.000   0.639
MADRS10   0.392  0.377  0.325 -0.212  0.154  0.257  0.153  0.389  0.639   1.000
```

Item set scalability= 0.339
It is immediately realised that MADRS4 is (or will be) a problematic item. It is probably non-contributing and might well be counterproductive due to some negative pairwise scalabilities. A preliminary dimensionality test (Mokken AISP) left MADRS4 outside the primary dimension.

Investigating the item scalabilities, see table IV.MADRS.1.2., points out more clearly the weakness of MADRS4 (scalability<.3).

Table IV.MADRS.1.2. Item scalabilities

MADRS1	MADRS2	MADRS3	**MADRS4**	MADRS5	MADRS6	MADRS7	MADRS8	MADRS9	MADRS10
0.469	0.442	0.419	**0.103**	0.338	0.358	0.281	0.388	0.371	0.257

The weakness of MADRS4 or weakness/strength of the other items might depend on effects from the sparse representation of data for some categories. Item scalabilities, based on the reduced scale, is presented in table IV.MADRS.1.3.

Table IV.MADRS.1.3. Reduced scale (0,0,1,1,2,2,2)
Item set scalability H= 0.328

MADRS1	MADRS2	MADRS3	**MADRS4**	MADRS5	MADRS6	MADRS7	MADRS8	MADRS9	MADRS10
0.450	0.414	0.453	**0.114**	0.316	0.316	0.244	0.408	0.371	0.239

As seen in table IV.MADRS.1.3., the reduced scale yields approximately the same structure of item scalabilities and the item set scalability is virtually the same. It seems as a scale reduction does not severely change the item scalability structure.

Deletion of MADRS4 might be considered.

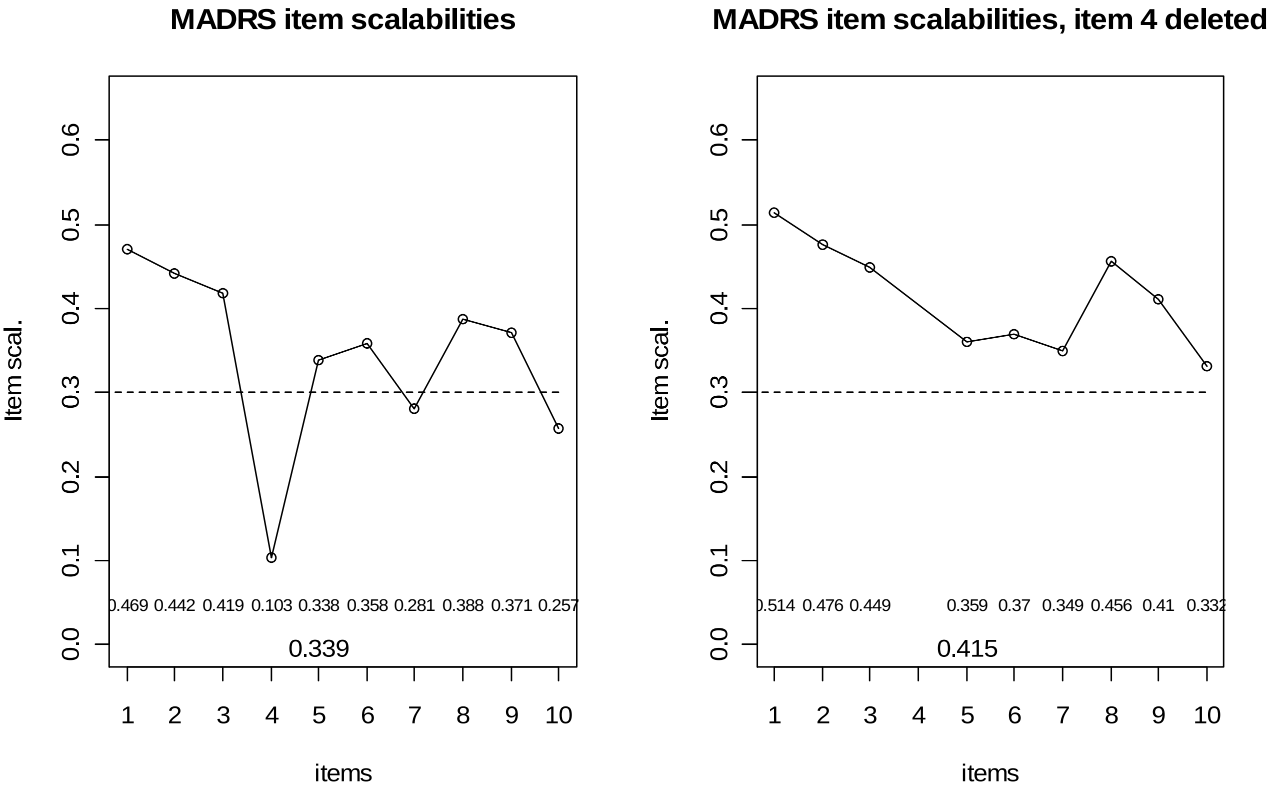
Table IV.MADRS.1.4. MADRS4 deleted H= 0.415 (the full scale)

MADRS1	MADRS2	MADRS3	MADRS5	MADRS6	MADRS7	MADRS8	MADRS9	MADRS10
0.514	0.476	0.449	0.359	0.370	0.349	0.456	0.410	0.332

Excluding MADRS4 seems to substantially improve the scalability structure, see table IV.MADRS.1.4.

Fig IV.MADRS.1.1. illustrates the marked difference between the full item set and a set without MADRS4.

Fig IV.MADRS.1.1.



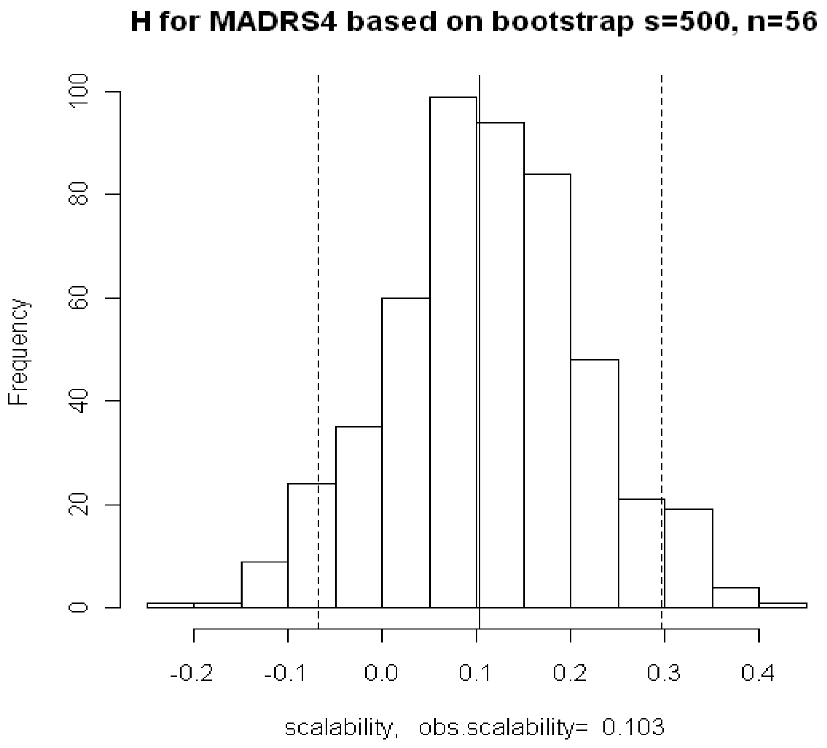
Should item MADRS4 be excluded from the questionnaire?

Exclusion of the disturbing item, MADRS4, ended in a significantly ($p<0.05$) improved overall scalability H of 0.415, C.I. [0.31, 0.51], implying a medium performing instrument. The test was done by bootstrapping the difference in scalability between the full item set and a set with MADRS4 excluded.

H=0.339 (with MADRS4 still in the item set) is just within the C.I. for H=0.415 (with MADRS4 excluded). This might look strange as a significant difference is statistically demonstrated, but is a consequence of the estimation procedure. C.I.:s for H and H(MADRS4 excl.) are based on different bootstrapped data sets (different empirical distributions), while the difference H - H(MADRS4 excl.) is directly estimated, based on a common set of bootstrapped samples, i.e. with an increased precision.

Looking on MADRS4 alone, see Fig. IV.MADRS.1.2., bootstrapping the distribution of its scalability indicates a possibility of a negligible or zero MADRS4 scalability. The C.I. [-0.069 0.297] says that MADRS4 might well contribute just noise.

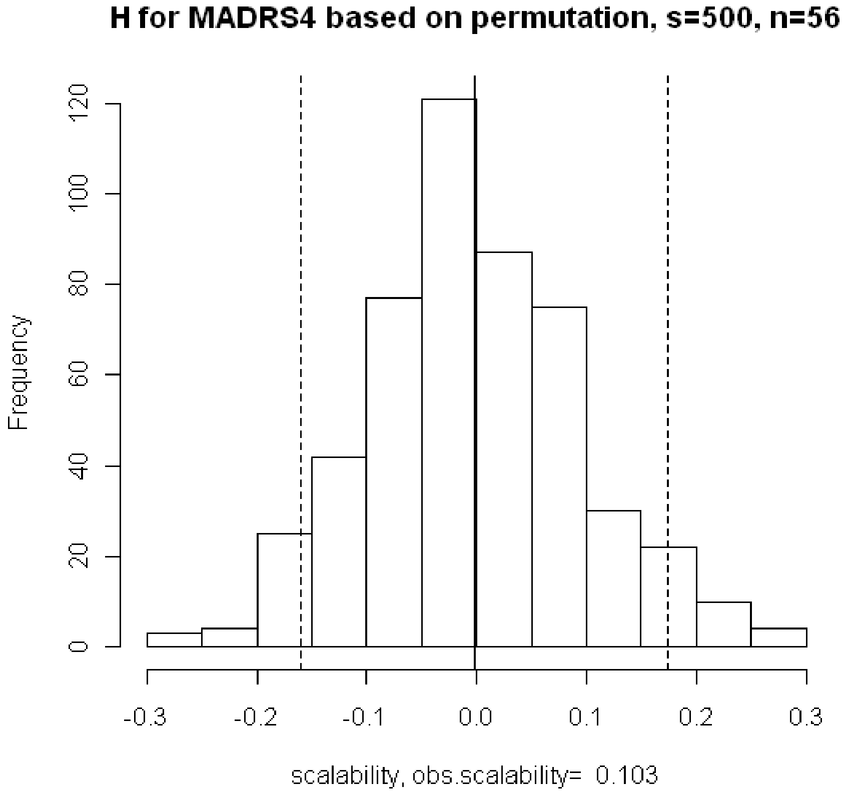
Fig. IV.MADRS.1.2.



```
  Min.  1st Qu.  Median  Mean   3rd Qu.    Max.      5%     95%
-0.218   0.047   0.112   0.111   0.178   0.436    -0.069   0.297
```

An alternative way of illustrating the weakness of MADRS4 is by permutation.
The response structure of the 56 answer profiles should be kept, but the answers on MADRS4 are randomly permutated to illustrate how an uncorrelated MADRS4 will appear.

Fig. IV.MADRS.1.3.



```
   Min.  1st Qu.  Median   Mean  3rd Qu.    Max.     5%      95%
 -0.259  -0.066  -0.008  -0.002   0.059   0.275  -0.160   0.174
```

The observed item scalability for MADRS4 is well within a C.I. for the MADRS4 made uncorrelated. This result together with the direct test suggests that the questionnaire would ‘feel better’ without MADRS4.

Monotonicity and non-intersection
Test of monotonicity revealed just a few non-significant violations.

Analysis of intersection, based on a reduced scale = (0,1,2,3,4,5,5)
Table IV.MADRS.1.5.

```
        ItemH  #vi  maxvi  sum  zmax #zsig
MADRS1   0.51   38   0.33 4.34  1.87     1
MADRS2   0.48   23   0.27 2.22  1.23     0
MADRS3   0.45   31   0.33 3.53  1.87     1
MADRS5   0.36   27   0.20 2.80  1.17     0
MADRS6   0.36   34   0.33 5.34  1.87     4
MADRS7   0.35   34   0.22 3.11  1.23     0
MADRS8   0.45   34   0.40 3.42  1.81     1
MADRS9   0.41   36   0.31 3.67  1.87     1
MADRS10  0.33   19   0.40 2.34  1.87     2
```

Analysis of non-intersection revealed a number of small violations and some more important (8 significant places). Some items are not perceived similarly when comparing ‘low’ and ‘high’ respondents. Noticeable intersections are observed between: 1 vs 5, 1 vs 6, 1 vs 7, 1 vs 9, 1 vs 10, 3 vs 5, 3 vs 6, 6 vs 8, 6 vs 9, and 8 vs 10.

However, only level 0, 2, 4 and 6 are labeled with a specific text, see Appendix C. If the scale is reduced by moving intermediate levels to the lower level with a specified text, we get
(0, 1, 2, 3, 4, 5, 6) →(0, 0, 1, 1, 2, 2, 2). Score 6 is moved to level 2 due to sparse data.

Table IV.MADRS.1.6.

```
        ItemH #vi  maxvi  sum  zmax #zsig
MADRS1   0.50   9   0.33 0.96  2.16     1
MADRS2   0.45   4   0.21 0.45  1.06     0
MADRS3   0.48   6   0.12 0.48  0.49     0
MADRS5   0.33   7   0.20 0.74  1.17     0
MADRS6   0.31  10   0.33 1.36  2.16     1
MADRS7   0.31   6   0.11 0.48  0.69     0
MADRS8   0.46   7   0.15 0.54  0.75     0
MADRS9   0.40   5   0.12 0.50  0.69     0
MADRS10  0.32   2   0.16 0.29  0.75     0
```

The number of violations is now considerably reduced. This might be explained by the new structure of questionnaire with text specified levels only. There are still problems with MADRS1 and MADRS6, which are differently ordered by low and high scoring subjects. The exclusion of MADRS4 and the rescaling is a considerable rearrangement of the original layout to get a reasonable questionnaire. Obviously, this exercise is too data driven’, but it illustrates the structure of the heterogeneity of the questionnaire and how it is perceived by the actual population.

However, the objective is to evaluate the questionnaire as it is. It is of interest to see how MADRS4 and the labelling of the answer levels are received by a parametric modelling approach.

Step 2 MADRS. Analysis by a Rasch model

Estimates based on a Rasch RSM.
Table IV.MADRS.2.1.

```
--------------------------------------------------------
|           MODEL|   INFIT  |EXACT MATCH|ESTIM|        |
|  MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM   |
|----------------+----------+-----------+-----+--------|
|    1.00     .11| .79  -1.3| 36.8  27.9|  .97| MADRS1 |
|    1.32     .12| .52  -3.0| 35.8  29.1| 1.17| MADRS2 |
|     .70     .11| .61  -2.6| 38.6  30.9| 1.03| MADRS3 |
|     .80     .11|2.03   4.8| 28.1  29.1|  .34| MADRS4 |
|    1.56     .13|1.08    .4| 35.1  39.5| 1.13| MADRS5 |
|     .37     .11|1.03    .2| 35.1  33.1| 1.07| MADRS6 |
|     .99     .11|1.15    .9| 26.3  27.9|  .87| MADRS7 |
|    1.21     .11| .84   -.9| 36.8  27.1| 1.12| MADRS8 |
|     .74     .11| .76  -1.5| 41.1  28.8| 1.17| MADRS9 |
|    1.56     .13|1.25   1.2| 40.4  39.5|  .98| MADRS10|
|----------------+----------+-----------+-----+--------|
|Mean 1.02       |Person reliability = 0.67
|S.D.  .37       |Item reliability   = 0.89
------------------
```

There were a lot of minor violations against the ordering of the categories. The main reasons are probably too many and sparsely observed categories together with too narrow categories, as perceived by the respondents.

Table IV.MADRS.2.2. Standardised residual variance (in Eigenvalue units)

```
Total raw variance in observations     =         16.9 100.0%
  Raw variance explained by measures   =          6.9  40.9%
    Raw variance explained by persons  =          4.6  27.2%
    Raw Variance explained by items    =          2.3  13.8%
  Raw unexplained variance (total)     =         10.0  59.1%
  Unexplained variance in 1st contrast =          1.9  11.1%
```

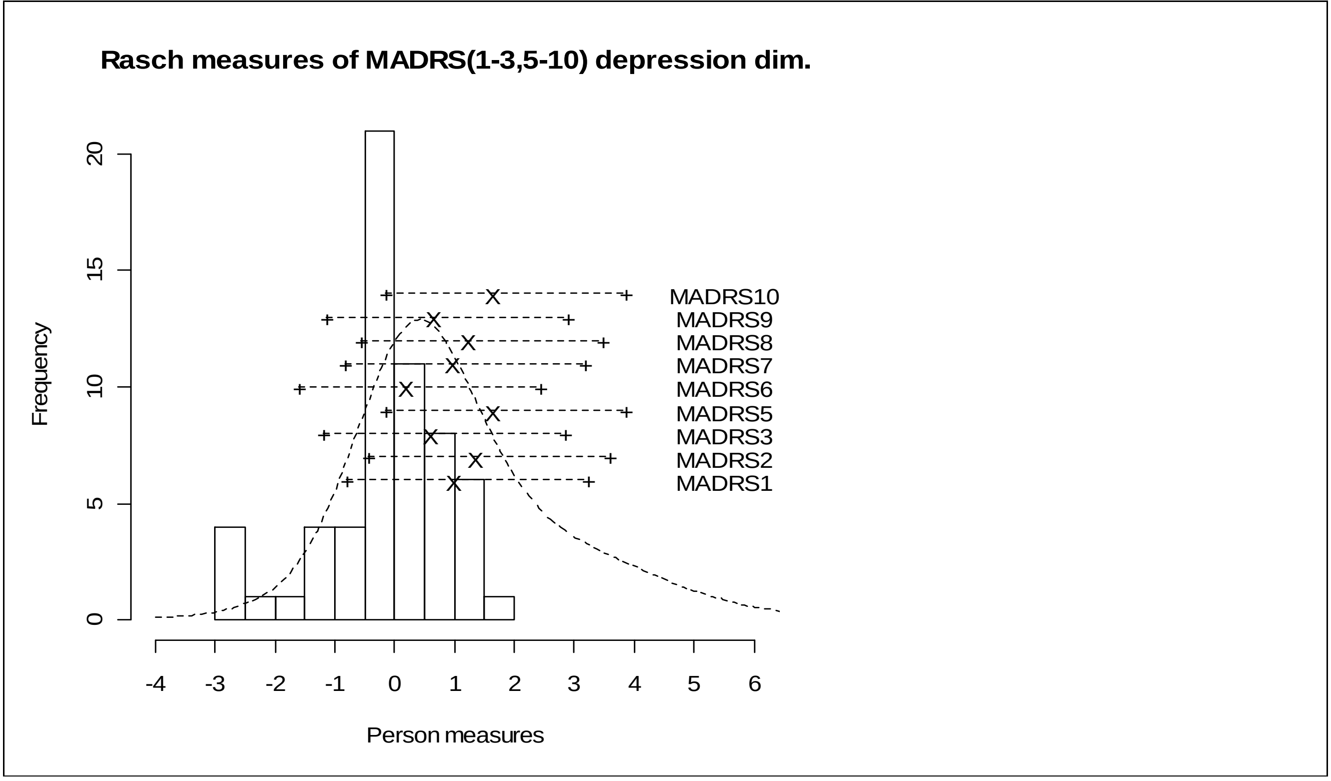
The person reliability, 0.67, is insufficient while the item reliability 0.89 indicates a relevant item set. However, MADRS4 does not fit into the model. There is much of the variation still unexplained although no second dimension could be recognised. Separate item thresholds or reduction of the scale did not change the structure shown in table IV.MADRS.2.1.and table IV.MADRS.2.2..
Moderate residual correlations were seen. Just one, cor(MADRS9, MADRS10)= 0.4 was > 0.3.

A Rasch RSM, with MADRS4 excluded, is shown in table IV.MADRS.2.3. and fig. IV.MADRS.2.1..

Table IV.MADRS.2.3. Common scale, MADRS4 deleted.

```
---------------------------------------------------------
|           MODEL|   INFIT   |EXACT MATCH|ESTIM|        |
|   MEASURE  S.E. |MNSQ  ZSTD| OBS%   EXP%|DISCR| ITEM   |
|----------------+----------+-----------+-----+--------|
|     1.21     .12| .89   -.6| 42.1  33.0|  .92| MADRS1 |
|     1.58     .13| .64  -2.1| 30.2  31.2| 1.13| MADRS2 |
|      .85     .12| .74  -1.6| 40.4  35.2|  .96| MADRS3 |
|   DELETED       |          |           |     | MADRS4 |
|     1.86     .14|1.26   1.2| 40.4  42.7| 1.02| MADRS5 |
|      .44     .12|1.30   1.6| 28.1  33.4|  .87| MADRS6 |
|     1.19     .12|1.25   1.4| 22.8  31.0|  .86| MADRS7 |
|     1.45     .12| .88   -.6| 40.4  32.9| 1.12| MADRS8 |
|      .90     .12| .89   -.6| 39.3  34.9| 1.08| MADRS9 |
|     1.86     .14|1.32   1.5| 40.4  42.7|  .96| MADRS10|
|----------------+----------+-----------+-----+--------|
|MEAN 1.26        | Person reliability = 0.75
|S.D.  .45        | Item reliability   = 0.91
-------------------
```

Fig. IV.MADRS.2.1.



The model reveals an insufficient coverage and a peaked test information.
The item discriminations, seen in table IV.MADRS.2.1. and IV.MADRS.2.3., are estimated after fitting the Rasch model and look fairly homogeneous and close to unity (besides MADRS4).

It should be noted that, in terms of person and item reliabilities, no essential information is lost when excluding MADRS4.
An extended model might give a different message.

Step 3 MADRS

An extended model, GRM with item specific slopes and a common set of category thresholds is fitted. A reduced scale (0,1,2,3,3,3,3) is used. Application of the model is shown in fig IV.MADRS.3.1.

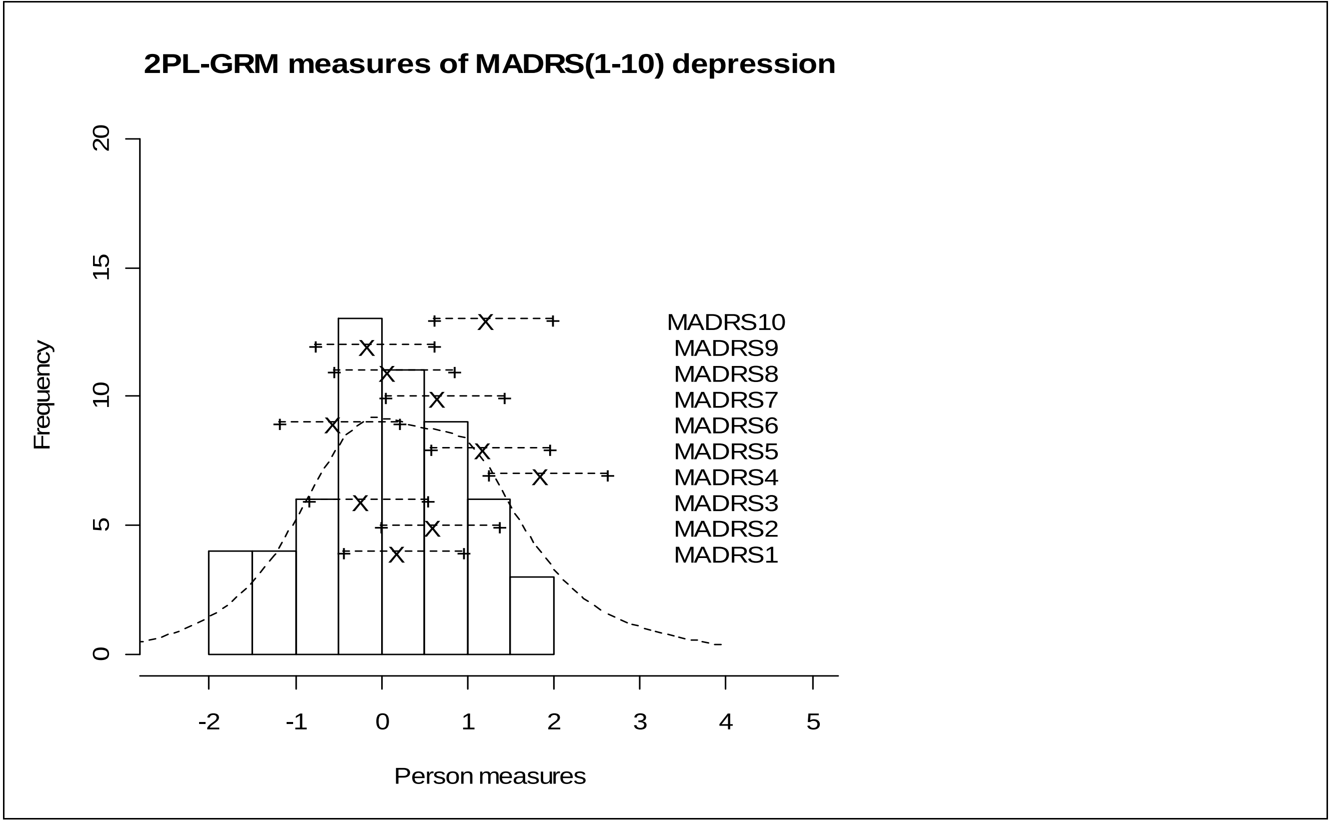
Table IV.MADRS.3.1. GRM with item specific slopes and a common set of thresholds.

