

Nielsen, T. & Kreiner, S. (2004). Modifying or replacing items: A suggestion for a strategy. Paper 1 i Nielsen, T. (2005a). Learning styles of Danish university students – Do they differ according To subject of study at the start of the first academic year? – Is there a subject specific socialization effect of one year of higher education? Development of and Research by means of The Danish Learning Styles Inventory (D-LSI) based on Sternberg's theory of mental self-government. PhD thesis, the Department of Educational Psychology, the Danish University of Education.

Paper 1

Modifying or replacing items: A suggestion for a strategy

Tine Nielsen

*Department of Educational Psychology, The Danish University of Education,
Copenhagen, Denmark*

Svend Kreiner

Department of Biostatistics, The University of Copenhagen, Copenhagen, Denmark

Paper presented at the 2nd *International Conference on Measurement in Health, Education, Psychology and Marketing: Developments in Rasch and Unfolding Models.*
Held at Murdoch University, Perth, Western Australia, January 20-22 2004.
Submitted for publication.

Abstract

Item analysis by graphical loglinear Rasch models was performed, in order to determine whether the items in the Danish edition of Wagner and Sternberg's MSG Thinking Styles Inventory "suffered" from differential item function, local dependence, or other problems.

From this work, we put forward a suggestion for a strategy for selecting items to be exchanged or modified, and deciding to which degree the modification should be, for a main or second pilot study based on the first pilot study. The strategy combines three types of information; 1) Statistical results from graphical loglinear Rasch modeling. 2) Subject matter knowledge of the theory behind the inventory scales. 3) Spontaneously written comments from participants in the pilot study. An example of the application of the strategy to a single scale is presented.

Application of the suggested strategy to the 13 subscales of the Danish edition of the MSG Thinking Styles Inventory resulted in considerable improvement of 10 scales in the sense that the revised scales had fewer misfitting items, and fewer differential item functioning and local dependence problems. One scale did not improve in any noteworthy way. Two scales declined slightly in quality; one in the sense that the revised scale had an additional local dependence problem, the other in the sense that one of the items in the revised scale turned out to be biased relative to gender.

1. Introduction

The family of Rasch models define a set of item response theory (IRT) models useful for evaluating the qualities of summated scales summarizing responses to a number of discrete items, because these models encapsulate a number of desirable and even optimal theoretical and technical properties of scales: construct validity, sufficiency and objectivity.

Construct validity is the vital requirement of summary scales implying that the indirect measurement provided by the scale measures nothing but an unidimensional latent variable corresponding to a specific theoretical construct. Construct validity implies according to Rosenbaum (1989) that items are locally independent, that item characteristic curves are monotonously increasing functions of the latent traits or properties measured by the scale, and that there is no differential item functioning in the sense that items are conditionally independent of exogenous variables given the latent variable. Given these requirements it follows that the summary scale will be positively associated with all variables known a priori to be positively related to the latent variable. Construct validity as defined by Rosenbaum therefore implies criterion validity, as it is usually defined in psychometrics. Subordinate to the requirement of construct validity one may raise a number of different technical questions relating to the quality of measurements and to the efficiency of the way the information in item responses have been used in the summary scale. Reliability is but one such requirement. Statistical sufficiency of the summated scale is another and more fundamental requirement in the sense that sufficiency implies that it is impossible to get more information out of the given set

of items than the intrinsic information in the total score. Increasing the reliability by using a different way to summarize the information in the items is therefore unrealistic.

Andersen (1977) shows that the so-called Rasch model (Rasch, 1960) for dichotomous items is the only IRT model for a construct valid set of items satisfying the requirement of sufficiency. From this follows another important technical property of the scale. Rasch (1968) shows that the Rasch model for dichotomous items is characterized by so-called specific objectivity, implying that there is a guarantee against systematic measurement errors when item responses meet the requirements of Rasch models. Similar results do not exist for polytomous (ordinal) items. The generalized Rasch model of Andersen (1977) otherwise known as the partial credit model of Masters (1982) is however the only known model for this kind of items also satisfying the requirements of sufficiency and objectivity.