```
CATEGORY PARAMETER  :     0.601     0.182    -0.783
              S.E.  :     0.062     0.060     0.071
+--------+---------+---------+---------+---------+
| ITEM   |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+========+=========+=========+=========+=========+
| MADRS1 |   1.927 |   0.719 |   0.160 |   0.198 |
| MADRS2 |   1.425 |   0.481 |   0.588 |   0.222 |
| MADRS3 |   1.175 |   0.239 |  -0.252 |   0.284 |
| MADRS4 |   0.273 |   0.047 |   1.832 |   1.372 |
| MADRS5 |   0.712 |   0.209 |   1.161 |   0.361 |
| MADRS6 |   0.659 |   0.180 |  -0.582 |   0.314 |
| MADRS7 |   0.783 |   0.134 |   0.641 |   0.439 |
| MADRS8 |   1.065 |   0.155 |   0.052 |   0.409 |
| MADRS9 |   0.898 |   0.259 |  -0.178 |   0.247 |
| MADRS10|   0.683 |   0.233 |   1.197 |   0.309 |
+------+---------+---------+---------+---------+
```

```
               ITEM FIT STATISTICS
 --------------------------------------
 | ITEM | CHI-SQUARE |  D.F. | PROB.   |
 --------------------------------------
 | MADRS1 |   11.66970 |    5. | 0.039 |
 | MADRS2 |    3.43424 |    6. | 0.755 |
 | MADRS3 |    5.57487 |    3. | 0.132 |
 | MADRS4 |    5.34485 |    3. | 0.146 |
 | MADRS5 |    3.02175 |    5. | 0.699 |
 | MADRS6 |    0.76800 |    1. | 0.385 |
 | MADRS7 |    8.62382 |    4. | 0.070 |
 | MADRS8 |    6.88413 |    4. | 0.141 |
 | MADRS9 |    6.75456 |    3. | 0.079 |
 | MADRS10|    4.43488 |    5. | 0.490 |
 --------------------------------------
 |  Total |   56.51078 |   39. | 0.034 |
 --------------------------------------
```

In table IV.MADRS.3.1, MADRS4 shows a strange estimate with a very high location and unreasonable S.E, which indicates a ‘random behaviour’. The slope is unreasonably low, which is in agreement with the findings in Step1 and Step2. The analysis reveals a rather high slope for MADRS1, which contradicts the Rasch model.

Fig. IV.MADRS.3.1. GRM with item specific slopes and a common set of thresholds.



Although the model cannot be rejected, according to the item fit statistics, it is hardly better than the Rasch model. However, the item slopes indicate that equal item discrimination is probably too parsimonious to be a reasonable approach. The relative item information, see table IV.MADRS.3.2., confirms this indication. However, general conclusions are hazardous, as this analysis is carried out on a greatly reduced scale.

The approximate item relative information, as estimated from the applied model, appears relevant when compared to estimates based on an unconstrained model with item specific slopes and item specific category thresholds.

Table IV.MADRS.3.2. Relative item information, GRM with common category thresholds (constrained model) vs an unconstrained model

Item	rel.info % GRM constr.	rel.info % GRM unconstr.
MADRS1	25.2	27.2
MADRS2	16.6	19.7
MADRS3	12.7	12.1
MADRS4	1.3	2.1
MADRS5	6.1	4.9
MADRS6	5.6	4.4
MADRS7	7.1	4.5
MADRS8	11.1	8
MADRS9	8.7	10.9
MADRS10	5.7	6.2

Irrespective of any model, MADRS4 is identified as an item with negligible contribution.

Conclusion about MADRS questionnaire

- A questionnaire with 10 items and 7 categories might well be too much for an evaluation with just 61 patients. Specified texts for only the half of the answer levels – does it really work? However, at least some major messages are worthwhile to consider.
- It would be too risky to recommend exclusion of MADR4, based on just 61 patients. I suggest that MADRS4 is kept in the questionnaire, but complementary analyses, with MADRS4 excluded, should be continuously followed in further evaluations. Prepare a parallel measure without MADS4 and follow what happens.
- Even if a Rasch model does not seem to fit very well, an extended model is hardly an improvement.
- The analysis, so far, indicates that item specific discriminations might be considered.

Summary conclusion about the three assessment instruments

A basic recommendation for all the three questionnaires: Widen the range of the items to improve the coverage.
In general, there are good item reliabilities, but further items are needed to improve the person reliability.
MADRS seems to be the weakest questionnaire in spite of, or due to, too many category levels.
Even if the authors' intention is to create a questionnaire with equally important items, it appears to be difficult to accomplish. The analyses indicate item specific weights. Although 'data driven' it should be interpreted as a hint to reformulate the items if a balanced questionnaire is desirable.
The MADRS and the ASD questionnaires should probably be improved by replacement of their weak items.

Study V

Adler Mats, Brodin Ulf. **An IRT validation of the Affective Self Rating Scale.**
Nordic Journal of Psychiatry, 2011, Vol 65 No 6, p 396-402.

The Affective Self Rating Scale (AS-18) is intended for getting information about bipolar outpatients. It includes subscales for the rating of depressive and manic type symptoms. It has previously been validated using methods from Classical Test Theory.
The aim of this study was to evaluate the psychometric properties of the AS-18 when used at an outpatient clinic for patients with bipolar disorder at routine visits, and to analyse the potential for improvement of the scale.
231 patients with mainly bipolar disorder doing ratings on routine visits at an affective disorder outpatient clinic were included.

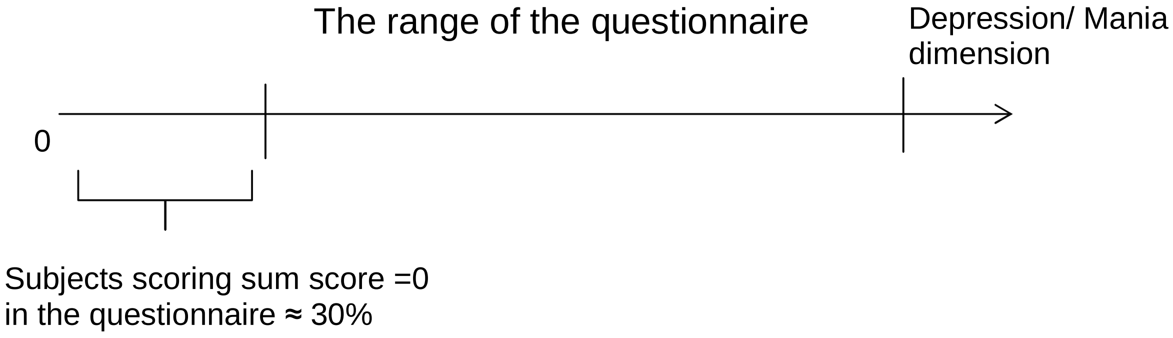
A large part of the patients scored zero at all items, 26% of the Depression and 33% of the Mania questionnaires. The objective was to evaluate the questionnaire. This means that the items should be relevant for the target population. As the questionnaire is aimed for a defined population, this also means that patients with extreme scores should be observed just occasionally. An extreme score is achieved when all items are responded in one of the extreme categories, i.e. all responses are in the first category or all responses are in the last category. Even if many IRT methods do not include extreme cases in the estimation process, a large amount of extreme cases might deteriorate the evaluation and hide important characteristics. However, when a questionnaire is taken in regular use later on, the extreme cases will not cause any problems except when there is a large part of respondents for which the questionnaire seems not to be suitable. In this case, it is the lower extreme which is of interest. Patients scoring zero at all items cannot be said to be ‘free of symptoms’ in any definite sense, but rather on a level outside the range of the questionnaire.

The first index in tables and figures, V, indicates study V.
The second index in tables and figures represent the intended dimension,
Depression=Dep, Mania =Man.
The third index in tables and figures represent the step within dimension. A ‘zero’ represent a general investigation before going into dimension or Step details.
The fourth index represent table or figure no. within the step.

In a preliminary Mokken dimensionality analysis, with all 18 items included (mania + depression), all items besides Dep2 were placed within the same scale. The analysis was based on n= 193 complete response profiles. This result indicates that the two dimensions may not be perceived by the responders, but rather defines a clinical separation. As this is a decision (or a problem) outside the scope of this thesis, the evaluation is continued by analysing the two phenomenon separately, as defined by the clinicians. This also means that extreme cases within the Depression or the Mania dimension are excluded in the respective evaluations.

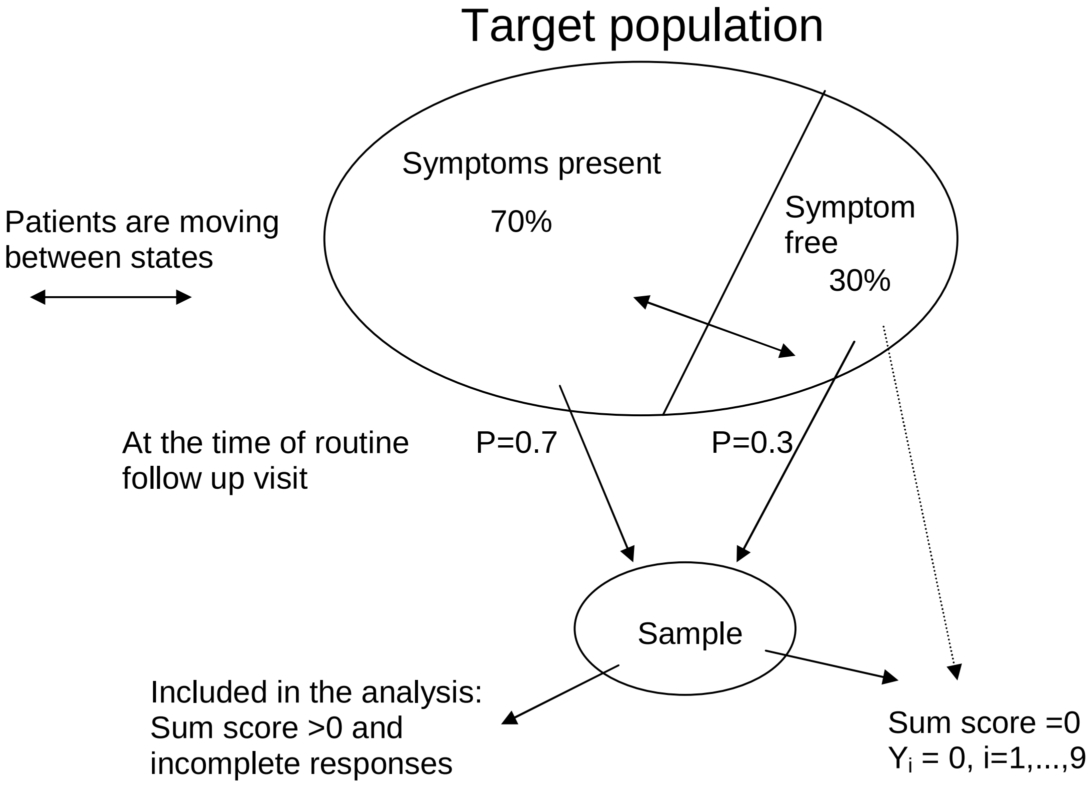
The questionnaire has a limit of detection, below which symptoms are not recognised.

Fig. V.Dep/Man.0.1.


The range of the questionnaire
Depression/ Mania dimension
0
Subjects scoring sum score =0 in the questionnaire ≈ 30%

The range of the questionnaire in fig. V.Dep/Man.0.1 also has an upper limit (all items scored in the highest category), which did not cause any problem as none of the patients in the actual sample reached the maximum sum score.

Fig. V.Dep/Man.0.2. Outline of the group ‘free of symptoms’ in the population and a sample.


Target population
Symptoms present
70%
Symptom free
30%
Patients are moving between states
At the time of routine follow up visit
P=0.7
P=0.3
Sample
Included in the analysis: Sum score >0 and incomplete responses
Sum score =0
Y_i = 0, i=1,...,9

A relatively large amount of patients scored zero at all depression or mania items, Y_i=0 for item i=1,…,9. Fig.V.Dep/Man.0.2. illustrates how the ‘symptom free’ group, about 30%, is treated. Patients are moving in and back from the target population. The ‘symptom free’ group will constitute a stochastic sub population. This means that a particular patient will be suitable/not suitable for the questionnaire at different points in time.

Analysis of the dimensions.

The intention of the questionnaire is to identify two dimensions, Depression and Mania. The items are mixed in the questionnaire. N= 109 persons had a sum score >0 over all the 18 items. The AISP procedure was performed for the 18 items. The result is presented in table Dep/Mani.0.1.

Table Dep/Mani.0.1. Lower bound H_{ij}= 0.3

Dimensionality analysis of the total questionnaire						
Item	scale	Item	scale		no scale	
Dep4	1	Man1	2		Dep2	0
Dep5	1	Man3	2		Man18	0
Dep10	1	Man16	2			
Dep11	1					
Dep12	1					
Dep13	1					
Dep14	1					
Dep17	1					
Man6	1					
Man7	1					
Man8	1					
Man9	1					
Man15	1					

The main part were placed in the first scale, while leaving two items outside any scale. If the lower bound was set to 0.2 all items were placed in a common scale, leaving Dep2 and Man1 outside. The procedure just illustrates that the two dimensions turn out to be close to each other.

Investigation of the ‘Depression’ questionnaire (Dep).

The 228 questionnaires have to be regrouped according to the frequency of non-responses and the ‘all-zero’ scoring patients. Unfortunately, this result in substantially different groups depending on which analysis is carried out. Rasch models tolerate non responses, while most of the other model do not. The subgroups are specified in table V.Dep.0.1.

Table V.Dep.0.1. Material for evaluation of the Depression questionnaire

Total sample	N= 228
Complete cases	n= 205
Cases with ∑Y(i)= 0, i=1,…,9	n= 54
Cases with ∑Y(i)> 0, i=1,…,9	n= 151
Patients included in the evaluation	n= 228-54= 174

Fig. V.Dep.0.1. Distribution of the sum score. Depression dimension. N= 205

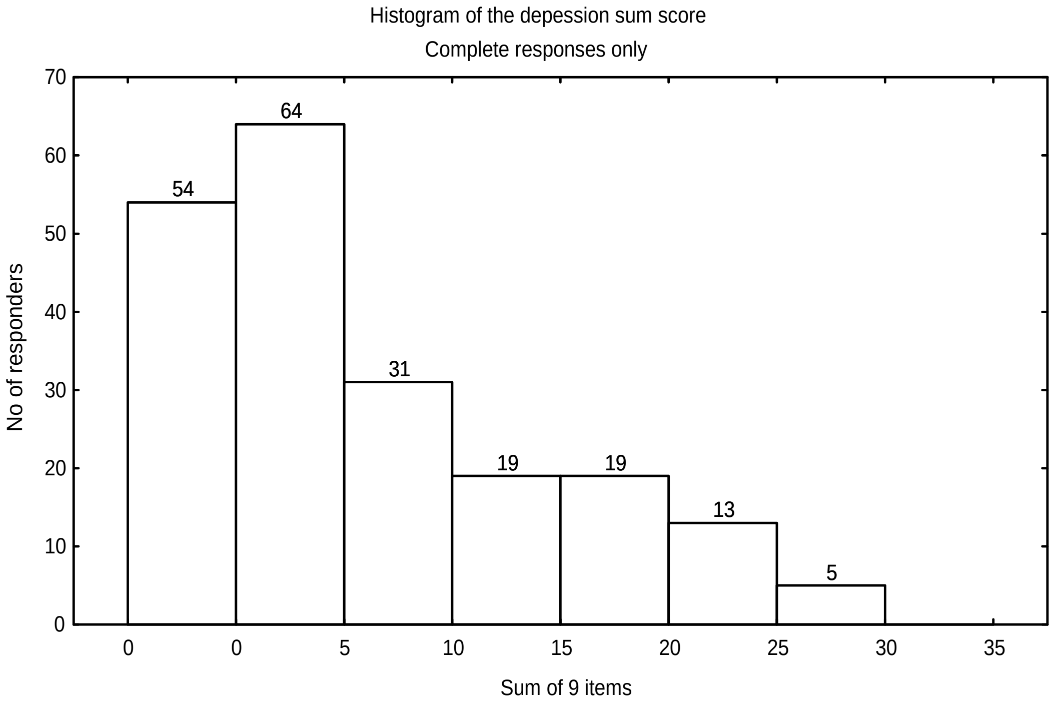


Fig. V.Dep.0.1. reveals a sum score distribution skewed to the lower end.

The sample of N= 174 subjects, see table V.Dep.0.1., consists of:
151 subjects with complete answer profiles and ∑Y(i)> 0.
18 subjects with ∑Y(i)> 0, but with one or more items with a ‘non response’.
5 subjects with ∑Y(i)= 0, but with one or more items with a ‘non response’. These 5 subjects are included as we are not able to say that they have responded zero to all items.
This group is analysed in Step2.

Table V.Dep.0.2. Response frequencies in categories. N= 174

Category	0	1	2	3	4	Non resp.
Dep2	82	31	33	20	6	2
Dep4	74	32	31	20	14	3
Dep5	95	33	30	7	6	3
Dep10	57	43	36	21	15	2
Dep11	72	35	43	15	7	2
Dep12	38	45	51	24	9	7
Dep13	79	28	28	22	9	8
Dep14	78	32	31	14	5	14
Dep17	136	11	8	4	3	12

As is seen from Table V.Dep.0.2., the proportion of non responses is <4% although their number, 54, is quit large.

<u>Step 1 Dep.</u> The Mokken scale analysis.

The 151 complete cases with ∑Y(i)> 0, i=1,…,9 are used in Step1.
All items showed scalabilities well above 0.3 and item pair scalabilities in the range [0.34, 0.67].
The item set scalability is = 0.562, which indicates a strong scale (table 3 in the article).
The analysis of monotonicity showed only a few violations. Dep2 had noticeable violations for low scoring subjects (table V.Dep.1.1.). The same can be said about Dep12.

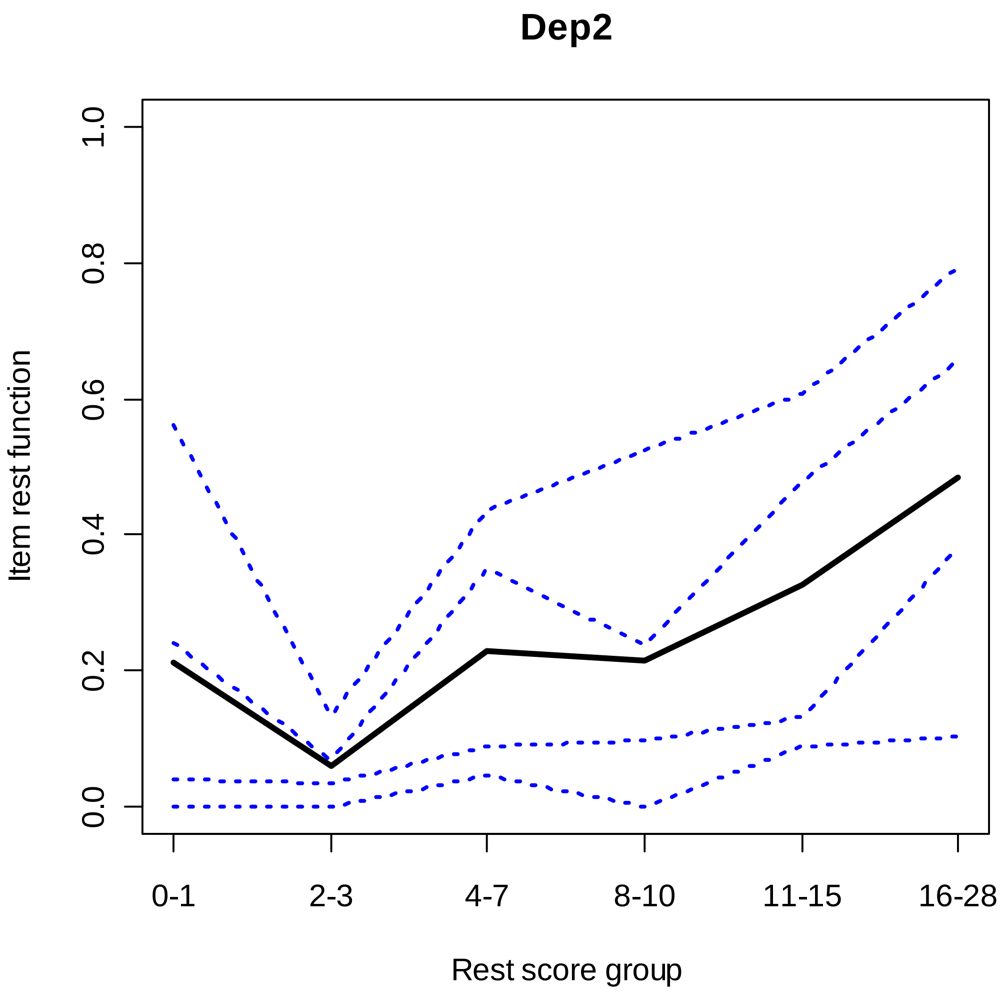
Analysis of monotonicity.

Table V.Dep.1.1. Analysis of monotonicity. Minimum size=20. N= 151.