From these arguments it follows that one may evaluate the qualities of a summated scale by an analysis of the extent to which the items meet the requirements of Rasch models. The requirement of no DIF is an inherent part of construct validity as defined by Rosenbaum (1989, formulas (4), (11) and (12)), but it is only implicitly assumed that it is required for all variables defining different subpopulations of subjects. Kreiner and Christensen (2002) suggest that it is useful to embed the Rasch model as a measurement model in a larger structural framework consisting of all variables that are important in connection with the specific intended application of the summated scale. They propose a so-called graphical model (Whittaker, 1990; Lauritzen, 1996) as a useful model for this set-up, because most of the properties of construct validity and Rasch models may be stated in terms of conditional independence. Kreiner and Christensen (2002) refer to this type of model as a graphical Rasch model. One advantage of graphical Rasch models is that it becomes explicitly clear which exogenous variables are considered as potential and important sources of differential item functioning. Other advantages are pursued in Kreiner and Bang Christensen (2004) and Christensen et.al. (2004).

Established paradigms for item analysis by Rasch models (Glas and Verhelst, 1994) often focus on person and item fit statistics implying that departures from the Rasch model should be remedied by elimination of less than perfect items not fitting the models. While this in principle will lead to valid and objective measurements, it should nevertheless be recognized that the price of objectivity may be a serious loss of reliability. This paper is concerned with two alternatives to item-elimination. The first has been explored by Kreiner and Christensen (2002) and Kreiner and Bang Christensen (2004) modeling departures from the graphical Rasch models by so-called graphical loglinear Rasch models (GLLRM). These models attempt to explain the departures from the Rasch model in terms of uniform DIF and uniform LD. If the modeling attempt is successful it follows that the summated score is still sufficient and it may be argued that measurements can be regarded as essentially objective. Some reliability is still lost, however, compared to the ideal situation with locally independent items, and the evaluation of the level on the latent scale becomes somewhat more complex than for a set of unbiased items. The other approach therefore uses the results from the item

analysis by graphical loglinear Rasch models as one of the starting points of a process aimed at rewriting or replacement of items, in order to retain both simplicity and reliability.

The purpose of this paper is to illustrate the second approach by examples from a pilot study of learning styles of university students. Section 1 introduces the background of the study. Section 2 provides some details on the study. The item analysis by graphical loglinear Rasch models is briefly described in Section 3. The details of the strategy combining results from the analysis by GLLRMs with subject matter considerations and an example of application to a scale are given in Sections 4 and 5. Finally, in section 6 and 7, the results are presented and discussed.

2. Learning styles of university students

The overall aim of the research project *Learning styles of university students* is to study the learning styles of two groups of university students (bachelor students of sociology and master students of educational psychology), and to examine the change in learning styles over the first year of their studies. The main purposes of the study is to determine which factors effect the students learning styles at the two points in time and to test hypotheses concerning the association between learning styles and the choice of major, the association between the major studied for a year and the learning styles, and the association between personality dimensions and the changes in learning style from study start to one year later.

The aim of the pilot study is to investigate the construct validity and psychometric properties of Danish editions of the MSG Thinking Style Inventory (Sternberg, 1997) together with an additional style construct, a new measure of flexibility and the Danish edition of the short version NEO Five-Factor Inventory (Costa & McCrae, 1992).

This paper will only be concerned with the development of a strategy for deciding which items should be replaced respectively modified, based on results of item analysis by graphical loglinear Rasch models (Kreiner and Christensen, 2002), results of subject matter analysis and comments from participants, and application of this strategy to the first Danish edition of the MSG Thinking Style Inventory.

The MSG Thinking Styles Inventory

The instrument of measurement used in this study is a Danish edition of the MSG Thinking Styles Inventory (Sternberg, 1997) developed by the first author of this paper.

The learning style construct used in the research project and this particular study is derived from the thinking style construct proposed by Robert J. Sternberg in his Theory of mental self-government (Sternberg, 1988, 1997). Within this theoretical framework, thinking style is defined as a profile of styles describing a person's preferred ways of thinking in specific contexts. The learning style construct used is defined as: A profile of styles describing the individual's *preferred* ways of thinking in the university learning context. These preferred ways of thinking represent different ways of perceiving and handling different types of problems in the learning context.