```
       ItemH  #vi  maxvi  sum   zmax #zsig
Dep2    0.34    6   0.43  0.91  3.06     1
Dep4    0.63    5   0.20  0.62  1.23     0
Dep5    0.53    1   0.07  0.07  0.23     0
Dep10   0.67    3   0.25  0.39  1.42     0
Dep11   0.64    0   0.00  0.00  0.00     0
Dep12   0.63    5   0.29  0.54  1.78     1
Dep13   0.57    1   0.12  0.12  0.56     0
Dep14   0.53    0   0.00  0.00  0.00     0
Dep17   0.45    4   0.14  0.33  1.06     0
```

The same result is achieved with the minimum group size set to 30, however not significant.

Fig. V.Dep.1.1. The probability of scoring on Dep2 vs the level of depression as estimated by the rest score groups.



As can be seen from fig. V.Dep.1.1. the probability of scoring >1 and >2 on Dep 2 is decreasing for increasing rest score levels for low scoring patients, which is a violation against monotonicity. There is also some disordering for the middle (8-10) rest scoring group. This should not be taken too seriously but indicates difficulties of Dep2 to agree with a parametric model later on. Dep12 and Dep17 also show some irregularities. In fig. V.Dep.1.1. the level of depression is represented by the subject's rest score, $R_{(\text{-Dep2})}$.

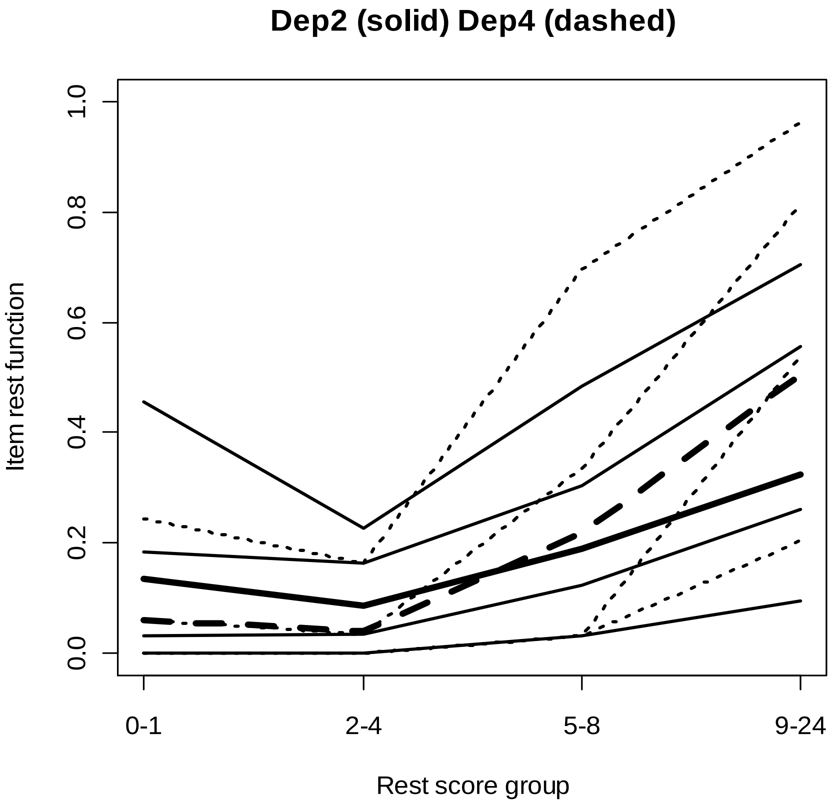
Analysis of non-intersection

There were a few but substantial violations against non-intersection. Even when increasing the violation threshold to 0.1, there were still a set of significant violations. A change of the minimum group size did not change the structure of the analysis.

Table V.Dep.1.2. Analysis of intersection, Violation threshold= 0.1. Min.groupsize=30.

	ItemH	#vi	maxvi	sum	zmax	#zsig
Dep2	**0.34**	**16**	**0.32**	**2.60**	**2.85**	**5**
Dep4	0.63	7	0.21	0.92	1.34	0
Dep5	0.53	2	0.15	0.26	1.60	0
Dep10	0.67	4	0.19	0.62	2.16	1
Dep11	0.64	4	0.22	0.56	2.22	1
Dep12	0.63	1	0.32	0.32	2.85	1
Dep13	0.57	3	0.23	0.51	2.42	2
Dep14	0.53	3	0.15	0.39	1.46	0
Dep17	0.45	0	0.00	0.00	0.00	0

Fig. V.Dep.1.2. The probability of scoring on Dep2 and Dep4 vs the level of depression as estimated by the rest score groups $R_{(-\text{Dep2},-\text{Dep4})}$.



The mean estimates for Dep2 crosses the mean estimates for Dep4. The thick solid and dashed lines are expected responses, rescaled to fit the range [0,1].

Dep2 and Dep4 are close to each other regarding the item location (-0.05 and -0.34 from a Rasch perspective). Thus, Dep2 is somewhat ‘easier’ than Dep4.

Consider the low scoring group with rest score 0-1.
P(Y≥c) means probability(the answer(Y) is in category c or higher)

```
             P(Y≥1)   P(Y≥2)   P(Y≥3)   P(≥4)
Dep2:        0.4545   0.1818   0.0303   0.0000
Dep4:        0.2424   0.0606   0.0000   0.0000
```

From the table above we realise that the probability of scoring high is larger for Dep2 than for Dep4. Let us move to the high scoring group with rest scores 9-24.

```
             P(Y≥1)   P(Y≥2)   P(Y≥3)   P(≥4)
Dep2:        0.7037   0.5556   0.2593   0.0926
Dep4:        0.9630   0.8148   0.5370   0.2037
```

Now we realise that the probability of scoring high is larger for Dep4 than for Dep2. This is an indication, and statistically significant, that the items are perceived differently when we compare low and high scoring groups. This means a ‘violation against IIO’. An IIO is desired for a good questionnaire.
The same structure is observed for Dep2 vs Dep10, Dep 11, Dep 13 and Dep 14.

This result says that the order of items, in terms of ‘difficulty’ is not homogeneous over the depression scale. This indicates further difficulties for Dep2 to agree with a parametric model where the order of items is expected to be invariant over the actual range of the dimension. A straightforward interpretation of a sum score is also deteriorated.
The result from Step 1 says that the sum scores are reasonable for ranking the subjects on the depression scale. The result does not exclude a parametric model. However, special attention should be paid to Dep2

A few item pair scalabilities are >0.7. Their interpretation will be revealed in Step 2 and Step 3.

<u>Step 2 Dep.</u> Analysis by a Rasch model

A Rasch RSM is fitted to the data, N=174.

Table V.Dep.2.1. Application of the Rasch RSM with 9 items

```
--------------------------------------------------------
|             MODEL|   INFIT  |EXACT MATCH|ESTIM|      |
|    MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|------------------+----------+-----------+-----+------|
|      -.05     .10|1.95   6.4| 44.6  54.3|  .13| Dep2 |
|      -.34     .10| .90   -.8| 50.9  50.7| 1.17| Dep4 |
|       .40     .11|1.04    .3| 56.9  58.5|  .99| Dep5 |
|      -.59     .09| .62  -3.9| 53.3  46.8| 1.33| Dep10|
|      -.19     .10| .61  -3.9| 62.3  52.4| 1.38| Dep11|
|      -.81     .09| .66  -3.5| 56.4  46.5| 1.14| Dep12|
|      -.17     .10|1.00    .0| 55.6  53.0| 1.11| Dep13|
|       .11     .11|1.12   1.0| 53.8  55.4|  .91| Dep14|
|      1.64     .15|1.93   4.3| 76.1  76.0|  .86| Dep17|
|------------------+----------+-----------+-----+------|
| MEAN   .00       |   Person reliability = 0.78
| S.D.   .67       |     Item reliability = 0.97
--------------------
```

It is clearly indicated from table V.Dep.2.1. that Dep2 contribute more noise than information.
A large infit statistic and a low estimated discrimination indicate exclusion of Dep2 from the Rasch model. Exclusion of an outlying patient, BP269 with an outfit=6.7 and a response vector (0,0,0,0,0,0,0,0,1), did not change the result.
No second dimension was indicated, and no disturbing positive residual correlations were found, see table V.Dep.2.2.

Table V.Dep.2.2. Largest standardized residual correlations used to identify dependent items

```
---------------------------
|CORREL-|        |        |
|  ATION| ITEM   | ITEM   |
|-------+--------+--------|
|   .29 | Dep4   | Dep10  |
|-------+--------+--------|
|  -.38 | Dep2   | Dep4   |
|  -.37 | Dep2   | Dep10  |
|  -.30 | Dep10  | Dep14  |
---------------------------
```

Exclusion of Dep2 might change both the estimated locations and the structure of residual correlations.

Excluding Dep2 deteriorates the fit for Dep17, but leaves the person and item reliability unchanged. No loss of information is seen by exclusion of Dep2.

Table V.Dep.2.3.

```
-----------------------------------------------------
|          MODEL|   INFIT   |EXACT MATCH|ESTIM|      |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM |
|---------------+----------+-----------+-----+------|
|               |          |           |     | Dep2 |
|   -.42     .11| .92   -.6| 51.9  52.5| 1.15| Dep4 |
|    .46     .12|1.32   2.3| 55.8  58.8|  .75| Dep5 |
|   -.70     .10| .63  -3.8| 57.7  48.9| 1.34| Dep10|
|   -.24     .11| .68  -3.1| 62.8  52.9| 1.34| Dep11|
|   -.94     .10| .81  -1.7| 52.6  48.6| 1.00| Dep12|
|   -.20     .11|1.09    .8| 58.2  54.5| 1.02| Dep13|
|    .14     .11|1.25   1.9| 52.7  57.4|  .77| Dep14|
|   1.90     .16|2.14   5.0| 74.2  76.1|  .71| Dep17|
|---------------+----------+-----------+-----+------|
|Mean 0.00        |Person rel.= 0.79
|S.D. 0.83        |  Item rel.= 0.97
-------------------
```

Analysis of residual variance (in Eigenvalue units)

```
Total raw variance in observations     =         22.5 100.0%
  Raw variance explained by measures   =         14.5  64.5%
    Raw variance explained by persons  =          8.3  36.7%
    Raw Variance explained by items    =          6.2  27.7%
  Raw unexplained variance (total)     =          8.0  35.5%
  Unexplained variance in 1st contrast =          1.9   8.3%
```

Dep17 does not agree with the Rasch RSM, MNSQ= 2.14.
MNSQ>2 means “Off-variable noise is greater than useful information. Degrades measurement” [Linacre J., 2008]
No second dimension is indicated, unexplained variance in 1st contrast < 2.0

Table V.Dep.2.4. Largest standardized residual correlations used to identify dependent items. Dep2 is excluded.

```
--------------------------
|CORREL-|        |       |
|  ATION| ITEM   |ITEM   |
|-------+--------+-------|
|  -.41 | Dep10  |Dep14  |
|  -.37 | Dep5   |Dep17  |
|  -.33 | Dep11  |Dep14  |
|  -.33 | Dep5   |Dep13  |
--------------------------
```

The bad fit of Dep 2 and Dep 17 to the Rasch RSM, and the low MNSQ:s for some items indicate that a model with equal discrimination might not be sufficient.
This will be further evaluated in Step 3.

Step 3 Dep

Fitting a GRM with common slopes.

Table V.Dep.3.1. A GRM with common slopes.

```
  CATEGORY PARAMETER  :     1.239     0.523    -0.441    -1.321
  S.E.                :     0.035     0.037     0.049     0.076
+--------+--------+---------+---------+---------+
| ITEM   | SLOPE  |   S.E.  |LOCATION |   S.E.  |
+========+========+=========+=========+=========+
| Dep2  |  1.183 |   0.035 |   1.163 |   0.127 |
| Dep4  |  1.183 |   0.035 |   0.833 |   0.123 |
| Dep5  |  1.183 |   0.035 |   1.414 |   0.129 |
| Dep10 |  1.183 |   0.035 |   0.844 |   0.135 |
| Dep11 |  1.183 |   0.035 |   0.829 |   0.126 |
| Dep12 |  1.183 |   0.035 |   0.589 |   0.140 |
| Dep13 |  1.183 |   0.035 |   0.991 |   0.121 |
| Dep14 |  1.183 |   0.035 |   1.154 |   0.127 |
| Dep17 |  1.183 |   0.035 |   2.479 |   0.154 |
+-------+--------+---------+---------+---------+
```

```
               ITEM FIT STATISTICS
---------------------------------------
| ITEM   | CHI-SQUARE |  D.F. | PROB. |
---------------------------------------
| Dep2  |   22.50713 |   13. | 0.048 |
| Dep4  |   13.94924 |   13. | 0.377 |
| Dep5  |   15.03587 |   11. | 0.180 |
| Dep10 |   19.56718 |   13. | 0.106 |
| Dep11 |   32.22437 |   13. | 0.002 |
| Dep12 |   28.58355 |   12. | 0.005 |
| Dep13 |   11.78149 |   14. | 0.625 |
| Dep14 |   16.48608 |   13. | 0.223 |
| Dep17 |   15.37478 |    9. | 0.081 |
---------------------------------------
| Total  |  175.50970 |  111. | 0.000 |
---------------------------------------
```

With the constraint of common slopes, Dep2 does not appear as bad as was earlier indicated. However, the constraint may hide special characteristics of Dep2.

Table V.Dep.3.2. A GRM with common slopes and Dep2 excluded.

```
   CATEGORY PARAMETER  :     1.160     0.471    -0.419    -1.212
   S.E.                :     0.033     0.035     0.046     0.070
+--------+--------+---------+---------+---------+
|  ITEM  | SLOPE  |   S.E.  |LOCATION |   S.E.  |
+========+========+=========+=========+=========+
| Dep4   |  1.393 |   0.044 |   0.790 |   0.114 |
| Dep5   |  1.393 |   0.044 |   1.332 |   0.122 |
| Dep10  |  1.393 |   0.044 |   0.740 |   0.124 |
| Dep11  |  1.393 |   0.044 |   0.826 |   0.117 |
| Dep12  |  1.393 |   0.044 |   0.530 |   0.132 |
| Dep13  |  1.393 |   0.044 |   0.920 |   0.112 |
| Dep14  |  1.393 |   0.044 |   1.074 |   0.119 |
| Dep17  |  1.393 |   0.044 |   2.321 |   0.144 |
+--------+--------+---------+---------+---------+
```

```
               ITEM FIT STATISTICS

 --------------------------------------
 |  ITEM  | CHI-SQUARE |  D.F. | PROB. |
 --------------------------------------
 | Dep4   |   10.67171 |   14. | 0.712 |
 | Dep5   |    9.36957 |   11. | 0.589 |
 | Dep10  |   21.42770 |   15. | 0.123 |
 | Dep11  |   24.66578 |   13. | 0.026 |
 | Dep12  |   29.78602 |   12. | 0.003 |
 | Dep13  |   11.97851 |   13. | 0.530 |
 | Dep14  |   14.45008 |   13. | 0.343 |
 | Dep17  |   16.56430 |   10. | 0.084 |
 --------------------------------------
 | Total  |  138.91368 |  101. | 0.007 |
 --------------------------------------
```

Exclusion of Dep2 yields a marginally better model (smaller S.E. for the item locations and a somewhat better fit statistic). A comparison of $\chi^2_{\text{d.f.=111}}$ vs $\chi^2_{\text{d.f.=101}}$ (Dep2 excluded) might be used as an indication of a better model where Dep2 is excluded. However, a formal comparison is not reliable due to dependence between items, caused by the constraints – common slopes and a common set of category thresholds.

Table V.Dep.3.3. A GRM with item specific slopes

```
CATEGORY PARAMETER  : 1.091     0.459    -0.392    -1.158
S.E.                : 0.029     0.031     0.040     0.064
+--------+---------+---------+---------+---------+
| ITEM   |  SLOPE  |   S.E.  |LOCATION |   S.E.  |
+========+=========+=========+=========+=========+
| Dep2   |   0.780 |   0.089 |   1.346 |   0.186 |
| Dep4   |   1.759 |   0.195 |   0.770 |   0.116 |
| Dep5   |   1.355 |   0.150 |   1.404 |   0.144 |
| Dep10  |   2.116 |   0.305 |   0.630 |   0.111 |
| Dep11  |   2.492 |   0.295 |   0.877 |   0.113 |
| Dep12  |   1.928 |   0.220 |   0.491 |   0.122 |
| Dep13  |   1.234 |   0.143 |   0.944 |   0.130 |
| Dep14  |   1.047 |   0.138 |   1.115 |   0.134 |
| Dep17  |   0.632 |   0.140 |   2.921 |   0.283 |
+--------+---------+---------+---------+---------+

               ITEM FIT STATISTICS

 --------------------------------------
 | ITEM   | CHI-SQUARE |  D.F. | PROB. |
 --------------------------------------
 | Dep2   |   44.58747 |   12. | 0.000 |
 | Dep4   |   10.79286 |   11. | 0.461 |
 | Dep5   |   17.68989 |   11. | 0.089 |
 | Dep10  |   15.31368 |   11. | 0.168 |
 | Dep11  |   13.38690 |   10. | 0.202 |
 | Dep12  |   34.20889 |   11. | 0.000 |
 | Dep13  |   12.69316 |   12. | 0.392 |
 | Dep14  |   16.04209 |   12. | 0.189 |
 | Dep17  |   14.77394 |    8. | 0.063 |
 --------------------------------------
 | Total  |  179.48886 |   98. | 0.000 |
 --------------------------------------
```

When the constraint of common slopes is released and item specific slopes are estimated, the GRM model shows a bad fit. In particular, Dep2 with a chi-square= 44 against 12 d.f. stands out as a bad fitting item and should be removed in the first place.

Table V.Dep.3.4. A GRM with item specific slopes and Dep2 excluded.

```
CATEGORY PARAMETER  :  1.077     0.439    -0.388    -1.128
S.E.                :  0.029     0.030     0.039     0.059
+--------+--------+---------+---------+---------+
| ITEM   | SLOPE  |   S.E.  |LOCATION |   S.E.  |
+========+========+=========+=========+=========+
| Dep4   |  1.762 |   0.207 |   0.716 |   0.109 |
| Dep5   |  1.310 |   0.151 |   1.327 |   0.140 |
| Dep10  |  2.367 |   0.327 |   0.575 |   0.107 |
| Dep11  |  2.480 |   0.295 |   0.794 |   0.108 |
| Dep12  |  1.863 |   0.218 |   0.432 |   0.120 |
| Dep13  |  1.270 |   0.149 |   0.873 |   0.124 |
| Dep14  |  1.142 |   0.136 |   1.031 |   0.133 |
| Dep17  |  0.667 |   0.143 |   2.786 |   0.277 |
+--------+--------+---------+---------+---------+
```