The theory of mental self-government is based on the idea that the forms of government we see in the world represent alternative ways of organizing our thinking. The learning style profile is therefore made up by 13 learning styles organized in five categories signifying the functions, forms, levels, scopes and leanings of mental self-government, as shown in Table 2.1 (Sternberg, 1997).

Table 2.1. *The 13 learning styles in the Danish Learning Styles Inventories*

Functions	Levels
Legislative	Global
Executive	Local
Judicial	Scopes
Forms	Internal
Monarchic	External
Hierarchic	Leanings
Oligarchic	Conservative
Anarchic	Liberal

The 13 styles are *always* represented in the learning style profile, but to varying degrees. For example, the level of mental self-government is not a question of preference for either the global ways of thinking or the local ways of thinking. Rather it is a question of degrees of preference for the global *and* the local ways of thinking. The preference to certain ways of thinking can therefore, and will usually, be of differing strength, but never absent.

The original MSG Thinking Styles Inventory is a self-report inventory and consists of subscales measuring 13 thinking styles (Sternberg, 1988, 1997), as illustrated in table 2.1. Each subscale consists of 8 items in the form of statements – like the example item shown below – to be evaluated with regard to how well it describes the person within a certain context, using a polytomous scale with 7 categories, with 1 corresponding to *not at all well* and 7 corresponding to *extremely well*.

“I prefer tasks or problems where I can grade the design or methods of others.” (Sternberg, 1997 p. 37).

The MSG Thinking Styles Inventory was translated into Danish in the following manner: 1) A translation into Danish was done by the first author of this paper as the subject matter expert. 2) A panel of 12 adults of different educational levels and professions answered and commented the inventory with regard to anything they found puzzling, difficult to understand or answer etc. 3) The first author refined the translation. 4) The English edition was compared to the Danish edition by a bilingual secretary as the language expert, and suggestions for modifications were made. 5) The modification suggestions were discussed and decided upon by three subject matter experts; the first author and her two supervisors.

The resulting Danish edition of the MSG Thinking Styles Inventory consists of the same 13 subscales as the original instrument. The changes made to the original instru-

ment in the Danish edition are: Addition of one subscale¹. Items were not divided into sections according to subscales. And in the written instructions the participants were instructed to rate the items according to how well they described themselves in learning situations within the context of their university study, thereby measuring more specifically learning style in the meaning “thinking style in learning situations”.

Samples and administrations

This study includes two separate data samples derived from the pilot study sample and the first of the two main study samples – hereafter pilot sample and follow up sample.

The pilot sample consists of 283 university students from two Danish universities, collected in 10 separate sessions within a period of 9 weeks during the months of February through April 2003, and 114 variables:

- 7 background variables: Gender, age, university, major subject of study, study level, number of prior completed educations/studies, number of prior incomplete educations/studies, and number of reason for choice of study subject.
- 3 design variables: An “order” variable stating whether the participants answered an “ordered by style” design or a random design of the Danish edition of the MSG Thinking Styles Inventory. A “method” variable stating whether the data collection was done integrated with the class subject, during class without subject integration, after a class, or handed out without instruction. An “information” variable stating whether or not the participants had been informed of the data collection prior to the class where it took place.
- 104 learning style variables making up the 13 scales with 8 items in each.

The follow up sample consists of 162 university students from two Danish universities, collected in 3 separate sessions within a period of 10 days in September 2003, and 114 variables:

- 9 background variables: Gender, age, university, major subject of study, study level, number of prior completed educations/studies, number of prior incomplete educations/studies, a variable “starter” stating whether or not they are university starters or have studied at university level earlier.
- 1 design variable: “Data collection number” giving the number in the row of data collections.
- 104 learning style variables making up the 13 scales with 8 items in each.

The distributions of the background and design variables of the two samples are shown in Tables 2.2 and 2.3.

¹ The additional subscale is an 8-item scale measuring a proposed additional learning style construct within the form-category: The Democratic style (Nielsen, Kreiner & Styles, 2005). This scale is not included in this study.

Table 2.2. *Pilot and follow up sample distributions of background variables.*

Variable	Categories/statistics	Pilot sample	Follow up sample
		N = 283 Percent	N = 162 Percent
University	Danish University of Education (DPU)	20.5	41.6
	University of Copenhagen (KU)	78.4	52.2
	Other ²	1.1	6.2
Total		100.0	100.0
Major³	Educational psychology	20.7	40.4
	Sociology	38.6	46.0
	Public health	18.2	
	Psychology	18.9	
	Other	3.8	13.7
Total		98.9	100.0
Study level	Bachelor	73.8	46.0
	<i>Kandidat</i> ⁴	26.2	40.4
	Uncertain		13.7
Total		100.0	100.0
Starter	Yes		59.6
	No		40.4
Total			100.0
# completed educations	0	67.1	55.9
	1	16.3	26.7
	2	10.6	14.3
	3	4.2	3.1
	more than 3	1.8	
Total		100.0	100.0
# incomplete educations	0	67.1	77.6
	1	27.9	21.7
	2	4.9	0.6
Total		100.0	100.0
Gender	Male	19.8	34.2
	Female	80.2	65.8
Total		100.0	100.0
Age	Range	18 – 58	19 – 68
	Mean	28.06	29.44
	Standard deviation	9.12	9.97
# reasons	Range		1-21
	Mean		9.20
	Standard deviation		4.23

² Students from other universities taking single classes at the universities participating in the study.

³ In the pilot sample, psychology majors are all bachelor level students and educational psychology majors are all *kandidat* level students, public health and sociology majors include students at both study levels. In the follow up sample, educational psychology majors are all *kandidat* level students and the sociology majors are all bachelor level students.

⁴ The *kandidate* degree is a two-year subject specific study placed after the three years of bachelor level studies, and the Germanic equivalent of the Anglo-Saxon master degree.

Table 2.3. *Pilot and follow up sample distributions of design variables.*

Variable	Categories	Pilot sample N = 283 percent	Follow up sample N = 162 percent
Order	Random	49.8	
	Ordered by style	51.2	100.0
Total		100.0	100.0
Method	In class integrated with subject	45.6	100.0
	In class without subject integration	34.3	
	After class	16.3	
	Without instruction	3.9	
Total		100.0	100.0
Information	Yes	34.3	100.0
	No	65.7	
Total		100.0	100.0
Data collection #	1		56.5
	2		19.9
	3		23.6
	Total		100.0

In the 10 data collection sessions making up the pilot sample, inventories were administered in connection with classes with the permission of individual teachers. The purpose of the pilot and the main study and the different inventories were explained briefly to the students. The definition of learning styles as used in the study was explained and an opportunity for choosing not to participate was given before the actual instructions for answering the inventories were administered. The inventories were administered anonymously.

In the three data collections making up the follow up sample, the participants had, prior to the data collection sessions, received a written invitation to participate in the main study and information on the study's aims, methods, perspectives, etc. The inventories were administered integrated in class subjects in different ways with the permission of teachers and subject coordinators. Permission to conduct the research had been granted by department boards prior to approaching students and teachers. The overall results of the pilot study and purpose of the main study and the different inventories were explained briefly to the students. The definition of learning styles as used in the study was explained and an opportunity for choosing not to participate was given before the actual instructions for answering the inventories were administered. For purposes of pairing data from the two collections of the main study and participant's possibility of obtaining their personal learning style profiles⁵, the inventories were marked with identification numbers.

3. Measurement models

The family of Rasch models for dichotomous and polytomous items (Fisher & Molenaar, 1995) is a formal representation of optimal measurements satisfying requirements of construct validity (Rosenbaum, 1989) and providing ideal measurement qualities in terms of objectivity and sufficiency.

⁵ In order not only to receive information from the participants, but also give information to them, the study includes an opportunity for the participants to obtain their personal learning style profiles on request, when both the data collections of the main study has been completed.