```
               ITEM FIT STATISTICS
 -------------------------------------
 |ITEM   | CHI-SQUARE |  D.F. | PROB. |
 -------------------------------------
 |Dep4   |    4.46206 |   11. | 0.954 |
 |Dep5   |   19.75089 |   10. | 0.032 |
 |Dep10  |    8.31840 |   11. | 0.686 |
 |Dep11  |   16.25657 |   11. | 0.131 |
 |Dep12  |   28.56251 |   11. | 0.003 |
 |Dep13  |   14.63607 |   11. | 0.199 |
 |Dep14  |   13.66151 |   12. | 0.323 |
 |Dep17  |   14.38452 |    8. | 0.072 |
 -------------------------------------
 | Total |  120.03252 |   85. | 0.007 |
 -------------------------------------
```

A somewhat better fit is achieved, but there is a large variation between the item slopes.

Table V.Dep.3.5. Approximate relative item information (%) based
on a GRM with a common set of category thresholds (constrained model)
vs an unconstrained model (8 items are used).

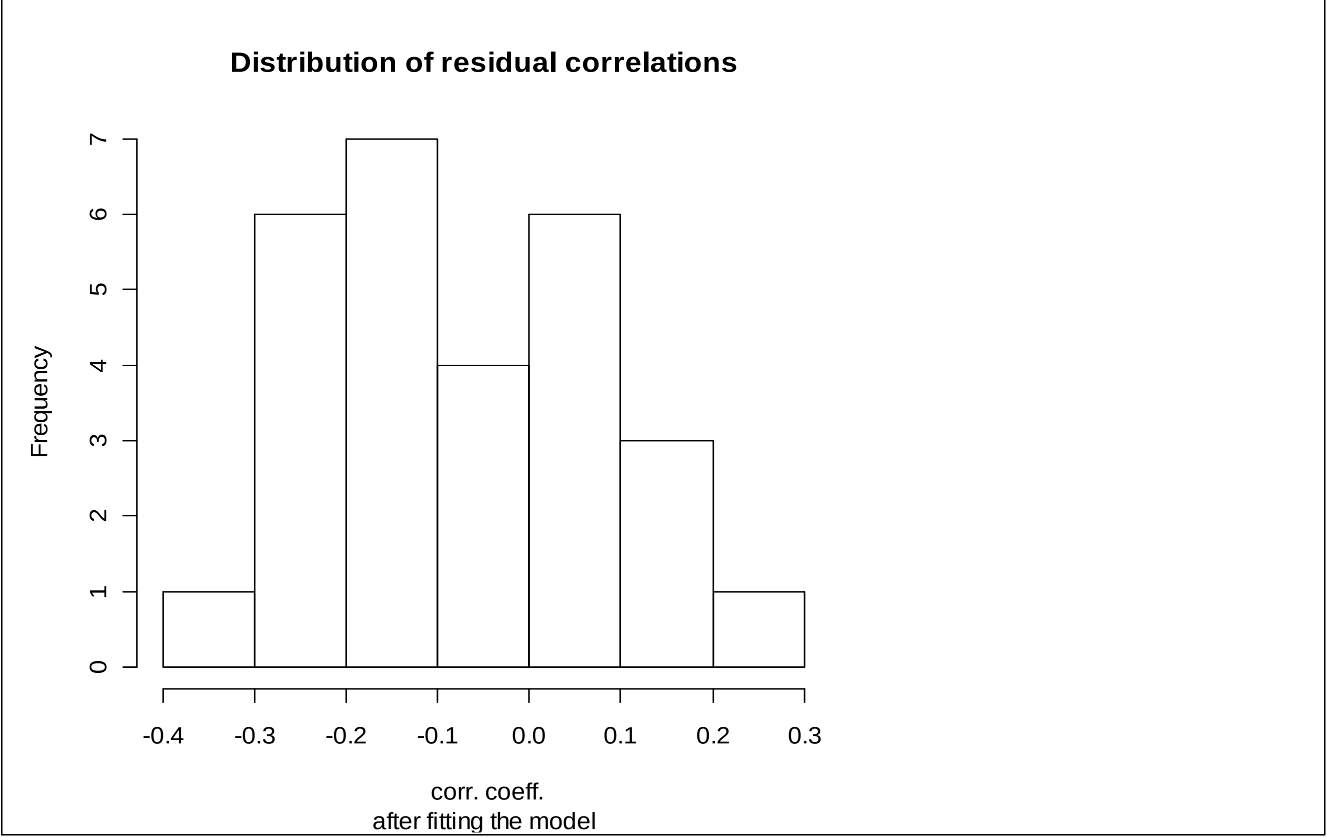
	Rel. info % GRM constr.	Rel. info% GRM unconstrained
Dep4	14.1	16.4
Dep5	8.1	7
Dep10	21.8	23.4
Dep11	23.2	19.8
Dep12	15.4	12.9
Dep13	8.3	9.4
Dep14	6.7	6.7
Dep17	2.5	4.4

The estimated size of relative item information, based on a GRM with item specific slopes, is reasonable as compared with an unconstrained model. Dep17 appears not to yield very much information. However, this item has a high location (2.786), is sparsely endorsed, and consequently is estimated with a relatively low precision, S.E.= 0.277, see Table V.Dep.3.4.

Table V.Dep.3.6. Residual correlations after fitting the GRM with 8 items

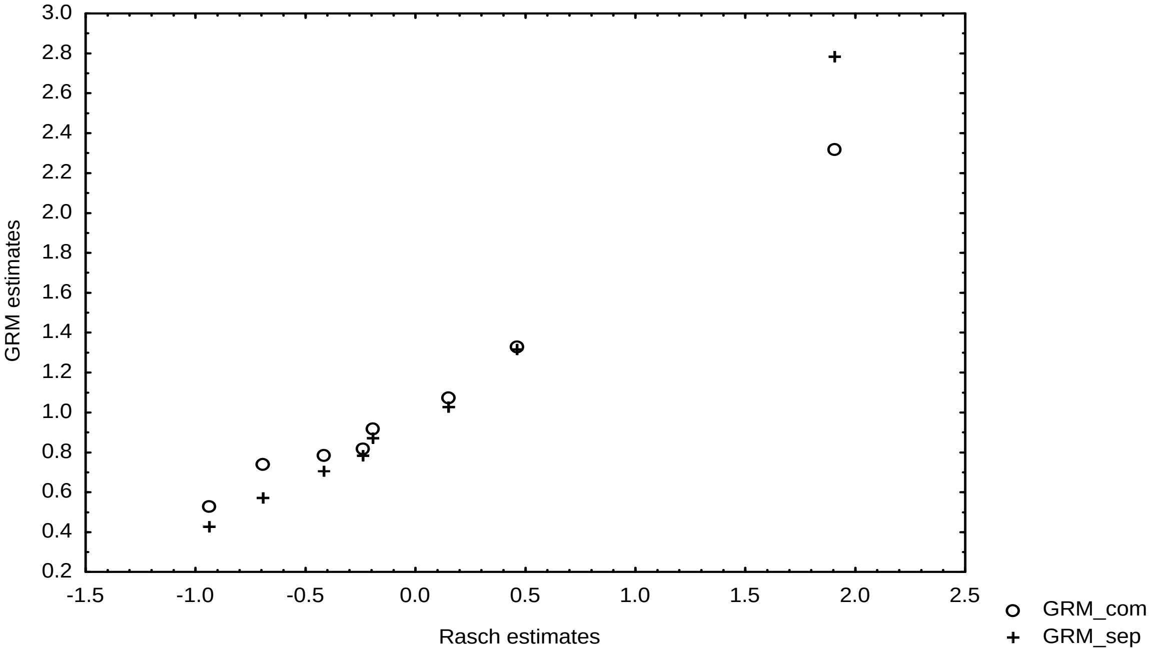
```
        Dep4  Dep5 Dep10  Dep11  Dep12  Dep13  Dep14  Dep17
Dep4    1.00 -0.09  0.18  -0.28  -0.27   0.07  -0.04   0.09
Dep5   -0.09  1.00 -0.20  -0.13  -0.04  -0.18   0.17  -0.26
Dep10   0.18 -0.20  1.00  -0.13  -0.39   0.03  -0.20   0.10
Dep11  -0.28 -0.13 -0.13   1.00   0.01  -0.22  -0.26  -0.18
Dep12  -0.27 -0.04 -0.39   0.01   1.00  -0.22  -0.12  -0.06
Dep13   0.07 -0.18  0.03  -0.22  -0.22   1.00   0.27   0.15
Dep14  -0.04  0.17 -0.20  -0.26  -0.12   0.27   1.00   0.01
Dep17   0.09 -0.26  0.10  -0.18  -0.06   0.15   0.01   1.00
```

Fig V.Dep.3.1.   Distribution of 8*(8-1)/2 residual correlations after fitting the GRM.



The distribution of residual correlations is fairly symmetric around zero.

Fig. V.Dep.3.2. Estimated item locations from the three models, Rash RSM (along the x-axis), GRM with common slopes(o) and GRM with item specific slopes(+), all with a common set of thresholds within the model. Dep2 is excluded.



In fig. V.Dep.3.2., the order of item locations (12,10,4,11,13,14,5,17) is the same for all three models as well as the relative position of the locations.

Table V.Dep.3.7. Test of item specific slopes in an unconstrained model (ltm).
LR test of a model with a common slope vs a model with item specific slopes. Item specific thresholds are used for both models.

```
Likelihood Ratio Table
                  AIC     BIC  log.Lik   LRT df p.value
constr_model   2612.93 2712.51 -1273.47
unconstr_model 2580.22 2700.91 -1250.11 46.72  7  <0.001
```

The result was approximately the same when the answer categories were reduced
(0,1,2,3,4) -> (0,1,2,3,3).

The LR test tells us that the need for item specific slopes was not caused by the constraint of a common set of categories.

Conclusion about the ‘Depression’ questionnaire

Although there were 228 patients included in the study, a substantial part of the answers did not yield any information about the properties of the questionnaire, besides its capacity to separate the respondents into symptoms/ no symptoms with an artificial limit given by the questionnaire. It can further be recognized that out of 228*9 = 2052 possible positive answers (an answer to an item above the lowest category), only 802 (≈ 40%) were in the category 1-4. If we exclude patients with incomplete questionnaires and with a sum score =0, there are only n=151 left for a reasonable evaluation of the properties of the questionnaire. Dep2 is detected, already in Step 1, as a problematic item, violation against monotonicity and non-intersection. It should be replaced by another item or substantially reformulated. Furthermore, from fig. V.Dep.3.2. it can be concluded, irrespective of any model, that there is a gap between the bulk of items and Dep17. As there are patients endorsing Dep17, this is an indication that items in this ‘empty interval’ are needed.

Common or item specific slopes?
From table V.Dep.2.3. and table V.Dep.3.1. it is indicated that a model with common slopes (Rasch or GRM) is unlikely. This is further confirmed by the LR test in table V.Dep.3.7. Large item slopes might stem from high correlation between residuals but table V.Dep.2.4. and table V.Dep.3.6. show moderate values. Unconstrained estimation of item information agrees with the GRM model (table V.Dep.3.5.), sending the message that Dep17, so far, does not seem to contribute very much. However, it cannot be rolled out as there were not many patients in the neighbourhood of that item’s location on the depression scale. This is also reflected by the relative large s.e. of the location estimate for Dep17, (Table V.Dep.3.4.).

The insufficient coverage for the Depression as well as for the Mania questionnaire is discussed and illustrated in the article.

Investigation of the ‘Mania’ questionnaire (Man)

The 227 questionnaires have to be regrouped according to the frequency of non-responses and the ‘all-zero’ scoring patients. Unfortunately, this result in substantially different groups depending on which analysis is carried out. Rasch models tolerate non responses, while most of the other models do not. The subgroups are specified in table V.Man.0.1.

Table V.Man.0.1. Material for evaluation of the maniac questionnaire

Total sample	N= 227
Complete cases	n= 200
Cases with ∑Y(i)= 0, i=1,…,9	n= 67
Cases with ∑Y(i)> 0, i=1,…,9	n= 133
Patients included in the evaluation	n= 227-67= 160

Fig. V.Man.0.1. Distribution of the sum score. Mania dimension. N= 200.

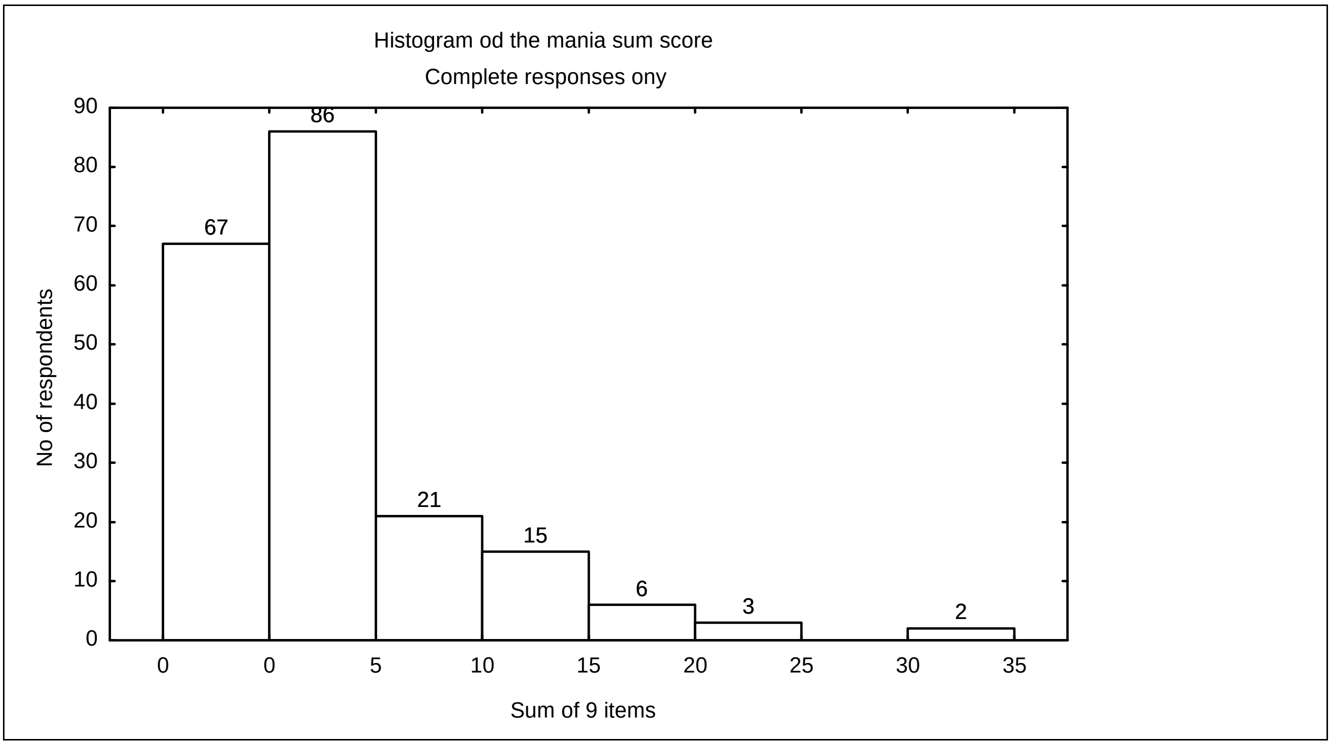


Fig. V.Man.0.1. reveals a sum score distribution concentrated to the lower end.

The sample of N= 160 subjects, see table V.Man.0.1., consists of:
133 subjects with complete answer profiles and ∑Y(i)> 0.
24 subjects with ∑Y(i)> 0, but with one or more items with a ‘non response’.
3 subjects with ∑Y(i)= 0, but with one or more items with a ‘non response’. These 3 subjects are included as we are not able to say that they have responded zero to all items.
This group is analysed in Step2.

Table V.Man.0.2. Response frequences in categories. N= 160 patients.

Category	0	1	2	3	4	Non resp.
Man1	101	31	19	3	3	3
Man3	98	23	23	7	2	7
Man6	101	30	20	3	4	2
Man7	91	30	21	9	4	5
Man8	79	37	23	11	7	3
Man9	49	42	43	13	6	7
Man15	115	17	15	1	2	10
Man16	109	18	14	5	3	11
Man18	114	20	8	5	2	11

The proportion of non responses, extracted from table V.Man.0.2., is just about 4% although their number, 59, is quit large.

Step 1 Man. The Mokken scale analysis.

All items show scalabilities well above 0.3 and in the range [0.37, 0.63]. Just two item pair scalabilities are in the range [0.7, 0.8] they will be further investigated in Step 2.
The item set scalability= 0.49, which indicates an at least moderate scale (table 3 in the article).

Table V.Man.1.1. Analysis of monotonicity. Min. group size= 20. N= 133.

```
       ItemH #vi  maxvi  sum  zmax #zsig
Man1    0.47   1   0.05 0.05  0.63     0
Man3    0.43   2   0.08 0.14  0.80     0
Man6    0.63   0   0.00 0.00  0.00     0
Man7    0.50   0   0.00 0.00  0.00     0
Man8    0.57   2   0.15 0.19  0.80     0
Man9    0.45   9   0.63 2.09  4.71     4
Man15   0.44   0   0.00 0.00  0.00     0
Man16   0.51   0   0.00 0.00  0.00     0
Man18   0.37   3   0.06 0.15  0.32     0
```

The analysis of monotonicity showed only a few violations. However, a number of violations were seen for Man9 (fig. V.Man.1.1.), stemming from decreasing probabilities for low scoring patients. The mania levels are represented by rest scores $R_{(\text{-Man9})}$.
The significance in the last column in table V.Man.1.1. depends, however, on the choice of the rest score grouping, i.e. the minimum group size. Significant violations disappear when the group size is set to 25 or larger. With a group size of 30 there will be only three rest score groups, which results in a more conservative analysis.

Fig. V.Man.1.1. Monotonicity of Man9 (table V.Man.1.1.) vs the mania levels represented by the rest score groups.

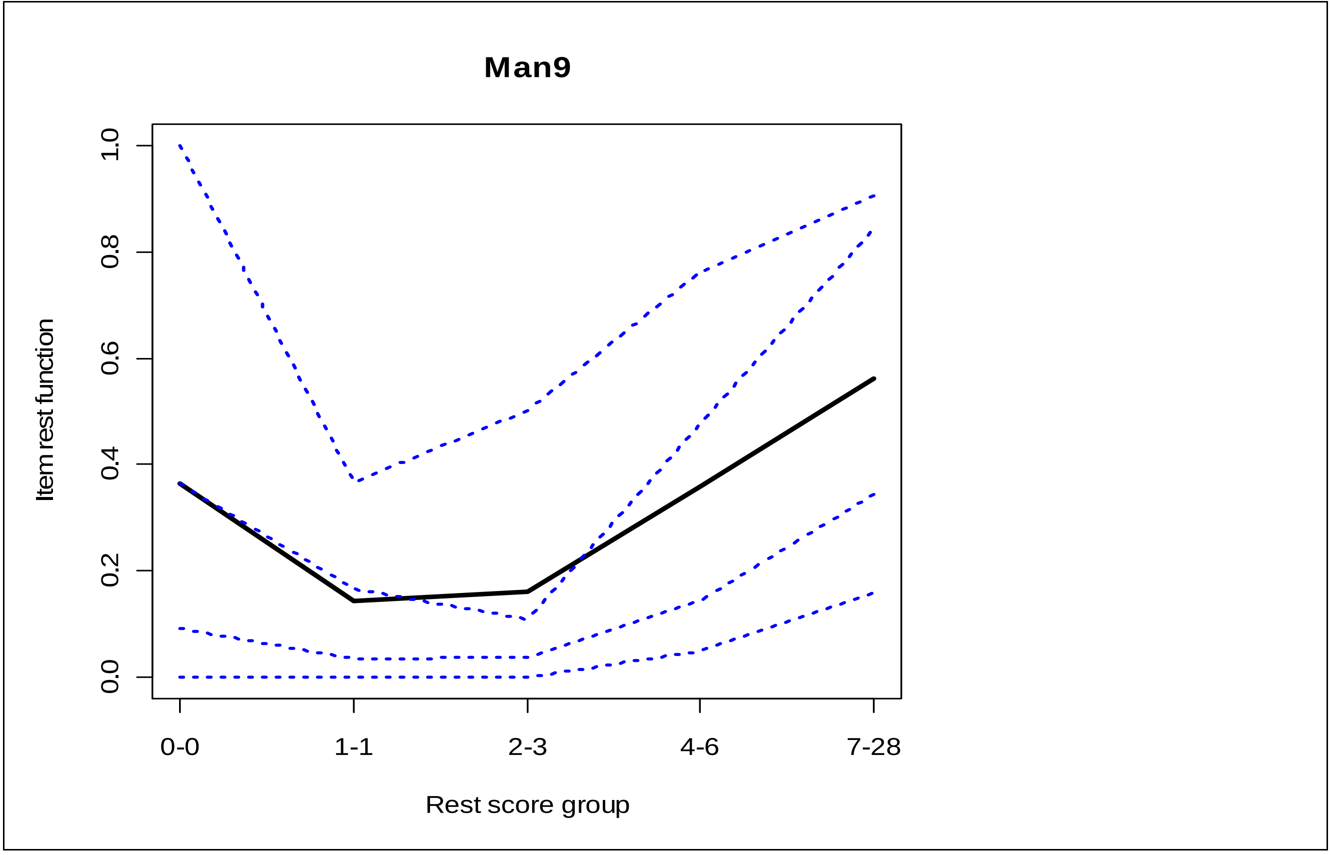
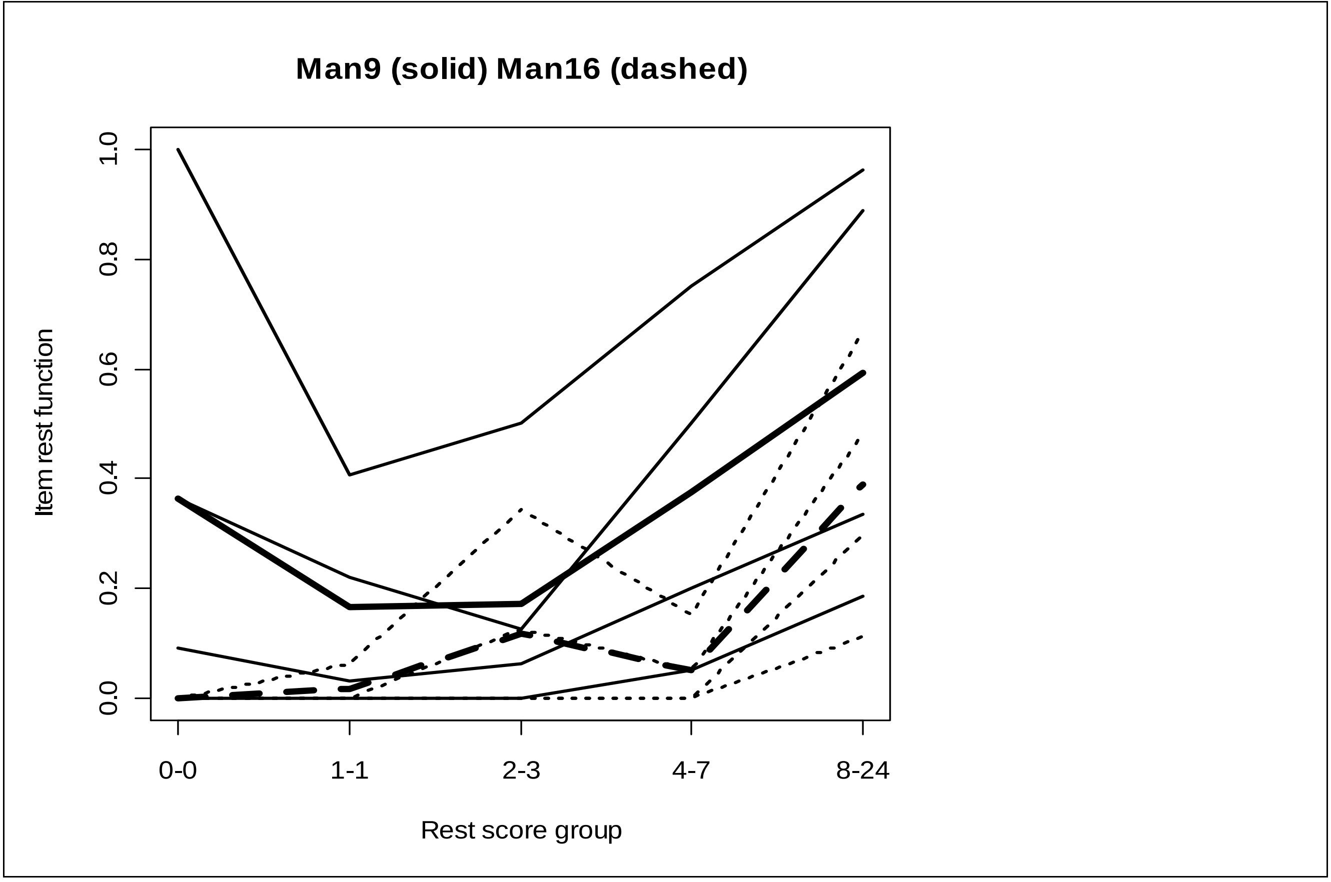


Table V.Man.1.2. Analysis of non-intersection. Minimum group size=20.

	ItemH	#ac	#vi	#vi/#ac	maxvi	sum	sum/#ac	zmax	#zsig
Man1	0.47	512	4	0.01	0.14	0.51	0	1.21	0
Man3	0.43	464	4	0.01	0.22	0.61	0	1.21	0
Man6	0.63	480	5	0.01	0.14	0.65	0	1.21	0
Man7	0.50	448	2	0.00	0.15	0.25	0	0.86	0
Man8	0.57	464	5	0.01	0.22	0.79	0	1.21	0
Man9	**0.45**	**512**	**4**	**0.01**	**0.22**	**0.66**	**0**	**2.07**	**1**
Man15	0.44	480	1	0.00	0.12	0.12	0	0.55	0
Man16	**0.51**	**496**	**7**	**0.01**	**0.22**	**0.97**	**0**	**2.07**	**1**
Man18	0.37	464	4	0.01	0.19	0.55	0	1.55	0

There are a few, but significant, violations against non-intersection. These persist in spite of a minimum violation threshold as high as 0.1. The same result appears even for a minimum group size=30.

Fig. V.Man.1.2. Non-intersection. Man 9 vs Man16 (table V.Man.1.2.). Minimum group size=20.



Although there are some itersections in fig. V.Man.1.2., the rescaled expected responses (the thick solid and the thick dashed line) do not intersect.

A parametric approach might be tried with special attention on Man9 and Man16. The ‘ordering structure’ in a parametric model should be investigated.

Step 2 Man. Analysis by a Rasch model

A Rasch RSM is fitted to the data.

Table V.Man.2.1. Application of the Rasch RSM with 9 items, n= 160.

```
------------------------------------------------------
|          MODEL|   INFIT   |EXACT MATCH|ESTIM|       |
| MEASURE  S.E. |MNSQ  ZSTD| OBS%  EXP%|DISCR| ITEM  |
|---------------+----------+-----------+-----+-------|
|     .14     .12|1.09    .6| 59.6  63.0|  .91| Man1  |
|     .00     .12|1.20   1.4| 62.6  62.3|  .88| Man3  |
|     .07     .12| .57  -3.6| 69.5  62.5| 1.28| Man6  |
|    -.21     .11|1.03    .3| 63.3  60.1| 1.02| Man7  |
|    -.52     .10| .78  -1.8| 62.9  55.6| 1.18| Man8  |
|   -1.06     .10|1.05    .5| 42.6  44.8|  .78| Man9  |
|     .65     .15|1.26   1.4| 75.3  72.5| 1.01| Man15 |
|     .33     .13|1.08    .6| 64.6  66.3| 1.09| Man16 |
|     .60     .14|1.61   3.1| 75.2  72.6|  .79| Man18 |
|---------------+----------+-----------+-----+-------|
|Mean   .00      |Person rel. = 0.62
|S.D.   .51      |  Item rel. = 0.93
------------------------------------
```

There are no immediate objections against the Rasch RSM. However, Man6 shows a possible redundancy and too much noise is indicated from Man18.
Raw variance explained= 52%. There was no concern about any strong second dimension.

Average depression score did not ascend with category score for Man 1 and Man18, but there were only minor violations in just a few categories..

Table V.Man.2.2.