One of the requirements of construct validity as defined by Rosenbaum (1989) is that items are conditionally independent of both criterion and other variables given the latent variables that the items purportedly measure. This condition may be interpreted as a requirement of no differential item functioning (DIF). The requirement that there is no DIF relative to any exogenous variables appears at one and the same time to be both unnecessarily vague and unnecessarily strict. The family of graphical Rasch models (GRM) defined by Kreiner and Christensen (2002) embeds the Rasch model in a structural framework defined by a so-called graphical model containing a number of exogenous variables in addition to the items and the latent variable and restricts the assumption of no DIF to the specific set of exogenous variables included in the model. The question of the validity of the measurement therefore becomes a question of validity within this specific framework.

Apart from the restriction on the requirements of no DIF, the GRM retains and even strengthens all the fundamental properties of construct valid ideal measurements. The structure is fundamentally unidimensional, items are still assumed to be locally independent and monotonously related to the latent variable, the raw score is sufficient and items are conditionally independent of exogenous variables given both the latent variable and the total score. We refer to Kreiner and Christensen (2002) and Kreiner and Bang Christensen (2004) for a comprehensive discussion of these models.

The most important purpose of item analysis by Rasch models is consequently to test the fit of the models to item responses and to identify the points of departure where item responses deviate from the requirements of Rasch models. If the Rasch model does not fit the distribution of item responses, the result is a catalogue of less than ideal measurement properties and a set of suggestions for improvement of items. Conventional item analyses by Rasch models often give special attention to different kinds of item fit statistics assuming implicitly that items characterized by misfits are flawed items that should be eliminated. The underlying idea behind the analyses presented in this paper is contrary to this point of view that departure from Rasch models does not automatically imply that items are formally flawed and that the Rasch model may be inadequate even when items are fundamentally sound. One such situation appears when a so-called graphical loglinear Rasch model (GLLRM) fits the data. A GLLRM extends the GRM by adding interaction between some items and some exogenous variables and/or interaction between some items. These interactions entail DIF and/or LD. DIF and LD are, however, assumed to be uniform in the sense that the size of interaction does not depend on the latent variable. Under this assumption it follows that the total score is still a sufficient statistic providing essential objective measurements. Scores will be less reliable than in the ideal situation with locally independent items, and comparisons of scores from different groups confounded by DIF are an inconvenient, but in no way impossible, exercise in test equating.

There are two underlying assumptions of GLLRMs: First, that items are fundamentally sound, but that inherent properties of items lead to departures from the formal requirements of Rasch models. Second, that these item properties do not depend on the latent

variable. The interaction parameters relating to uniform DIF and/or LD may be regarded as additional item parameters that may be estimated independently of the latent variable in fundamentally the same way as the ordinary item parameters of Rasch models. Whether or not one decides to keep these items becomes a question of convenience and not a fundamental question of measurement validity. If one retains the items, one must accept some inconvenience due to DIF and/or reduced reliability due to LD. If this is not acceptable, then the information that items are basically sound, although infected with problems that have nothing to do with what the items purport to measure, will be useful information that may imply some rewriting rather than forthright elimination of items.

Conditional inference for conventional Rasch models (Andersen, 1970, 1971, 1972, 1973a, 1973b, 1994) applies in exactly the same way for the GLLRM. Details have been worked out by Kelderman (1984, 1989), Kelderman and Rijkens (1994), and Kelderman and Steen (1988). Further developments connecting so-called Mantel-Haenszel analyses of DIF and LD with analysis by GLLRM may be found in Kreiner and Christensen (2002) and Kreiner and Bang Christensen (2004). When items are polytomous rather than dichotomous we use Partial gamma coefficients (Davis, 1967; Agresti, 1985) for analyses of DIF and LD instead of the conventional Mantel-Haenszel estimates of odds-ratios.

Item analysis

The graphical loglinear Rasch analysis performed on the two samples include all the variables described earlier in a sample specific recursive order. For some variables only the major categories are entered and some variables are categorized in the respective analyses.

In the analyses of the pilot samples the following variables are entered into the graphical loglinear Rasch analyses as exogenous: University (KU, DPU), major (educational psychology, psychology, public health, sociology), study level (bachelor, *kandidat*), # completed educations (0, 1, more than 1), # incomplete educations (0, 1, 2), gender (male, female), age (18-21, 22-25, 26-35, 36-58), order (random, ordered), method (in class integrated with subject, in class with outsubject integration, after class), and information (yes, no). The 8 variables making up the specific subscale are entered as polytomous items with 7 categories. The pilot sample variables are analyzed with the recursive structure shown in Figure 3.1.

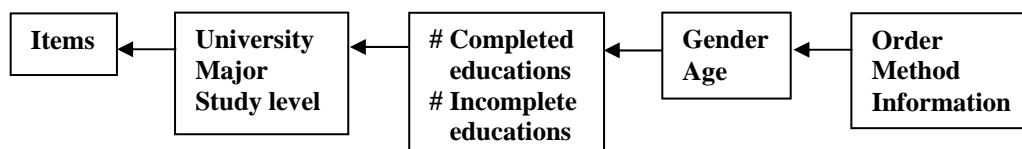


Figure 3.1. The recursive structure of the graphical loglinear Rasch analysis performed on the pilot sample.

In the analyses of the follow up sample the following variables are entered into the graphical loglinear Rasch analyses as exogenous: University (KU, DPU, other), major

(educational psychology, sociology), study level (bachelor, uncertain, *kandidat*), starter (yes, no) # completed educations (0, 1, 2, 3), # incomplete educations (0, 1, 2), # reasons for subject choice (1-5, 6-10, 11-15, 16-21), gender (male, female), age (19-21, 22-25, 26-35, 36-68), and data collection number (1, 2, 3). The 8 variables making up the specific subscale are entered as polytomous items with 7 categories. The follow up sample variables are analyzed with the recursive structure shown in Figure 3.2.

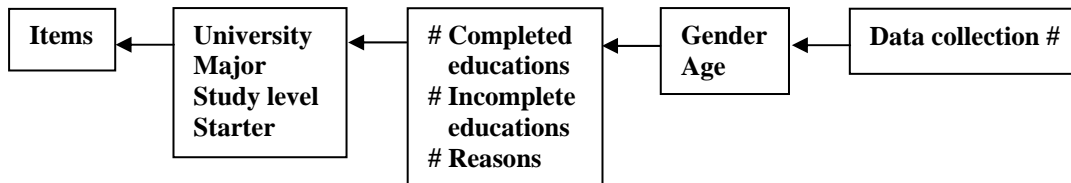


Figure 3.2. The recursive structure of the graphical loglinear Rasch analysis performed on the follow up sample.

4. A combination strategy

Graphical loglinear Rasch models (Kreiner & Christensen, 2002; Kreiner & Bang Christensen, 2004) enable us to model departures from the Rasch model in terms of uniform DIF and uniform LD. If graphical loglinear Rasch modeling is not possible with the complete item scale – that is if there is non-uniform DIF or LD or if there are more fundamental problems with items – then the options are to eliminate items in order to achieve a graphical loglinear Rasch model (or a pure Rasch model) with fewer items. Once a graphical loglinear Rasch model is found, this can serve as one of two bases on which the decision to rewrite (in the sense modify) or replace items in order to improve the scale, while retaining both simplicity and reliability, can be made. The other basis for these decisions is a subject matter analysis where the items are analyzed within the particular theoretical framework of the study/scale in order to determine problems with or between items on the theoretical level. The suggested modification strategy is an attempt to systematize the relationship between the results of the statistical analysis by graphical loglinear Rasch models, the subject matter analysis, and the decision to rewrite or replace items into an applicable strategy for scale improvement. Underlying the ideas behind the strategy are two overriding concerns. The first is that evidence from subject matter analysis suggesting theoretical errors during item construction should always lead to elimination of items. The second is that a very complicated loglinear structure with extensive DIF and LD among items always indicates that there must be serious problems with the fundamental ideas which suggested a construct valid score with no DIF and no LD. The strategy therefore basically addresses the situation where the departures from the Rasch model are limited and where the face validity of items is plausible even after careful rethinking of the theoretical foundations in the light of the results of the initial analysis.