```
Largest standardized residual correlations
used to identify dependent items
---------------------------
|CORREL-|        |        |
|  ATION| ITEM   | ITEM   |
|-------+--------+--------|
|   .32 | Man15  | Man16  |
|-------+--------+--------|
|  -.32 | Man6   | Man9   |
|  -.30 | Man3   | Man9   |
---------------------------
```

A slight correlation between residuals cor(Man15, Man16)= 0.32.
The item set seems adequate, item reliability 0.93, but the person reliability is not sufficient. The Rasch RSM might not be the best. An extended model might give further information.

Step 3 Man

An extended model, a GRM with equal slopes was applied. The set of categories had to be reduced, (1,2,3,4,5) -> (1,2,3,4,4), due to sparse data in category 4 and 5.

Table V.Man.3.1. GRM with equal slopes and a common set of categories.

```
   CATEGORY PARAMETER  :     0.953     0.068    -1.021
   S.E.                :     0.045     0.055     0.085
+--------+--------+---------+---------+---------+
| ITEM   | SLOPE  |   S.E.  |LOCATION |   S.E.  |
+========+========+=========+=========+=========+
| Man1   |  0.956 |   0.038 |   1.553 |   0.172 |
| Man3   |  0.956 |   0.038 |   1.384 |   0.163 |
| Man6   |  0.956 |   0.038 |   1.366 |   0.166 |
| Man7   |  0.956 |   0.038 |   1.255 |   0.166 |
| Man8   |  0.956 |   0.038 |   0.991 |   0.156 |
| Man9   |  0.956 |   0.038 |   0.381 |   0.162 |
| Man15  |  0.956 |   0.038 |   2.024 |   0.184 |
| Man16  |  0.956 |   0.038 |   1.784 |   0.172 |
| Man18  |  0.956 |   0.038 |   1.999 |   0.183 |
+--------+--------+---------+---------+---------+

      ITEM FIT STATISTICS
 --------------------------------------
 | ITEM   | CHI-SQUARE |  D.F. | PROB. |
 --------------------------------------
 | Man1   |    8.53224 |    8. | 0.383 |
 | Man3   |    6.65026 |    8. | 0.576 |
 | Man6   |    9.69223 |    8. | 0.287 |
 | Man7   |    6.09833 |    8. | 0.638 |
 | Man8   |   11.77232 |    9. | 0.226 |
 | Man9   |    6.96373 |    9. | 0.642 |
 | Man15  |    6.82219 |    7. | 0.448 |
 | Man16  |   22.95386 |    7. | 0.002 |
 | Man18  |    7.90057 |    7. | 0.341 |
 --------------------------------------
 |  TOTAL |   87.38574 |   71. | 0.091 |
 --------------------------------------
```

Equal slopes seems OK besides Man16.

In order to explore the structure of item slopes, a GRM with item specific slopes is fitted.

Table V.Man.3.2. A GRM with item specific slopes. The common set of categories is maintained.

```
   CATEGORY PARAMETER  :     0.924     0.069    -0.993
   S.E.                :     0.040     0.050     0.081
+--------+--------+---------+---------+---------+
| ITEM   | SLOPE  |   S.E.  |LOCATION |   S.E.  |
+========+========+=========+=========+=========+
| Man1   |  1.066 |   0.149 |   1.527 |   0.184 |
| Man3   |  0.855 |   0.125 |   1.503 |   0.204 |
| Man6   |  2.386 |   0.678 |   1.210 |   0.123 |
| Man7   |  1.163 |   0.138 |   1.105 |   0.173 |
| Man8   |  1.337 |   0.199 |   0.950 |   0.136 |
| Man9   |  0.917 |   0.103 |   0.413 |   0.175 |
| Man15  |  0.993 |   0.144 |   2.421 |   0.255 |
| Man16  |  0.537 |   0.181 |   1.995 |   0.215 |
| Man18  |  0.722 |   0.098 |   2.157 |   0.288 |
+--------+--------+---------+---------+---------+
```

```
      ITEM FIT STATISTICS
 --------------------------------------
 | ITEM   | CHI-SQUARE |  D.F. | PROB. |
 --------------------------------------
 | Man1   |   11.31320 |    7. | 0.125 |
 | Man3   |    5.44094 |    8. | 0.711 |
 | Man6   |    7.74981 |    7. | 0.355 |
 | Man7   |    7.47629 |    9. | 0.589 |
 | Man8   |   13.88530 |    8. | 0.084 |
 | Man9   |   18.87032 |   11. | 0.063 |
 | Man15  |   14.09861 |    7. | 0.049 |
 | Man16  |   17.25715 |    7. | 0.016 |
 | Man18  |   11.29360 |    7. | 0.125 |
 --------------------------------------
 | TOTAL  |  107.38521 |   71. | 0.003 |
 --------------------------------------
```

Man16 can still be questioned. Man15 and Man16 are closely related, residual cor(15,16)= 0.47.

Table V.Man.3.3. Approximate relative item information (%) based on a GRM (reduced scale) with a common set of category thresholds (constrained model) vs an unconstrained model.

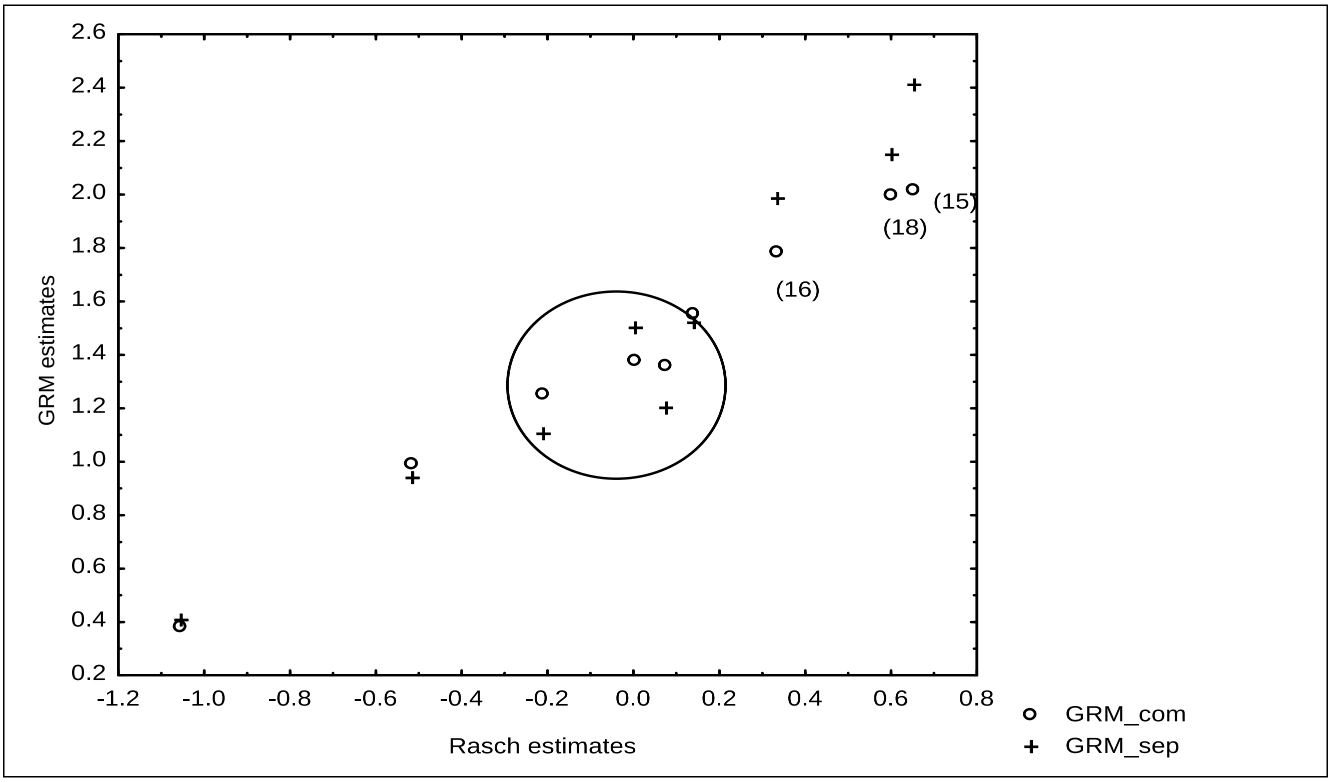
	Rel.info % GRM constr.	Rel.info % GRM unconstrained
Man1	10.2	8.2
Man3	7.3	8.6
Man6	31.4	24.6
Man7	11.7	12.2
Man8	14.4	16.4
Man9	8.4	4.5
Man15	8.2	8.9
Man16	3.3	8.9
Man18	5.2	7.6

Table V.Man.3.3. indicates that the GRM agrees with the unconstrained model except for Man16. In the GRM approach Man16 seems to be an item with very little contribution while the unconstrained model consider Man16 at the same information level as most of the other items.
LR test of common vs item specific slopes in a model with a common scale (one set of category thresholds). (PSL):
Common slope vs specific slopes: diff(-2 LogL) = 2116.4 - 2088.8 = 27.6, d.f ≈ 8, p<0.01.

LR test of common vs item specific slopes in a model with separate scales (item specific category thresholds) (ltm):
Common slope vs specific slopes diff(-2 LogL) = 2094.0 – 2081.8 = 12.2, d.f ≈ 8, p≈ 0.143.

The two tests do not agree. The above analysis says that if the constraint $\tau_{ik} = \tau_k$ i=1,…,9; k=1,…,4 is imposed on the categories, item specific slopes are needed. Otherwise, item specific slopes are not justified.

Fig. V.Man.3.1. Estimated item locations from the three models, Rash RSM (along the x-axis), GRM with common slopes(o) and GRM with item specific slopes(+), all with a common set of thresholds within the model.



The order of items is approximately maintained for all three models but is marginally reversed for Man3, Man6 and Man1 (encircled in fig. V.Man.3.1.).

Conclusion about the ‘Mani’ questionnaire

Even if the relative information seems to vary substantially between items, we cannot distinguish between common and item specific slopes. Man16 should be reformulated. A non-negligible residual correlation with Man15 is indicated. More items in the lower part of the scale is desirable, either by moving (reformulation of items) or (which should be preferred) by adding further items, to improve the coverage.

Conclusion about the Affective Self Rating Scale questionnaire

The analyses reveal moderate to strong questionnaires.
In general, the items’ reliabilities are sufficient but the person reliabilities are just moderate, which require more items. The gap in the Depression item locations have to be filled in with some suitable items.
The opposition between a ‘cut off’ and an ‘estimate of a level’ of Depression/Mani questionnaire should be addressed.

Discussion

The suggested explorative strategy, outlined in three steps, is specifically aimed for analyses of symmetric questionnaires, with a common set of ordered response categories, in combination with small samples of respondents.

It can be noticed from all five studies that the first step, the Mokken analysis, reveals immediately the most basic characteristics of the questionnaires. The following two steps are mostly a refinement and an adjustment to a parametric environment, as interval scaled person and item measures usually is a requirement (or at least the intention) from the researchers. Their object is to use the estimated person and item measures for presentation and further analyses in agreement with what we might call the 'conventional statistics' with all its methods according to a CTT approach. In the light of the researchers' intentions, Step 1 would be treated as a step backwards and Step 3 as a step forward beyond the Step 2, which might be considered as 'the straight forward sum score' approach.
Much of the analysis is focused on detection of unanticipated patterns in the data (the questionnaire), yielding a range of suggestions for an improvement of the questionnaire. This should be seen as an explorative process, where the questionnaire, in the form of a structure of an outcome space, should rely on a common set of measurement units and hopefully constitute an orthogonal space, conditioned on the person measure θ on the intended dimension.
On the other hand, the researcher generally has a hidden/unexpressed hypothesis that his/her questionnaire is suitable for a simple sum score approach. Such a hypothesis can be formally investigated by statistical tests which, via sometimes significant objections against the hypothesis, should lead to deletion or radical reformulation of particular items. This is what we would call a confirmatory analysis. However, with a small sample in mind and the explorative nature of the '3- step' strategy, I am satisfied with a recommendation to thoroughly follow specified items or structures when further persons are involved to respond to the questionnaire. Furthermore, if the goal is to create a questionnaire, equally applicable for specified subgroups, such as gender or clinically defined, suitable for a straight forward sum score approach, the researcher should take action to actually change the questionnaire according to the recommendations. 'Any sensible investigator will realise the need for both explanatory and confirmatory techniques, and many methods will often be useful in both roles' [Everitt & Dunn, 1991].

Step 1: The person measure, as represented by the raw sum score $T_n = \sum Y_{ni}$, for person n, n=1,2,…,N, on the items i=1,2,…,I, together with individual item scores is the starting point for this step – without any particular assumptions.
The primary role of the ordinal statistic T is a reasonable measure for arranging the individuals on the intended latent dimension according to the size of T.

1. An analysis of the dimensionality, performed in Step 1, depends heavily on the chosen scalability cut off. Although hazardous, it might well give some valuable information and a clue to what will happen in Step 2 and Step 3.
2. Reasonable item scalabilities, $0.2 < H_i < 0.5$ were mostly found. $H_i < 0.2$ pointed out weak items, which turned out to cause problems in the subsequent steps. However, the authors of the questionnaires usually succeed in a coherent set of items. Contradictory items, $H_i \leq 0$, are relatively rare. Conspicuous, large pair scalabilities, frequently seen in the Mokken analysis may hide a disturbing local dependence. However, it turns out that most of them are 'dissolved'

by a parametric model, while some will resist and point out local dependence conditioned on the chosen model.

3. Cleaning the questionnaire from non-contributing items already in step 1 seems a good advice. Such a measure usually leads to reasonable monotonicity and facilitates the parametric approach.
4. The information from the non-intersection investigation is difficult to estimate due to the small and mostly too few rest score groups. At the best, the violation of IIO can be traced back to items with low scalability. Unless there are serious violations, no action should be taken.

Incoherent answer profiles are difficult to unmask in Step 1 as there are no 'residuals ' or 'expected values'. However, with a small sample, the estimation of the scalabilities are sometimes much influenced by just one or two profiles. Screening the distribution of an item scalability by a 'jackknife' approach, which means jackknifing persons, might extract influential individuals.

Not much can be definitely proved or disapproved in a small sample setting, but in some cases there are convincing statistical demonstrations against basic assumptions, mostly a deviant behaviour of certain items. These should be earmarked and thoroughly followed in Step 2 and Step 3.

From Study I – V we can conclude that Step 1 is a necessary procedure before trying a parametric model. The results from Step 1 help us to dissect problems in the succeeding steps.

Step 2: Step 2 uses the person sum score to create an interval scaled measure, based on the Rasch approach (the person sum score is a sufficient statistic). In the first place (with a positive result from Step 1), the fit statistics and estimated item discriminations is an essential information about a reasonable model. Unidimensionality, or rather, verification of a dominating dimension, is indicated by small contrast values when investigating the residuals, given the chosen model.
Secondly, even if the model is relevant it might not be sufficient. As is often the case, as seen in Study I (AA and AM), Study II (Social dimension) and Study V, the item reliability is sufficient while the person reliability is rather low. This is a strong indication that more items are needed, usually for covering certain parts of the dimension, which is poorly represented by the given item set. Low item reliability implies low person reliability due to the lack of a suitable set of items.
We have also seen that deficiencies in Step 1 are frequently carried on to Step 2, causing problems with the parametric model. These problems can now be identified as depending on the modelling procedure as such and not traced back to more basic assumptions, represented by Step 1. A low item scalability together with weak monotonicity usually leads to a low discrimination. A strong residual correlation (> 0.4), as a sign of local dependence, is connected with a large scalability in Step1 (Study II, III and IV). However, a large scalability does not necessary imply a strong residual correlation (Study III). Furthermore, violation of non-intersection predicts bad item fit in a Rasch model (Study III)
A large amount of ambiguity is expected to be inherited in a small sample study and a decision on a suitable model structure of the questionnaire might be difficult. Continuing to Step 3 might cast further light on the ambiguities.

Step 3: In Step 3, the sum score as a sufficient statistic is released and the model is made more flexible. Inevitably, the data will 'agree' substantially better with a model with item specific weights and even more with a model with item specific category thresholds.

Some problematic items appear more clearly in the Step 3 approach (AttDef2 and AttDef12 in Study III) and some items seem to require an unconstrained model (AttDef12 in Study III). Some residual correlations pop up in this step although not detected in Step 2 (ASD6 vs ASD8 in Study IV). However, they are usually seen as strong pair scalabilities in Step 1.
With a more 'data driven' model, this step should be seen as an even more explorative analysis. Unrestricted estimation, but still with the constraint of local independence, might confirm the feasibility of the more restricted approach in Step2. Applying an unconstrained model might be helpful when investigating the role of item specific weights.
A certain amount of knowledge concerning the 'item information' can also be gained from this step. At the best, this can be interpreted as the relative importance of an item in the questionnaire. Moving between a parsimonious Rasch model and a model with item specific weights should normally be sufficient to reveal the structure of the questionnaire. The forth and back procedure (Step 2 ↔ Step 3) constitutes essentially of the question of deleting or keeping particular items, which are suspected to hide an underlying common structure, constituted by the rest of the items.

General remarks.

Outlying answer profiles
Outlying persons, i.e. outliers in terms of incoherent answer profiles, are, as in all statistical procedures, disturbing and a cause of misinterpretation and possibly wrong decisions. Some outliers were identified in the studies II and III, but the indication is that they do not much harm when the analysis is directed to the characteristics of the questionnaire. However, they are disturbing in the evaluation process and when estimating parameters aimed for placing individuals on the interval scaled variable. Sometimes, these outliers are identified already in Step 1, which is an advantage, as they are detected independently of any of the parametric model. In general, these should be put aside when evaluating the questionnaire.

Local independence
Local independence cannot be expected to be met. This is a general problem and not inherited in a small sample situation. Frequently, residual correlations in the range 0.3 -0.4 is seen. The importance of these correlations are related to the square of the correlation, which means 0.09 – 0.16 (9% - 16%)
Of course, a small sample might give raise to occasionally large residual correlations, but the problem might also be a consequence of the estimating procedures.
The estimation by likelihoods is based on strictly local independence as the item likelihoods are straight forward multiplied without any covariance terms taken into account. This has implications in some of the statistical tests, which are included in the available programs or can be extracted.
The local independence is also assumed for the chi-square statistics of the overall model fit, where the item chi-square statistics are summed up and the d.f. is the sum of the items' d.f. These tests are of good help but should be interpreted by a wide margin.

Coverage:
It can be concluded from all the five studies that an insufficient coverage is a general problem. This seems to be frequently overlooked. It can be remedied to some extent by use of long rating scales, but they are difficult to construct and might give rise to uncertainty for the respondent. Although the estimation of person locations will be very approximate, extracted from a model based on an 'insufficient' sample, a graph illustrating the relationship between the person location distribution and

the range covered by the items (particularly the item locations) will give valuable hints when reformulating or adding items to the questionnaire. It seems as the distribution of item locations is fairly robust over different models.

Three main types of insufficient/ineffective coverage were noticed:

1. Gaps in item location: The item locations are unevenly distributed, leaving gaps where the intended scale is not represented by any item (PHQ in Study IV).
2. Cluttered item locations: This means a narrow range of a set of items in some of the scales. If the researcher is anxious about too many items in a questionnaire, item locations squeezed together will be waste of resources, as these items yield virtually the same information (Internalising behaviour in Study III). Two items at the same location might be motivated when one of the items has the role of controlling the other one.
3. An insufficient agreement between estimated person measures and item locations: The items do not cover the actual population as represented by the sample. (Social dimension in Study II).

Dimensionality:

The questionnaires in some of the studies are divided into parts, intended to represent different (but certainly close) dimensions. Even if it is not the main objective for this thesis, it is in the scope of ‘evaluation of the questionnaire’ whether the respondents actually perceive the dimensions as defined by the authors. There are methods, within the IRT concept, to investigate how the items actually define the intended dimensions. It is well worthwhile to use these methods in an exploratory way, but we should be aware of the risk that just a few ‘odd’ answer profiles may move items between the dimensions, as they are defined. Unless there are very strong indications, we should refrain from splitting the dimensions as they are decided by the authors. In Study II no clues could be found about the social and psychological dimensions, which was expected due to low scalabilities and heterogeneous subgroups. In study III, with 131 subjects and reasonable scalabilities, at least some information about the dimensions was possible to achieve. In Study IV, a supposed nearness of Depression and Mania was just confirmed.

Statistical tests:

Many IRT statistical programs yield large amount of test statistics and p-values. They appear as sets corresponding to the number of items. These sets should be interpreted with caution.

All statistical tests in the three steps should be viewed as tests of a general hypothesis:
$\mathbf{H_0}$**:** The questionnaire is a reasonable basis for calculating a valid person measure on an interval scale, based on the ‘raw sum score’ or the answer profile for each person in the intended population.

In Step 1, indications against $\mathbf{H_0}$ signify zero or negative scalabilities and violations against monotonicity and non-intersection. These violations are often numerous but should not be taken too seriously unless they are concentrated to particular items, in which case they cannot be denied. Corrections, like Bonferroni correction, have no meaning as ‘each violation in itself provides evidence against the model assumption’ [Molenaar I.W. & Sijtma K., 2000]. Furthermore, the different tables behind the tests have a complex structure of dependency, while independence is generally assumed when using the Bonferroni method. Furthermore, the tests of violations of monotonicity and non-intersection are based on a normal approximation of a hypergeometric distribution, which in turn is based on independence between items. Thus, hesitation against ‘z > 1.65’ should be called for. As is seen in all five studies, the size of the rest score groups is mostly 15 or 20, which is probably not sufficient for a good normal approximation. As a consequence, isolated violations should not be taken too seriously. The z-scores can be used as indicators and used as a warning when they are gathered together for one particular item. It would of course be possible to use exact probabilities, but such a procedure would imply a lot of calculation and programming work. I consider the pragmatic interpretation of the z-scores reasonable for the purpose.

In Step 2, we are directly estimating an interval scaled variable, based on the ‘raw sum score’. The tests are mainly about item fit to a parsimonious model and whether particular items reasonably contribute to the aggregated measure.

The tests in Step 3 might be viewed as a further investigation of the model stated in Step 2, concerning item weights, the item relative information and, once again, identification of negligibly contributing items.
Thus, the statistical tests in Step 1 →Step 2 → Step 3 will constitute a chain of partially dependent tests, where the consistency of an item’s ‘violation against what is assumed’ should be looked for rather than isolated ‘statistical significances’.
Thus, a ‘significance’ is an indication of a ‘deviant behaviour’ rather than a ‘rejection/ non rejection’ of the hypothesis.
As is seen from the investigations of the thirteen scales, I am hesitant to actually suggest removal of particular items, but rather recommend a reformulation and a follow up.

Item and total information
The value of item information can be questioned in a small sample setting but might give some additional explorative information, particularly when we look along the range of person measurements. Principally weak, but also strong items can be verified.

Some special remarks

Study I. The evaluation of this questionnaire clearly shows that Step 1 is indispensable for a pragmatic and scientifically correct evaluation. If a ‘Mokken scale’ cannot be established (as was the case for AR), the researcher should ‘refrain from using the ordinal measurement properties implied by the model’ [van der Ark, 2012] and there is neither a sound basis for a further parametric approach.

Study II. Already Step 1 points out the ambiguity of an attempt to evaluate the questionnaire when there are only about 30 subjects in three specified groups, particularly for the psychological dimension. The large amount of empty cells gives a hint of coming difficulties. A direct start with Step 2, sometimes seen in the literature, might give the impression that a Rasch model would be suitable, but the analysis shows that just a few strong indications can be extracted. We have to be prepared for much of ambiguity when an already small sample has to be divided in differently behaving subgroups (about 30 individuals/group), but this should not refrain us from continuing the evaluation. At least something ought to be gained from such an early evaluation and possible improvements of the questionnaire can certainly be extracted. With a combination of differently behaving, small sized subgroups and weak scalabilities we begin to see the lower limit of what is possible to achieve with the suggested strategy.

Study III. According to recent developments [van der Ark 2012], a screening for dimensions by the old search procedure in 'Mokken scale analysis i n R' might not be the best. However, in a small sample setting, we cannot be expected to get a solution with any precision, but the more simple approach was considered sufficient for our purpose. In essence, the three dimensions, defined as intended, were maintained. The analysis revealed that certain differences between boys and girls have to be taken into account when formulating the questionnaires.

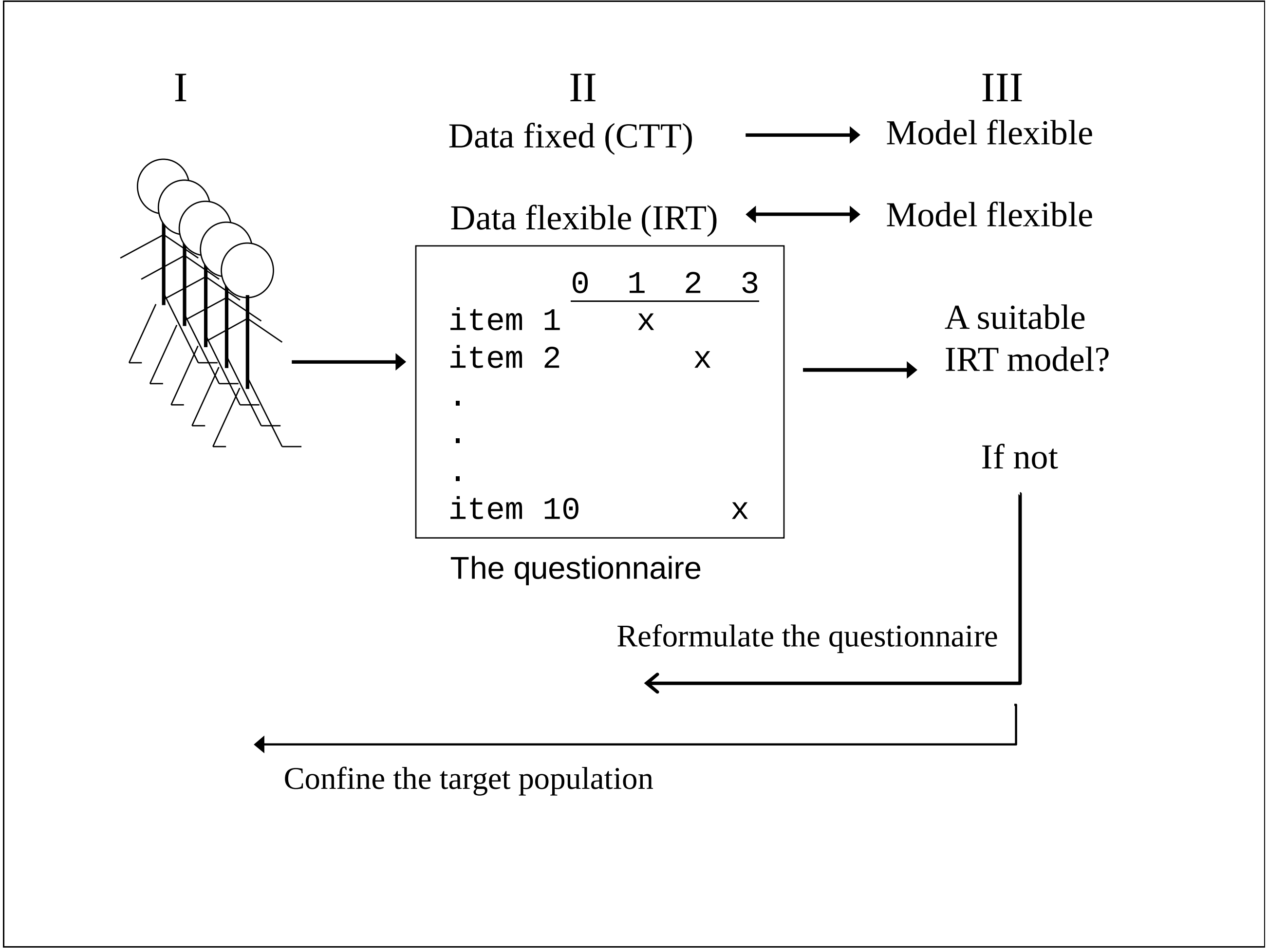
Study IV. The item MADRS4 shows a very deviant behaviour compared to the rest of the items. An exclusion is strongly indicated. However, creating two parallel measures, with and without MADRS4, is easy to perform. If MADRS4 turns out to be non-informative, it does not harm the questionnaire besides its occupation of an 'item place'.

Study V. In this study, we perceive contradictive purposes of a questionnaire. It tries to act both as a 'cut off test' and an instrument to create estimates on a latent scale. If the latter is the purpose of the questionnaire, the treatment of extreme answer profiles is not straight forward. Normally, they should be just a few. Complete 'zero profiles' were excluded but other attempts are of cause possible.
The questionnaire appears too 'difficult' (endorsing high categories). The items should occupy a wider range and a more even distribution. If the questionnaire is aimed for a separation into two groups, the items should be concentrated around a specified cut off point.

The '3 step' strategy in an iterative process

In CTT we usually consider observed data as fixed and the model flexible in order to fit the data. In our case, the data are governed by the structure of the questionnaire, which we would call the structure of the outcome space in mathematical terms. Conditioned on the individuals under investigation, the data and the analyses depend on the formulation of the questionnaire, which makes the questionnaire flexible as well. In other words, if the data do not fit the model, there might be problems we the data. By changing the data (formulation of the questionnaire), a parsimonious model might fit. It is quite possible to decide on the model and 'fit' the questionnaire. This idea was mentioned already by Rasch (according to Andrich, 1988). Furthermore, a manageable questionnaire might require a redefined, delimited target population. Then, in a broader sense, the analysis of a questionnaire can be looked at as an iterative refinement process of phase I (the target population), phase II (the questionnaire) and phase III (the aggregated person measure according to a model) as illustrated in figure 8. A change in one or more of these three phases has its implications and will certainly be recognised by the '3 step' strategy investigation. If we adopt a new approach to the problem, the goal of an iterative process is to bring the three phases into an agreement.
However, if applied, an annoying consequence is that subsequent groups of respondents will respond to slightly different questionnaires, but this problem is a matter beyond the scope of this thesis.

Fig. 8. Phase I, II and III in an iterative refinement of a questionnaire



Summary

The aim of this work, as stated in ‘Background/ The aim of the thesis’, was to create a strategy for evaluation of questionnaires based on a limited sample of respondents. By going through 13 questionnaires, I have suggested a pragmatic and reasonable way to handle ‘small questionnaire studies’ by an application of a strategy, which starts from scratch (Step 1) and then continues with a follow up by a modelling approach. The aim was not to find a suitable model but rather to reveal characteristics of the questionnaire. However, I have to admit that the strategy has its limitations. In case of too small samples in combination with low scalabilities (Step 1), tentatively $N \leq 30$ and an item set scalability < 0.3, Step 2 and 3 might be very hazardous and firm conclusions are hardly possible. But, this should not restrain us from going through the process. There is always a grain of information to gain with the objective to improve a questionnaire at an early state.
It might look strange to move to a more complicated model (Step2 → Step 3 with item specific slopes and category thresholds) when a more parsimonious model is statistically reasonable. However, a fairly large expected type II – error, connected with the fit statistics and a small sample size, and the intention to reveal useful characteristics of the questionnaire, motivate such an extension – ‘prevention is better than cure’.
This thesis suggests that the three steps should always be considered, when feasible. They are strongly related but pay attention to different characteristics of a questionnaire.

‘Don’t say you have found the truth, but rather, I have found a truth.’

Future directions

As pointed out in the introduction, small studies like those seen in this work will frequently be carried out even in the future. It is necessary to have at least some strategic procedure for an on-going evaluation of the questionnaire as the process continues.

It should be possible to investigate how the strategy works with a growing sample size and in connection with varying quality of the questionnaire, such as the degree of scalability. The lower limit of a functioning sample size is of course impossible to know, but with decreasing sample size in some realized studies, it would be worthwhile to investigate how an evaluation strategy like this one loses its capability. This capability is probably strongly connected to the degree of scalability but it has to be further evaluated in detail.

An interesting project, which is more accessible, would be to evaluate the strategy according to an increasing sample size in studies where we have, let us say, about 500 respondents. An evaluation scheme of N = 50, 100, 150, …, 500 should be considered. The samples should be chosen along the order of inclusion of the persons and not randomised from the total sample at the end of the study. This will be a process similar to sequential trials as performed in clinical studies. As an example, a deviant item should not be kept in the questionnaire longer than to an evidence of its 'non contribution' to the measure of interest.

It is quite possible that the characteristics of the process will change with time. The staff involved will be more adapted to the process etc. Simulation is of course another strategy, but such a procedure suffers from the dependence on a theoretical model which is not very convincing. Furthermore, answer profiles not observed will be created.

The structure of the investigation in this thesis is restricted to questionnaires with a symmetric form of ordered response alternatives with a 'sum score' in mind, but the suggested strategy should be tried on other types of questionnaires.

Another aspect, which has been just occasionally addressed in this work, is the question of redundant items in case of a set of items within a very short range. It is quite possible that the same amount of information might be gained by including just a few items, which have their location in this narrow range. On the other hand, this particular range might be hard to grasp with just a few variables. We do not know.

Although suggested as an explorative tool (in Step 3), the usefulness of item information is unclear. It would be worthwhile to bootstrap an unconstrained model together with a constrained model to get an idea of the variability of the items' relative information.

A general aspect: Statistical considerations regarding small groups answering a questionnaire is very much overlooked, in spite of the overwhelming number of clinical studies based on fewer than, let us say, 100 subjects. This problem should be further addressed. As can be seen, the statistical programs mostly use approximations, which are valid for sufficiently large samples. It would be of interest to develop and evaluate routines for small samples, in line with what is done for other types of data. The 'StatXact' [Cytel Software Corporation] is a program dedicated to exact non parametric statistical inference suited for analyses of continuous data and contingency tables based on small data sets.

IRT routines, entirely based on a bootstrap approach, which means statistical analyses and estimations based solely on empirical distributions, would be of value. They should be more extensively applied than what is done in this thesis. This is of particular interest when dealing with small samples, as most of the existing theories rely on large sample approximation of distributions. This will however require a lot of programming work as long as these methods are not included in user friendly statistical packages.

Acknowledgements

For all support in my work of this thesis I wish to express my gratitude to:

Rolf Bergin, my first supervisor, but unfortunately no longer with us, who encouraged me to start research about evaluation of questionnaires, a branch which for a long time has worried my statistical thoughts.

Uno Fors, my current supervisor, for his patient support and for his excellent continual review of my writing and presentation of the results as well as all help with administrative regulations connected with the formulation of the thesis.