The suggested modification strategy

The modification strategy we suggest is illustrated in figure 4.1. On the left hand side of figure 4.1, the results of the *loglinear Rasch analyses* are placed. From the top down, this part of the model also illustrates the actual process of analysis, starting with the full

8-item scale and, if modeling is not possible, moving on to scales with a lesser number of items, until modeling of one or more loglinear Rasch models is possible, and continuing to exclude items until a pure Rasch model can be modeled. The resulting loglinear Rasch models could include uniform DIF and/or uniform LD of varying severity (signified by the thin black arrows). The results of the loglinear Rasch analyses point out which items should be paid special attention in the subject matter analysis, the results of which are seen on the right hand side of Figure 4.1. The subject matter analyses of the items of the first Danish edition of the MSG Thinking Styles Inventory disclosed a number of problems, all of which could be classified as belonging to one of four types, presented in order of increasing severity: Type A problems, which are minor wording errors. Type B problems, which is when two or more items are partly similar in wording even though measuring separate style characteristics. Type C problems, which is when two or more items are measuring the same style characteristic. And type D problems, which is when an item is considered an unclear measure of style characteristic, either in the sense that we were not able to pinpoint the characteristic being measured, or in the sense that the item measured a characteristic of another style than intended. At the bottom of Figure 4.1 the participants' comments are placed. The modificational actions placed in the middle of Figure 4.1 (the circles) are reached either by evaluation of participant's comments, a single analysis result, or by combinations of results from the two types of analyses (signified by the thick black arrows and the circles with the &-character). The possible modificational actions are: No change, modification of item(s), and exchange of item(s).

The loglinear Rasch analysis can, as illustrated, result in one or a range of different models for the scale. The best result possible is a pure Rasch model with the complete number of items included, in which case the construct validity for the particular scale is considered good and no modification is needed. Unfortunately, this is often not achieved for all 8 items, but only with a lesser number, in which case there might be a preceding result in the form of one or more Loglinear Rasch models for 8 or less items.

The model we seek to improve by applying the strategy is the loglinear Rasch model for the highest number of items. The loglinear Rasch models for a lower number of items than the one being improved act as support-tools when working with the actual modification, in the sense that they might support (or not) which items are to be examined closely in the subject matter analysis. Returning to the loglinear Rasch model for the highest number of items, this can include uniform LD and/or uniform DIF of differing severity. Severity of LD and DIF is, in this study, determined from the partial γ -coefficients measuring the conditional association between pairs of items or items and exogenous variables. We have accordingly divided LD problems into three groups of severity: *Severe problems* with items are likely to be present when $0,2 \leq |\gamma|$ – this should always lead to scrutinization of items in subject matter analysis. *Moderate problems* with items are likely to be present when $0,1 \leq |\gamma| < 0,2$ – this should always lead to scrutinization of items in subject matter analysis. *Negligible problems* with items are likely to be present when $\gamma < |0,1|$. These cut points are of course arbitrary and can be determined by the individual researcher. Our choice of cut points, however reflect an additional finding of this study; that there is a high degree of agreement be-

tween the types of problems found in the loglinear Rasch analysis and the subject matter analysis. The same classification of severity can be applied to DIF problems⁶. This type of classification of severity of DIF should, however, not stand alone. It should be used in conjunction with a qualitative evaluation of the meaning, importance and practical implications of the DIF, and a statistical evaluation of the impact of DIF with respect to score equating. Differential item functioning will have serious consequences for objectivity if it is not dealt with in the correct manner and will always be a cause of inconvenience, even when it is recognized. Elimination of biased items will often be the preferable solution, even though inclusion of uniformly biased items in a scale to our mind is acceptable if proper care is taken to take DIF into consideration when scores are calculated.