Klara Bolander-Laksov, Margret Nisell, Gunilla M. Olsson and **Mats Adler** for providing the data their collaboration about the questionnaires, which constitute the basis of this work.

Elisabeth Berg for statistical discussions and patient support in my writing of the thesis.

Frank Miller for valuable statistical comments.

Nabil Zary, head of CLK, for pushing for continuously pushing me to complete this work.

Mats Brommels, head of department, for having made it possible for me to complete this thesis.

Margareta Krook-Brandt, Eva Hagel, Magnus Backheden and **Mesfin Tessma** at the Medical Statistics Unit for their interest and kindly support.

Samuel Edelbring for interesting discussions about questionnaires.

Ingrid Smedberg for administrative assistance.

Ludvig Andersson and **Erik Attoff** for technical assistance.

References

Adler Mats, Brodin Ulf. An IRT validation of the Affective Self Rating Scale. Nordic Journal of Psychiatry, 2011, Vol 65 No 6, p 396-402.

Adler M., Hetta J., Isacsson G., Brodin U. An Item Response Theory evaluation of three depression assessment instruments in a clinical sample. BMC Med Res Methodol. 2012 Jun 21;12(1):84

Adler M 1§, Hetta J, Isacsson G 1, Brodin U. An Item Response Theory evaluation of three depression assessment instruments in a clinical sample. BMC Med Res Methodol. 2012 Jun 21;12(1):84.

Andrich David. Rasch Models for Measurement. SAGE Publications 1988, vol. 68.

van der Ark L Andries. Mokken Scale Analysis in R. Journal of statistical computing, Febr. 2007,Vol. 20, Issue 11.

van der Ark L Andries. New developments in Mokken Scale Analysis in R. Journal of statistical computing, Maj 2012,Vol. 48, Issue 5.

de Ayala R.J.. The Theory and Practice of Item Response Theory, The Guilford Press, New York 2009.

Bagby RM, Ryder AG, Schuller DR, Marshall MB: The Hamilton Depression Rating Scale: has the gold standard become a lead weight? Am J Psychiatry 2004, 161(12):2163-2177.

Bond Trevor G. & Fox Christine M. Applying the Rasch Model, Fundamental Measurement in Human Science. Lawrence Erlbaum Associates, Publishers, London

Bot S D M, Tervee C B, van der Windt D A W M, Bouter L M, Dekker J, deWet H C W. Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature. Ann. Rheum Dis 2004; 63:335-341 2001

Brodin Ulf, Fors Uno, Bolander Laksov Klara. The application of Item Response Theory on a teaching strategy profile questionnaire. BMC Biomedical Education 2010, 10:14.

Brodin U., Fors Uno GH, Olsson Gunilla M. Adolescent Adjustment Profile - revised and investigated by means of an Item Response Theory approach. (Manuscript).

Davison A.C. & Hinkley D.V.. *Bootstrap Methods and their Application*. Cambridge University Press 1997.

Edelen Maria Orlando, Reeve Bryce B., Applying item response theory (IRT) modeling to questionnaire development, evaluation and refinement. Qual Life Res (2007) 16:5-18.

Efron Bradley, Tibshirani Robert J.. *An Introduction to the Bootstrap*. Chapman & Hall 1993

Embretsson E. Susan, Reise Steven P., Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Publishers 2000, ch2.

Everitt Brian S & Dunn Graham. *Applied Multivariate Data Analysis.* Edward Arnold, University of Cambridge, 1991.

Good Philip I, Hardin James W. Common Errors in Statistics, John Wiley & Sons 2009, p. 125.

Goodman Robert & Scott Stephen. Comparing the Strengths and Difficulties Questionnaire and the Child Behaviour Checklist: Is Small Beatiful? Journal of Abnormal Child Psychology, Vol 27, No. 1, 1999,pp 17-24

Jamieson Susan. Likert scales: how to (ab)use them. Medical Education 2004; 38: 1217-1218.

Linacre J. Optimizing Rating Scale Category Effectiveness. Ch 11 in Smith EV and Smith R.M.: Introduction to Rasch measurement : theory, models and applications. Maple Grove, Minn.: JAM Press 2004.

Linacre J. Winsteps 3.66, Rasch-Model Computer Program. In: Book Winstep 3.66, Rasch-Model Computer Program (Editor ed.^eds.). City; 2008.

Manly Bryan F.J.. *Randomisation, Bootstrapping and Monte Carlo methods in Biology,* Chapman & Hall, 1997.

Molenaar I.W. & Sijtma K. Users Manual MSP5 for Windows. Grningen: iecProGAMMA, 2000.

Nissell M., Brodin, U., Christensen K., Rydelius P-A. The Imperforate Anus Psychosocial Questionnaire (IAPSQ): Its construction and psychometric properties. Child and Adolescent Psychiatry and Mental Health 2009, 3:15 (14 May 2009).

Ostini Remo & Nering Michael L.. Polytomous Item Response Models. SAGE Publications 2006, vol. 144, ch. 3-4.

Olsson, G. M., Mårild, S., Alm, J., Brodin, U., Rydelius, P.-A., & Marcus, C.. The Adolescent Adjustment Profile (AAP) in comparisons of patients with obesity, phenylketonuria or neurobehavioral disorders. *Nordic Journal of Psychiatry, 62*, 66-76, 2008

Parscale 4.1 for Windows by Eiji Muraki, Darrel Bock, 2003. www.scienceplus.nl.

Reeve Bryce B. & Fayers Peter. Applying item response theory modelling for evaluating questionnaire item and scale properties kompl med title och årtal.

Reeve Bryce B., Hays Ron D., Bjorner Jacob B., Cook Karon F., Crane Paul K., Teresi Jeanne A., Thissen David, Revicki Dennis A., Weiss David J., Hambleton Ronald K., Liu Honghu, Grshon Richard, Reise Steven P., Lai Jin-shei, Cella David. Psychometric Evaluation and Calibration of Health-Related Quality of Lif Item Banks. Plans for the Patient-Reported Outcomes Measurement Information System (PROMIS). Medical Care, Vol 45 No.5 suppl.1, May 2007.

Rizopoulos Dimitris. Ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. Journal of Statistical Software 2006, Vol 17, Issue 5.

Selvin Steve. *Modern Applied Biostatistical Methods,* Oxford University Press, 1998

Schumaker, Randall E. Rasch Measurement: The Dichotomous Model. Ch 10 in Smith EV and Smith R.M.: Introduction to Rasch measurement : theory, models and applications. Maple Grove, Minnesota.: JAM Press 2004.

Sijtsma Klaas & Molenaar Ivo W. Introduction to Nonparametric Item Response Theory. SAGE Publications 2002, vol. 5, ch. 4.

Smith Everett V. Jr. & Smith Richard.M.: Introduction to Rasch measurement : theory, models and applications. Maple Grove, Minnesota: JAM Press 2004.

Smith Everett V. Jr. Representing Treatment Effects with Variable Maps. Ch 12 in Bezruczko Nikolaus. Rasch Measurement in Health Science, JAM Press 2005, Maple Groove Minesota.

Smith Richard M.. Rasch measurement models: Interpreting WINSTEPS/BIGSTEPS and FACETS Output. JAM Press Maple Groove, Minesota, 2003.

Snedecor G.W., Cochran W.G. Statistical methods. The Iowa State University Press 1980, ch. 10.

STATISTICA. StatSoft, Inc. (2011). STATISTICA (data analysis software system), version 10. www.statsoft.com.

StatXact. Statisticl Software for Exact Nonparametric Inference. Cytel Software, 675 Massachusetts Ave, Cambridge, USA

Wilson Mark. Constructing Measures. Lawrence Erlbaum Associates, Publishers 2005.

Wilson Mark. On chosing a model for measuring. Ch. 6 in Smith Everett V. Jr. & Smith Richard.M.: Introduction to Rasch measurement : theory, models and applications. Maple Grove, Minneasota.: JAM Press 2004.

Wright, B.D.. Misunderstanding the Rasch model.Journal of Educational Measurement,14, p. 97-116, 1997.

Appendices

Appendix A

The questionnaire used for study I

Title: The application of Item Response Theory on a teaching strategy profile questionnaire

This questionnaire is about activities that teachers undertake when they teach. Please read every statement carefully and then indicate the degree to which you use the stated activity in your teaching(circle the number). The numbers behind the statements have the following meaning:
(© Jan D. Vermunt, ICON – Graduater School of Education, Leiden University, Netherlands)

1= I do this seldom or never,	2= I do this sometimes	3= I do this regularly
4= I do this often	5= I do this almost always	

Extract from the questionnaire:

Question nr Swedish version	Question nr original version	The statement	Answers
…			1,2,3,4,5
Q2	Q31	Tell students exactly what they have to do	
Q4	Q33	Ask about the relevance of the subject matter for real life	
Q6	Q35	Make students formulate their own point of view	
Q7	Q36	Let students make connections with their own experiences	
Q9	Q38	Ask detailed questions	
Q10	Q39	Give student’s assignment of making a diagram of the subject matter	
Q13	Q42	Give exams that test factual knowledge	
Q15	Q44	Ask for similarities and differences between concepts	
Q17	Q46	Let students solve real life problems	

The scores are then summarized as indicated in the text.

<u>The questionnaire used for study II.</u>

	Items in IAPSQ Five faces on a pictogram / five point Likert scale
No and Item name	***Psychological***
	Emotional
1 FRIE16	How much do your friends love/like you?
2 MOTH17	How much does your mother love you?
3 FATH18	How much does your father love you?
4 SELF19	How much do you love/like yourself?
5 BODY20	How much do you like your body?
6 HUG21	How do you like being hugged by your mother?
7 HUG22	How do you like being hugged by your father?
8 FEEL23	How do you feel in general?
9 FEEL24	How will you feel when you become grown-up?
10 HAPP25	How often do you feel happy? *
11 ANGR26	How often do you feel angry?
12 SAD27	How often do you feel sad?
	Emotional/Cognition
13 FEEL13	How do you feel when you think of your condition?
14 MOTH14	How do you think your mother feels when she thinks of your condition?
15 FATH15	How do you think your father feels when he thinks of your condition?
16 PROBL37	Do you think of your condition?
17 MOTH38	Does your mother think of your condition?
18 FATH39	Does your father think of your condition?
19 THINK42	Do you think of your body?
	Self determination
20 DECID36	How much can you decide about your condition at home?*
21 TELL43	Do you say what you really want?*
22 DO44	Can you do as you like?*
23 GET45	Do you get what you want?*
	Social
1 SCHOO4	How do you like school?
2 TEACH5	How is your relation with the teacher?
3 FRIEND6	How is your relation with friends?
4 GYMN7	How do you like physical activity in school?
5 SHOW8	How do you like to take a shower after physical activity?
6 BREAK9	How do you like the breaks at school?
7 ACTI10	How do you like activities after school?
8 FRIEN28	How often are you together with friends? *
9 DECI29	How much can you decide when being with friends? *
10 TEAS30	Have you been teased at school?
11 BULLY31	Have you been bullied at school?
12 TOGETH32	Do you have a best friend and if you have a best friend, and how often are you together*?

* Reversed items

The questionnaire used for study III

Adolescent Adjustment Profile - revised and investigated by means of an Item Response Theory approach. (Manuscript).

Domain 1, Self rating Attention deficit

Item	Text answer = Always, Often, Sometimes, Seldom, Never
1	Easy to concentrate
2	Insecure about new tasks*
3	Tries his/her best
4	Persistent if needed
5	Interested in new tasks
6	Tries difficult tasks
7	Works quickly
8	Difficult to finish tasks*
9	Gives up at hard tasks*
10	Thorough
11	Satisfied with his/her task
12	Wants to complete tasks
13	Does homework

* Reversed items

Domain 2, Self rating Externalising behaviour

Item	Text answer = Never, Seldom, Sometimes, Often, Always
1	Bullies
2	Has stolen money
3	Has shoplifted
4	Smokes
5	Has disapproved friends
6	Destroys things voluntarily
7	Behaves thoughtlessly
8	Outburst of anger
9	Stays out past curfew
10	Plays truant from school
11	Lies to get out of trouble

Domain 3, Self rating Internalising behaviour

Item	Text answer = Never, Seldom, Sometimes, Often, Always
1	Lonely and sad
2	Does not feel liked
3	Sleeps badly/seems tired
4	Feels unhappy
5	Lacks friends
6	Does not feel good enough
7	Weak and powerless
8	Enjoys life*

* Reversed items

<u>The questionnaire used for study V and one of the three questionaires used in study IV.</u>

Items 2, 4, 5, 10 - 14 and 17 refer to study IVand V, the Depression dimension.
Items 1, 3, 6 – 9, 15, 16 and 18 refer to study V, the Mania dimension.

Affektiva mottagningen M59
Fax 585 866 30
Tel 585 866 26, 585 866 34

AS-18

Affektiv självskattningsskala

NAMN: . . . .
PERS-NR: . . . . DATUM: . . . .

Hur stora problem har du haft under den senaste veckan med:
(Ringa in det alternativ som stämmer bäst).

1 Att du varit så pratsam att andra tyckt det varit svårt att komma till tals.
2 Att du sovit mer än vanligt.
3 Att du behövt sova mindre och ändå varit pigg.
4 Att du känt hopplöshet.
5 Att du rört dig långsammare än vanligt.
6 Att du varit uppvarvad eller överaktiv.
7 Att du varit kroppsligt rastlös så att det har varit svårt att sitta stilla.
8 Att dina tankar rusat snabbt i huvudet.
9 Att du varit lättirriterad.
10 Att du känt dig nedstämd eller deprimerad.
11 Att du inte kunnat glädja dig eller intressera dig för sådant du annars tycker om.
12 Att du saknat energi.
13 Att du har haft skuldkänslor och känt dig värdelös.
14 Att dina tankar har gått trögt och långsamt.
15 Att du haft alltför hög självkänsla.
16 Att du har haft alltför stark känsla av glädje och intresse.
17 Att du haft tankar på att skada dig själv eller ta ditt liv.
18 Att du tagit risker, t ex med pengar, i trafiken eller i kontakten med andra människor.

Svarsalternativ:

Inga	Små	Måttliga	Stora	Mycket stora
0	1	2	3	4

OBS! Alla frågor gäller problem du haft under den senaste veckan.

Mats Adler, Affektiva mottagningen M59, Karolinska Universitetssjukhuset i Huddinge, 070612.

The second of the three questionnaires used for study IV.

Affektiva mottagningen M59 PHQ-9

Fax 585 866 30
Tel 585 866 26, 585 866 34

PATIENTUPPGIFTER:
NAMN: . . . .
PERS-NR: . . . .
SIGNATUR: . . . . DATUM: . . . .

Depressionsenkät (PHQ-9)

Detta frågeformulär är viktigt för att kunna ge dig bästa möjliga hälsovård. Dina svar kommer att underlätta förståelsen för problem som du kan ha.

1. Under de senaste två veckorna, hur ofta har du besvärats av något av följande problem.

a. Lite intresse eller glädje i att göra saker.
b. Känt dig nedstämd, deprimerad eller känt att framtiden ser hopplös ut.
c. Problem att somna eller att du vaknat i förtid, eller sovit för mycket.
d. Känt dig trött eller energilös.
e. Dålig aptit eller att du ätit för mycket .
f. Dålig självkänsla – eller att du känt dig misslyckad eller att du svikit dig själv eller din familj.
g. Svårigheter att koncentrera dig, till exempel när du läst tidningen eller sett på TV.
h. Att du rört dig eller talat så långsamt att andra noterat det? Eller motsatsen – att du varit så nervös eller rastlös att du rört dig mer än vanligt.
i. Tankar att det skulle vara bättre om du var död eller att du skulle skada dig på något sätt.

Svarsalternativ:

Inte alls	Flera dagar	Mer än hälften av dagarna	Nästan varje dag
0	1	2	3

2. Om du kryssat för att du haft något av dessa problem, hur stora svårigheter har dessa problem förorsakat dig på arbetet, eller för att ta hand om sysslor hemma, eller i kontakten med andra människor?

Inga svårigheter	Vissa svårigheter	Stora svårigheter	Extrema svårigheter
0	1	2	3

The third of the three questionnaires used for study IV.

Affektiva mottagningen M59
Fax 08-585 866 30

Tel 08-585 866 26/34

MADRS-M

Montgomery-Åsberg Depressionsskala

(Montgomery A. och Åsberg M. British Journal of Psychiatry 1979; 134: 382-389)

PATIENTUPPGIFTER:

SKATTAS AV: . . . .
DATUM: . . . .
NAMN: . . . .
PERS-NR: . . . .

1. SÄNKT GRUNDSTÄMNING – Avser en sänkning av det emotionella grundläget (till skillnad från situationsutlösta affekter). Omfattar dysterhet, tungsinne och nedstämdhet, som manifesterar sig i mimik, kroppshållning och rörelsemönster. Bedömningen baseras på utpräglingsgrad och avledbarhet. (Förhöjd grundstämning skattas ”0”).

0 Neutral stämningsläge.
1
2 Ser genomgående nedstämd ut, men kan tillfälligt växla till ljusare sinnesstämning.
3
4 Ser nedstämd och olycklig ut oavsett samtalsämne.
5
6 Genomgående uttryck för extrem dysterhet, tungsinne eller förtvivlad olycka.

2. NEDSTÄMDHET – Avser uppgift om sänkt grundstämning oavsett om den tar sig uttryck eller ej. Omfattar känslor av sorgsenhet, olycklighet, hopplöshet och hjälplöshet. Bedömningen baseras på intensitet, varaktighet och i vilken grad sinnesstämningen påverkas av yttre omständigheter.
(Förhöjd sinnesstämning skattas ”0”).

0 Neutralt stämningsläge. Kan känna såväl tillfällig munterhet som nedstämdhet, allt efter omständigheterna, utan övervikt för ena eller andra stämningsläget.
1
2 Övervägande upplevelser av nedstämdhet men ljusare stunder förekommer.
3
4 Genomgående nedstämdhet och dyster till sinnes. Sinnesstämningen påverkas föga av yttre omständigheter.
5
6 Genomgående upplevelser av maximal nedstämdhet.

3. ÅNGESTKÄNSLOR – Avser känslor av vag psykisk olust, inre oro eller obehaglig inre spänning, ångest, skräck eller inre oro, som kan stegras till panik. Bedömningen baseras på intensitet, frekvens, duration och ehov av hjälp. Särhålles från Nedstämdhet (2).

0 Mestadels lugn.
1
2 Tillfälliga känslor av obehaglig psykisk spänning.
3
4 Ständig känsla av inre oro, någon gång stegrad till panik, som endast med viss svårighet kan bemästras.
5
6 Långdragna panikattcker. Överväldigande känslor av skräck eller dödsångest, som ej kan bemästras på egen hand.

4. MINSKAD NATTSÖMN – Avser uppgifter om minskad sömntid eller sömndjup i förhållande till de ordinära sömnvanorna. (Ökad sömn skattas ”0”).

0 Sover som vanligt.
1
2 Måttliga insomningssvårigheter eller kortare, ytligare eller oroligare sömn än vanligt.
3
4 Minskad sömntid (minst två timmar mindre än normalt). Vaknar ofta under natten även utan yttre störningar.
5
6 Mindre än två till tre timmars nattsömn totalt.

5. MINSKAD APTIT – Avser upplevelser av att aptiten är sämre än normalt.

0 Normal eller ökad aptit.
1
2 Dålig matlust.
3
4 Aptit saknas nästan helt, maten smakar inte, måste tvinga sig att äta.
5
6 Måste övertalas att äta något överhuvudtaget. Matvägran.

6. KONCENTRATIONSSVÅRIGHETER – Avser svårigheter att samla tankarna eller koncentrera sig. Bedömningen baseras på intensitet, frekvens och i vilken mån olika aktiviteter försvåras.

0 Inga koncentrationssvårigheter.
1
2 Tillfälligt svårt att hålla tankarna samlade vid t ex läsning eller TV-tittande.
3
4 Uppenbara koncentrationssvårigheter som försvårar läsning eller samtal.
5
6 Kontinuerliga, invalidiserande koncentrationssvårigheter.

7. INITIATIVLÖSHET – Avser den subjektiva upplevelsen av initiativlöshet, känslan av att behöva övervinna ett motstånd, innan en aktivitet kan påbörjas.

0 Ingen svårighet att ta itu med nya uppgifter.
1
2 Lätta igångsättningssvårigheter.
3
4 Svårt att komma igång även med enkla rutinuppgifter, som nu kräver stor ansträngning.
5
6 Oförmögen att ta initiativ till de enklaste aktiviteter. Kan inte påbörja någon verksamhet på egen hand.

8. MINSKAT KÄNSLOMÄSSIGT ENGAGEMANG – Avser upplevelser av minskat intresse för omvärlden eller för sådana aktiviteter som vanligen bereder nöje eller glädje. Subjektiv oförmåga att reagera känslomässigt inför människor eller företeelser i omgivningen.

0 Normalt intresse för omvärlden och för andra människor.
1
2 Svårigheter att finna nöje i sådant som vanligen väcker intresse. Minskad förmåga att bli arg eller irriterad.
3
4 Ointresserad av omvärlden. Upplevelser av likgiltighet inför vänner och bekanta.
5
6 Total oförmåga att känna adekvat sorg eller vrede. Totalt eller smärtsam likgiltighet och oförmåga att uppleva känslor även för närstående.

9. DEPRESSIVT TANKEINNEHÅLL – Avser självförebråelser, självanklagelser, föreställningar om synd och skuld, mindervärdighet och ekonomisk ruin.
0 Inga pessimistiska tankar.
1
2 Fluktuerande självförebråelser och mindervärdesidéer.
3
4 Ständiga självanklagelser. Klara, men inte orimliga, tankar om synd eller skuld. Uttalat pessimistisk framtidssyn.
5
6 Absurda föreställningar om ekonomisk ruin och oförlåtliga synder. Absurda självanklagelser.

10. LIVSLEDA OCH SJÄLMORDSTANKAR – Avser upplevelser av livsleda, dödsönskningar och självmordstankar samt förberedelser för självmord. Eventuella självmordsförsök påverkar ej i sig skattningen.
0 Ordinär livslust. Inga självmordstankar.
1
2 Livsleda, men inga eller endast vaga dödsönskningar.
3
4 Självmordstankar förekommer och självmord betraktas som en tänkbar utväg, men ingen bestämd självmordsavsikt.
5
6 Uttalande avsikter att begå självmord, när tillfälle bjuds. Aktiva förberedelser för självmord.

TOTALPOÄNG av MADRS: ....

11. MINNESSTÖRNINGAR – Avser upplevelser av försämrat minne i förhållande till det för den skattade ordinära. Särhålls från Koncentrationssvårigheter (6).
0 Ingen subjektiv minnesstörning.
1
2 Tillfälliga minnesstörningar.
3
4 Besvärande till generande minnesstörningar.
5
6 Upplevelser av total oförmåga att minnas.

MINNESPOÄNG: ....

Appendix B

Some details of the methods in Step1

The scalability coefficients correspond to correlations in multiple regression analysis.

The scalability item pair i and j: $$H_{ij} = \frac{\mathrm{Cov}(Y_i, Y_j)}{\mathrm{Cov}_{max}(Y_i, Y_j)} \quad \text{for i and j} = 1,\ldots,I.\ \ i \neq j$$

Consider items with the response alternatives 1,2,3,4,5. For items i and j we get a 5x5 cross table. The calculation of the Covariance treat Y_i and Y_j as continuous variables. Under fairly general conditions, this means that the observations can be moved to get a correlation = 1 while keeping the mean and variance unchanged. However, in our case the observations are bound to the five integers. In a 5x5 cross table, (5-1)*(5-1) cells (i.e. the degrees of freedom) can be adjusted to get the maximum Covariance under the restriction of keeping the marginal totals unchanged. Thus, $H_{ij} \geq \mathrm{cor}(Y_i, Y_j)$ as $\mathrm{cor}_{max}(Y_i, Y_j) \leq 1$.

It then follows that the item scalability $$H_i = \frac{\mathrm{Cov}(Y_i, R_{(i)})}{\mathrm{Cov}_{max}(Y_i, R_{(i)})} \quad \text{for i} = 1,\ldots,I.\ ,$$

And the item set scalability $$H = \frac{\sum \mathrm{Cov}(Y_i, R_{(i)})}{\sum \mathrm{Cov}_{max}(Y_i, R_{(i)})} \quad \text{the sum over i} = 1,\ldots,I.\ ,$$

Some properties of the scalability coefficients: For all indices (i,j)

$\min_i H_{ij} \leq H_i \leq \max_j H_{ij}\,;$ $\min_i H_i \leq H \leq \max_i H_i;$ and $\min_{ij} H_{ij} \leq H \leq \max_{ij} H_{ij}$

and furthermore, if MHM holds: $0 \leq H_{ij} \leq 1$ for all i and j = 1,…,I, $i \neq j$ as well as
$0 \leq H_i \leq 1$ and : $0 \leq H \leq 1$
However, both H_{ij} and H_i , but hardly H, can be slightly below zero due to sampling fluctuations.

AISP (Automated Item Selection Procedure).
Within the '3- step strategy', this procedure is aimed for verification of the unidimensionality, as defined by the author(s) of the questionnaire. The analysis is carried out under the assumption of a MHM. This implies all $H_i > 0$ and all $H_{ij} > 0$. A first scale is constructed, based on the two items with the largest H_{ij} among all the I items. The H_i for the rest of the I-2 items are then recalculated with respect to the already chosen kernel. To qualify as a member of the scale, a constant c, is chosen as a threshold to guarantee that the H> c for the 'new scale' and thus measure a common trait with reasonable discrimination power to order the persons using the total sum score. When no item outside the new scale qualifies according to c, a construction of another scale is started. It might happen that certain items do not qualify for any scale, giving an indication that they do not yield any contribution. In practice, different c:s should be chosen as well different 'starting pairs' of items. Particularly for small samples, we have to anticipate an unstable situation with some items oscillating between scales without rejection of an intended main dimension.

Some additional remarks on the models used in Step2 and Step3.

Step2: In the following, θ (without an index) is a simplified general notation for the person measure θ_n , where n= 1,2,…, N.
In polytomous models there are two types of conditional probabilities:

a. The probability of responding in a given category.
b. The probability of responding positively rather than negatively at a given boundary between two categories. etc.

In essence, there are two types of polytomous models, referred to as Rasch vs Thurstone/Samejima models. The two concepts are illustrated in fig. B1.
In the Rasch approach, the estimation is then carried out on successive odds. This means, for each item, the odds of answering in category i+1 rather than in category i is considered. Such a dichotomization involves only the categories above and below a particular category boundary. This procedure is then continued by category i+2 vs i+1etc.
In the Samejima models, dichotomization involves all possible responses categories above and below a particular category boundary.
For further details, see Ostini & Nering, 2006, p. 11-16.

The Rasch Rating Scale Model (RSM).

RSM.

- All items are equally weighted, i.e. are set to have the same discrimination ability.
- All items share a common set of category thresholds.
- All items have the same wording for the answer categories.

An answer in category k of a K category item implies that the respondent has answered positively to categories 1 to k but negatively to category k+1. He/she has to 'pass' k thresholds.
Ex: Consider a five category item, i, with k= 0, 1, 2, 3, 4 and a respondent at level θ.
For short, set the category threshold $\delta_{ik} = (\delta_i + \tau_k)$ δ_i = location of item i, τ_κ = the distance to the category threshold k from the item location.
The probability of an answer in category 2 is modelled as follows:

cat. '0'   '1'   '2'

$$P(Y=2 \mid \theta) = \frac{\exp[\,0 + \theta-\delta_{i1} + \theta-\delta_{i2}]}{\exp[0] + \exp[\theta-\delta_{i1}] + \exp[2\theta-\delta_{i1}-\delta_{i2}] + \exp[3\theta-\delta_{i1}-\delta_{i2}-\delta_{i3}] + \exp[4\theta-\delta_{i1}-\delta_{i2}-\delta_{i3}-\delta_{i4}]} \qquad (1)$$

cat. '0'   '1'   '2'   '3'   '4'

where the numerator represents 'coming from '0', pass threshold '1' and pass threshold '2' but not '3'.
The denominator represents the sum of all categories. $\sum P(Y= k \mid \theta) =1$, k= 0,..,4.
$P(Y= 0 \mid \theta)$ for a respondent with a measure $\theta << (\delta_i + \tau_1)$, the first threshold, is dominated by exp[0]. Then $P(Y= 0 \mid \theta) = e^0/(e^0+\varepsilon) => 1$ for $\varepsilon => 0$, where ε is the denominator in (1) except exp[0]. For a respondent with a very large θ, $P(Y= 4 \mid \theta)$ is dominated by exp[4θ] in both the numerator and the denominator, which means that $P(Y= 4 \mid \theta) => 1$ for $\theta => \infty$.

A simplification is achieved if we consider odds in place of probabilities.

The odds of being in category 2 rather than in category 1 then becomes

$$\frac{P(Y= 2 \mid \theta)}{P(Y= 1 \mid \theta)} = \frac{\exp[\, 0 + \theta - \delta_{i1} + \theta - \delta_{i2}]}{\exp[\, 0 + \theta - \delta_{i1}]} = \exp[\, \theta - \delta_{i2}] = \exp[\, \theta - (\delta_i + \tau_2)] \qquad (2)$$

$P(Y= 0 \mid \theta)$ means that no threshold is passed and is set to 1/C, where C is the denominator of (1). This can be derived from the dichotomous case with two categories, “0” and “1”.
Let δ be the threshold for endorsing category “1”. Then: $P(Y=1| \theta) = \exp(\theta - \delta)/(1+\exp(\theta-\delta)$

$$P(Y=0) = 1 - P(Y=1) = 1 - \frac{\exp(\theta - \delta)}{1+\exp(\theta-\delta)} = \frac{1+\exp(\theta-\delta) - \exp(\theta - \delta)}{1+\exp(\theta-\delta)} = 1/(1+\exp(\theta-\delta))$$

The Winsteps program [Linacre J., Winsteps 3.66, 2008] produces a lot of output, which might cause worry how to start the evaluation. The program has an excellent help section, where the results and indices for evaluation are explained. However, a short presentation of some of the frequently uses expressions might help.
MNSQ: this is the chi-square statistic divided by its degrees of freedom and serves as an indicator of ‘the value’ of the item in the Rasch modelling procedure. MNSQ is the mean-square infit or outfit statistic with expectation equal to1. Values substantially less than 1 indicate dependency in the data; values substantially greater than 1 indicate noise. For our purpose I have concentrated on the infit statistic, as it is less sensible to outliers. The MNSQ should be familiar to the researchers as it corresponds to the usual measure of deviation known from the CTT.

Linacre suggests the following interpretation of the MNSQ:

Value	Meaning
>2.0	Off-variable noise is greater than useful information. Degrades measurement. Always remedy the large misfits first.
>1.5	Noticeable off-variable noise. Neither constructs nor degrades measurement
0.5 - 1.5	Productive of measurement
<0.5	Overly predictable. Misleads us into thinking we are measuring better than we really are. (Attenuation paradox.). Misfits <1.0 are only of concern when shortening a test

The reliability index (as described by [de Ayala R.J. 2009] and [Linacre J. 2008]:
High person reliability means that a person with a high estimated measure actually is on a higher level than a person with a lower estimated measure. The implication is that, with a high reliability, we are able to reasonably distinguish persons from each other on the latent scale. To achieve high reliability

we should have a sample with a large ability range and an instrument with many items. Linacre indicates ≥0.9 to be able to discriminate the sample into 4 levels. It is also essential that we have a good sample-item targeting, which means that the items should be fit for the target population.
High item reliability is a sign that we have an instrument adequate for our purpose – to catch the latent dimension of interest. We then need a wide range of the item locations and, which is one of our main difficulties and a great inconvenience, a large sample. This also means, in practise, that we are able to reasonably estimate the item locations in such a way that the structure of the items (the order the items in terms of difficulty) remains when another sample is collected or when the instrument is taken in regular use. Given a small sample, low item reliability usually means a too narrow range of item locations.
For the evaluation of a Rasch model, as carried out in the Winstep's program, the analysis of the residual variance is of general interest, similar to what should be performed according to the CTT. The following example is taken from Study III. With our small sample sizes in mind we should not look too deep in this analysis, but a few messages from the analysis are informative.

Table III.AttDef.2.1. Standardized residual variance (in Eigenvalue units)

```
                                                  Empirical
Total raw variance in observations   =         23.9 100.0%
Raw variance explained by measures   =         10.9  45.7%
  Raw variance explained by persons  =          9.8  41.0%
  Raw Variance explained by items    =          1.1   4.7%
Raw unexplained variance (total)     =         13.0  54.3%
Unexplained variance in 1st contrast =          2.5  10.4%
```

The total (Raw) variance, interpreted as the total information from the collected answers, is divided in two parts.

1. The percentage explained by the model, 45.7%
2. The percentage not explained by the model, 54.3 %, i.e. the residual information.

This residual information is expected to constitute a random set of numbers. As we are assuming just one latent trait as the cause behind the answer profiles, identified by the model, no systematic components should be inherited in the residual information. This is investigated by the 'contrasts'. Quoted from the help section in Winstep's program: 'The threat is that there is another non-Rasch explanation for the "unexplained". This is what the "contrasts" are reporting.'
The 'unexplained variance' is rescaled into eigenvalues, i.e. into a space where we have the number of orthogonal dimensions equal to the number of items. However, we should concentrate on the percentages rather than on the eigenvalues, except for the contrasts. As much as possible of the residual information is caught in the first contrast, representing a second dimension. A third dimension is represented by the second contrast and so on. A laborious analysis of the residuals is laid out by Everett V. Smith Jr [Smith Everett V. Jr. & Smith Richard.M, 2004 ch. 22], where he uses PCA and other techniques. However, in a small sample environment and with a short instrument we should rely on some few robust messages from the above sited table and combine the information with what is achieved in Step 1. A PCA, creating the contrasts, is performed on the observed residuals, i.e. as if they were metric variables. Approximate results is then achieved, but let us say that with a 1:st contrast <3 we should not be worried, "1:st contrast eigenvalue <3 → probably unidimensional" [Linacre J. Winsteps 3.66, 2008]. "However, multidimensionality only becomes a problem when data represent two or more dimensions so disparate or distinct that it is no longer clear what the Rasch model is defining …" [Smith Everett V. Jr. & Smith Richard.M, 2004, ch. 22, p.548]. The small amount of

variance explained by items, as seen in the above sited table, is usually a sign of an insufficient set of items, too few and/or too narrow in distribution.

The total raw score R_n is considered a 'sufficient statistic' for person n. This implies that ' non response' for some items can be handled within the estimation procedure. The estimation is evaluated against R_n, which is the sum score for the items actually endorsed.. This means that the logLikelihood for the person n data is expressed as

$\log L_n$ = observed sum score – expected sum score according to the model

$\log L_n = R_n - \sum_{i=1}^{L}\sum_{k=0}^{m} k\, P_{nik}$ where i= 1,…,L only for the items actually observed.

Step 3: the Graded Response Model (GRM).

GRM. Items may have their own weights but share a common set of category thresholds.
In this approach, the model is based on the probability of being in category k or *higher* .

$$P(Y\geq k \mid \theta) = \frac{\exp[\alpha_i\,(\theta - (\delta_i + \tau_k))]}{1 + \exp[\alpha_i\,(\theta - (\delta_i + \tau_k))]} \qquad P(Y\geq 0 \mid \theta) = 1 \qquad (3)$$

The probability of answering in category k then becomes:

$P(Y = k \mid \theta) = P(Y\geq k \mid \theta) - P(Y\geq k+1 \mid \theta)$ and the odds can be written

$$\frac{P(Y\geq k \mid \theta)}{P(Y< k \mid \theta)} = \frac{P(Y\geq k \mid \theta)}{1 - P(Y\geq k \mid \theta)} = \exp[\alpha_i\,(\theta - (\delta_i + \tau_k))]$$

where α_i is the item specific discrimination (weight).

This GRM is sometimes called M-GRM, the modified GRM. The basic GRM is presented with $\alpha_i\,(\theta - \delta_{ik})$ instead of $\alpha_i\,(\theta - (\delta_i + \tau_k))$, (Embretson 2000, ch. 5). The motivation for M-GRM is a suitable description of responses to Likert type scales, where the location and the thresholds are separated. This parameterisation is particular useful for questionnaires where all items have a common set of ordered response categories. The GRM can be further elaborated by introducing item specific thresholds. The term $\alpha_i\,(\theta - (\delta_i + \tau_k))$ then becomes $\alpha_i\,(\theta - (\delta_i + \tau_{ik}))$. This is an ‘unconstrained’ model in the sense that the items do not have any parameter in common. This model is required to estimate the ‘relative item information value’ of each item. This model is to complex (too many parameters) for a small sample but may serve as a check of the more parsimonious (constrained) GRM. In particular, the relative item information value, which can be approximately estimated by the constrained GRM, might be validated.
The nomenclature is not unique. Even this model is sometimes called a ‘Rating Scale Model’, as we are dealing with ‘rating scale type’ of data.

A brief description of the estimation procedures.

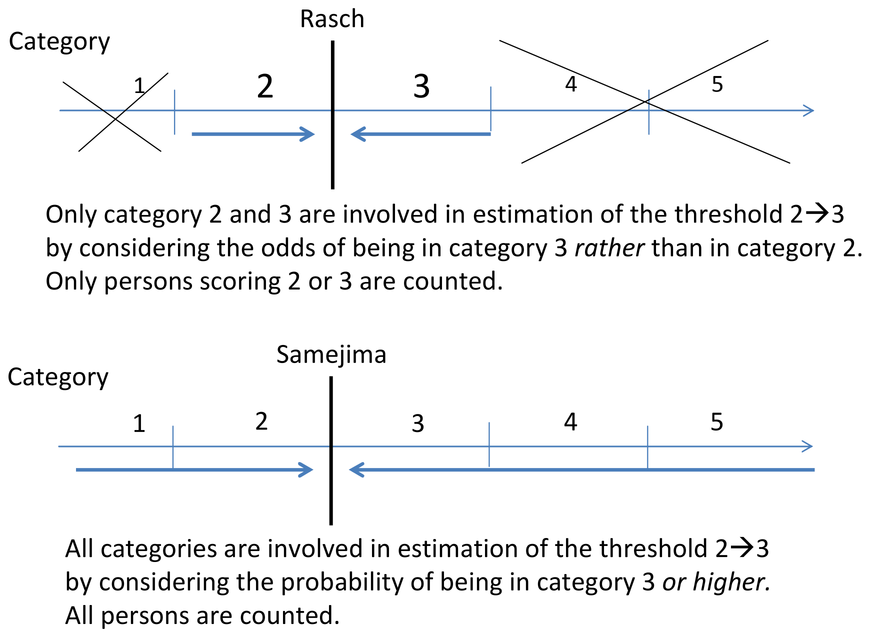
Consider a general probability function
P(a person with a level = θ answers in category j on item i) = $P(Y_{i,j}\,|\theta)$.
When there are two items, item 1 and item 2, we get $P(Y_{1,j}\,|\theta)$ and $P(Y_{2,k}\,|\theta)$ for the answer j on item1 and k on item 2. The joint probability of scoring (j,k) on these two items is
$P(Y_{1,j}\,|\theta) * P(Y_{2,k}\,|\theta)*[1 + g(_{1,2)}]$, where $g_{1,2}$ is a general residual term of dependence (covariance) between the two items, and governed by correlation(1,2), after having taken the fixed value θ into account. When there are three items, 1,2, and 3 we get
$P(Y_{1,j}|\theta) \text{ x } P(Y_{2,k}|\theta) \text{ x } P(Y_{3,l}|\theta) + g_{(1,2)} + g_{(1,3)} + g_{(2,3)} + g_{(1,2,3)}$ for the answer profile (j, k, l). In a questionnaire with 10 items let us say, we get the probability of the answer profile for person n as follows:
P(n) = P(the observed answer profile for person n) = $\prod_{i=1}^{10} P(Y_{i,j(i)}\,|\theta) + G_n$ where j(i) is the answer in category j on item i (item specific j). G_n is the general structure of covariance for a person with a location θ_n. It is reasonable to assume a simplification where this structure is the same for all individuals, thus $G_n = G$.

The persons in a sample of N individuals are assumed to be independent. The probability of the overall sample answer profile then becomes

$$P\begin{bmatrix} I_{1,1} \cdots\cdots\cdots\cdots I_{1,10} \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ I_{N,1} \cdots\cdots\cdots\cdots I_{N,10} \end{bmatrix} = \prod_{n=1}^{N} . \prod_{i=1}^{10} [P(Y_{i,j(i)}|\theta_n) + G]$$

where $I_{n,i}$ is the answer from person n on item i. The term G is governed by the correlation structure between the 10 items and will disappear in case local independence, i.e. all dependence can be conditioned on θ_n. We are now faced with a probability function for the total sample answer profile where P(n) is replaced by the IRT model under evaluation.

Fig. B.1.



Only category 2 and 3 are involved in estimation of the threshold 2→3 by considering the odds of being in category 3 *rather* than in category 2. Only persons scoring 2 or 3 are counted.

All categories are involved in estimation of the threshold 2→3 by considering the probability of being in category 3 *or higher.* All persons are counted.

Correlated residuals.

The residual correlations are calculated after the fit of a chosen model. In theory, LID means that the expected value of such a correlation should be equal to zero. For two items, i and j, the correlation is based on the covariance between the residuals (Y_{in} – E(Y_{in} | the applied model and the estimated θ_n)) and (Y_{jn} – E(Y_{jn} | the applied model and the estimated θ_n)) for the persons n=1,…,N. However, E(Y_{in}) and E(Y_{jn}) are dependent on the observed values Y_{in} and Y_{jn}, which leads to a biased expected value. As they are both included in the estimation of the model parameters, they are competing for the available information. As a consequence, the expected value is less than zero
According to (Yen, 1993), the expected value of an observed pairwise item residual correlation is approximately -1/(L-1), based on some simplified assumptions and a true local independence. L is the number of items in the questionnaire. This means a quit substantial deviance from zero in case of short questionnaires. A correction of an observed residual correlation by the term -1/(L-1), is possible.
This correction term can be viewed by the analogy with the correction N/(N-k) when calculating the residual variance in linear regression, where k is the number of estimated parameters.
However, such a correction might be motivated when there is a well-founded model, based on a sufficiently large sample. In our case, we are searching, via different types of models, the basic characteristics of a questionnaire by use of a very limited sample. This means that the correction might be questioned in our case.

Reliability.

The basic idea of the reliability: ‘The degree of stability of a respondent’s test score across independent replications of an administration of the questionnaire’.

This idea is also applicable for the items. The structure and the perception of the items should be the same at such a replication. However, at a repeated administration of a questionnaire (to the same group of respondents), it is very likely that they remember the items and their answers given at the first administration. The respondent may also change her/his attitude in the meantime as well as having been influenced by the questionnaire itself. Therefore, a measure of reliability cannot be directly calculated. It has to be estimated based on a single administration, and will, as such, be a statistical measure, not an observed characteristic. The most well-known estimator is Cronbach’s alpha, which nowadays has been proven to be ‘one of the worst estimators’ (Sijtsma 2009). There is a set of other estimators, but from our point of view, which is a small sample situation, the differences between the estimators are of less importance. In Step 1, a ‘test score reliability’ can be estimated within the Mokken scale analysis. In Step 2, both person and item reliability can be estimated, based on a basic Rasch model. For our purpose, it is sufficient to look at the estimates from the Rasch model, as they are based on the idea of a sum score as a sufficient statistic. The difference between a reliability from Step 1 and Step 2 is negligible for our purpose. Similarly, the difference between different types of estimates is of minor interest. Therefore, only the person and item reliability from Step 2 is reported and sometimes commented. The item reliability can be used as an indication whether the items are relevant and contribute to the creation of a reasonable person measure. Unfortunately, a low item reliability may as well stem from an insufficient (too small) sample.

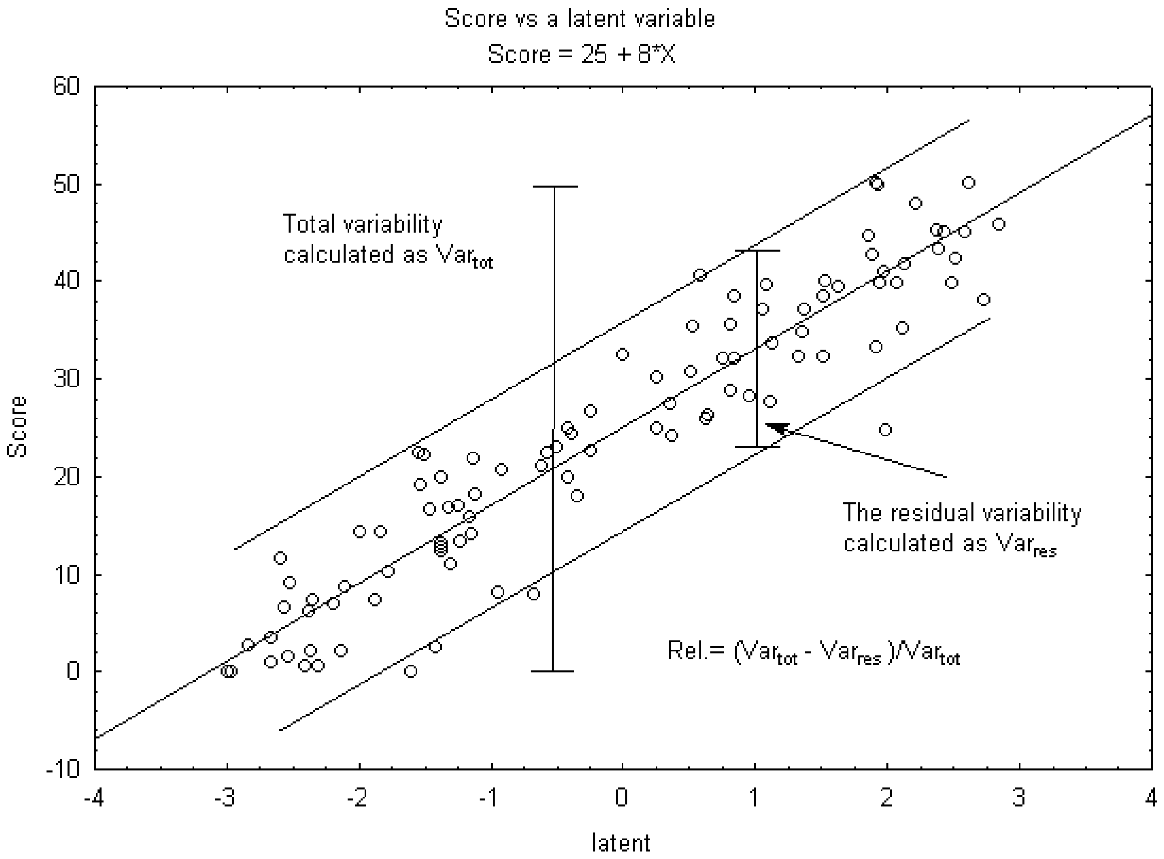
How large a person, or item, reliability is sufficient? For parametric models, the concept is based on a separation index (de Ayala R.J., 2009, Linacre J., 2008) which allows us to separate the respondents in classes – low, medium, high- (or even more classes) on the created one dimensional scale. It turns out that a person reliability of about 0.9 or more seems to be a ‘good reliability’, telling us that the questionnaire works well. According to Linacre, (Linacre J., 2008) a person reliability of about 0.9 indicates that the questionnaire can discriminate the sample into three or four ordered classes. The same idea can be applied to the items. Are the items well separated in terms of their location on the latent scale? Although very loosely founded, an item reliability ≥ 0.9 is certainly ‘an indication of a reasonable set of items as well as a not too small sample for an informative judgement’. However, in spite of large item reliability, the set of items may not be sufficient for good person estimates, further items might be needed to achieve a good questionnaire. In other words, good item reliability is necessary but not sufficient.

The calculation of the person and the item reliability is well outlined by deAyala (de Ayala R.J., 2009, Appendix E), and also described by Linacre (Linacre J., 2008). In essence, the person reliability is the relation between an adjusted person variance and the total observed variance.

$$\text{Person reliability} = \frac{\text{The total observed variance} - \text{The residual variance from the model}}{\text{The total observed variance}}$$

provided we have a plausible model. It corresponds to the correlation in a simple linear model. The idea is illustrated in fig. B.2.

Fig. B.2.



The item reliability is calculated in a manner somewhat parallel to the person reliability.

$$\text{Item rel.} = \frac{\text{The total ('observed') variance between estimated item locations} - \text{RMSE}^2 \text{ for items}}{\text{The total ('observed') variance between estimated item locations}}$$

The term RMSE^2, computed over items, is not straight forward. There are two alternative approaches, REAL or MODEL, where REAL represents a more pragmatic ‘worst case’. The REAL RMSE is the root-means-square average of the standard errors for the items, corrected by the infit statistics. A more detailed discussion of this subject can be found in the help section of Winsteps (Linacre J., 2008)

Appendix C

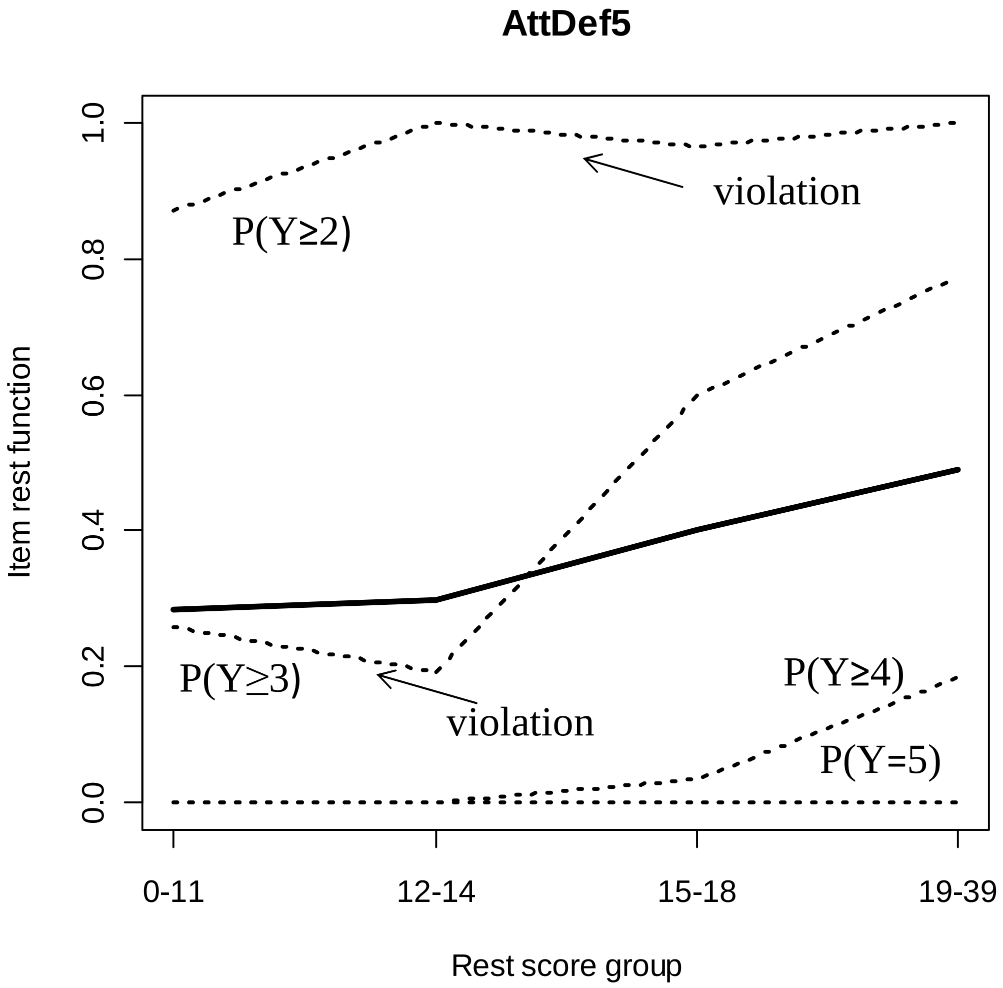
The rest score method:

The Item Response Function (IRF) for a particular item 'i' can be estimated from the data by first estimating the person measure ,θ , based on all k items on the latent scale. In a non-parametric setting, estimates of θ are not available. We have to rely on the person total score from the other k-1 items, the rest score $R_{(i)}$ with item i excluded, as a proxy for the latent scale. For a contributing item,
$P(X_i \geq c \mid R_{(i)})$ should be non- decreasing along the $R_{(i)}$ for c=(1,2,3,4,5). In a 13 item questionnaire (AttDef), with answer alternatives (1,2,3,4,5), the range of rest scores is [12, 60].
In principle, $P(X_i \geq c \mid R_{(i)})$ should be calculated for every sum score [12,60]. This is not feasible in a small study as there is a need for a reasonable number of subjects to estimate a probability by a relative frequency. Therefore, the subjects have to be grouped, where each group represents a range of scores. In such a procedure we have to state a minimum group size, which can be modified as the process goes on. An example of 'inspection of monotonicity' is presented in fig C1, where AttDef5 is investigated. $R_{(5)}$ is the sum score where item 5 is excluded.
Fig. C2 shows the comparison between AttDef3 and AttDef4. The rest score, $R_{(3,4)}$, is now the sum score based on the set of items not including items 3 and 4. AttDef3 (solid line) is more 'difficult' than AttDef4 (dashed line). This should be maintained for every rest score group and for every probability estimate in order to fulfill the 'item order invariance'. However, in a small sample we have to accept a few minor violations due to sample variations.

Note: Fig C1 and C2 are basically produced by (van der Ark, 2007) and then extended by additional information.

Investigation of monotonicity.

Fig C1

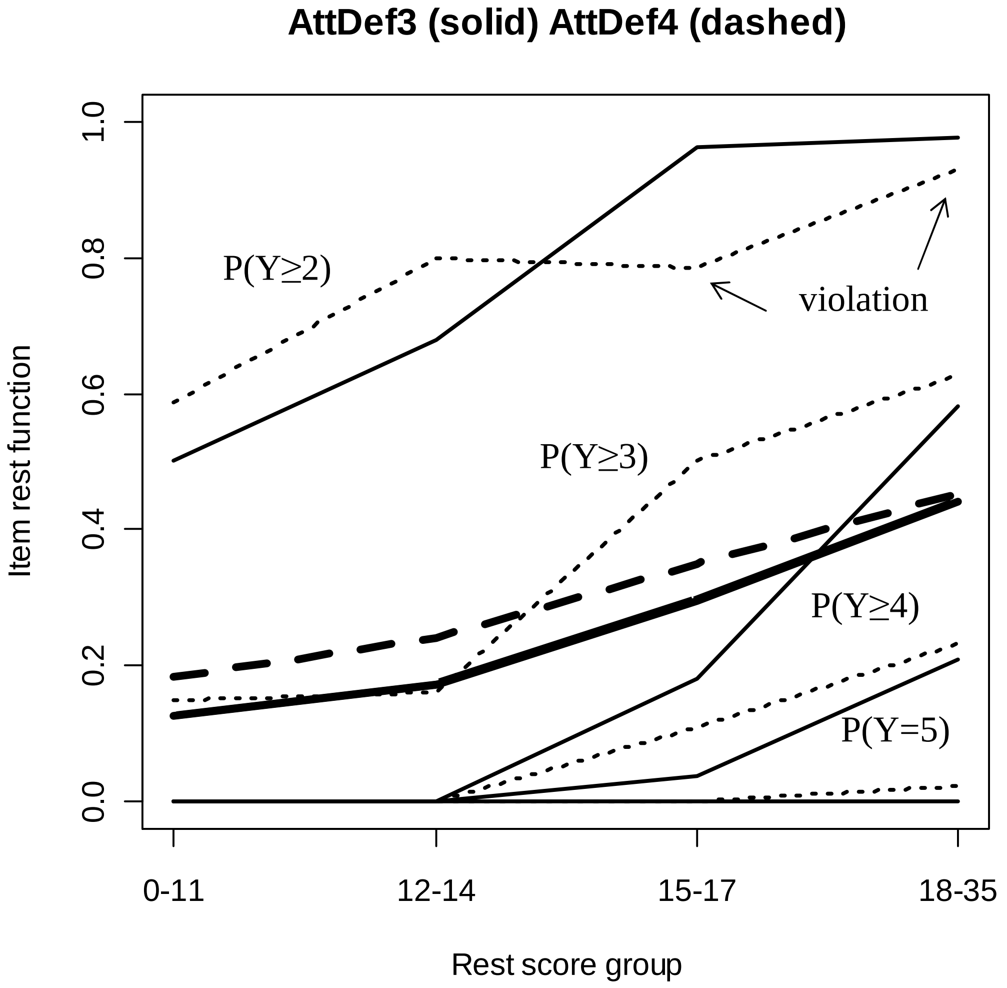


There are 4 rest score groups, $R_{(5)}$, and 2 minor violations. The solid line is the mean response function, rescaled to fit the scale [0, 1].

P(Y≥k), k= 2,…,5 should be non-decreasing when plotted against the person measure as represented by the rest score groups. Furthermore, the probability functions should not intersect, which means, for example, that P (Y≥3) should be larger than P(Y≥4) for all points of the rest score measure. No such an intersection is seen in fig. C1.

Investigation of non-intersection

Fig. C2



There are 4 rest score groups, $R_{(3,4)}$, and 2 violations. The thick lines are the mean response functions, rescaled to fit the scale [0, 1].

As AttDef3 is estimated to be more difficult than AttDef4, this should be reflected in all response functions. This means that the solid line should be inferior to the dashed line for each of the response functions. As can be seen in fig. C2, this is fulfilled for the mean response functions (the thick lines), but for the category response functions (Y≥2| AttDef3) and (Y≥2| AttDef4), there are 2 violations.