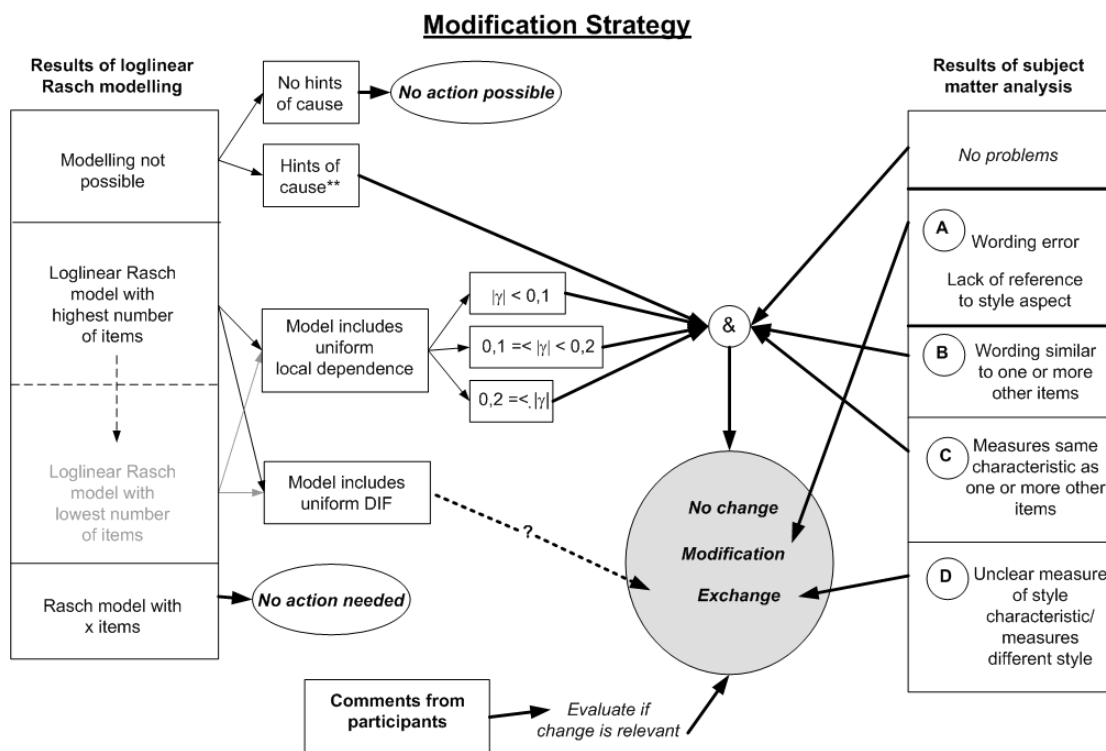


Figure 4.1. The suggested modification strategy.

⁶ This study, however, revealed only few of these problems for which reason a classification could not be investigated properly.

Finally, analyses where loglinear Rasch modeling is not possible are of course also relevant for the modification strategy. The reason for the failure of the loglinear Rasch model may be that DIF and/or LD is non-uniform, that items are differentially discriminating, or that one or more items are seriously defective. The worst possible case – lack of construct validity – may, of course, also be a possibility. The analysis will provide some hints as to the cause of the lack of possibility of modeling, but the reason for the failure of the loglinear Rasch model can only be determined by examination of items in the subject matter analysis. The results of analyses by loglinear Rasch models point out which items should be of special interest in the subject matter analysis. It should, however, be stressed that all items should be subjected to subject matter analysis if the loglinear Rasch analyses show any problems with a scale⁷.

The subject matter analysis can, as illustrated in Figure 4.1, show that there are no problems with the items of the scale, in which case no changes to the scale are needed. Furthermore, the subject matter analyses can pinpoint a range of scale problems, which can be grouped in two. *Less severe problems* (type A), such as wording errors which might have arisen from the translation or are simply the result of inadequate proof-reading, or in this particular study a lack of explicit reference to the style aspect in the statements. *Severe problems* of different degrees, such as: (type B) item(s) that is/are similar in wording to one or more other items of the scale – being the least severe of the three problems. (Type C) item(s) that measure(s) the same style characteristic as one or more other items of the scale. And (type D) item is an unclear measure of style characteristic – being the most severe of the three problems.

The participants' comments should always be evaluated and can within the strategy lead to different modificational actions depending on the outcome of the evaluation.

In summation: When applying the modification strategy in Figure 4.1, the starting point is the results of the loglinear Rasch analyses on the left side of the figure. The next step is the subject matter analysis and its results on the right side of the figure. The last step is deciding on which modificational action is appropriate, based on results of one or both analyses performed on the scale. Before going into detail with the specific relationships between analyses results and choice of modificational actions, as summarized in Table 4.1, it should be noted that modificational action is only warranted when the items pointed to as problematic by the two analyses are identical – for example a log-linear Rasch model with high uniform LD between items F and G and the subject matter analysis showing that item F and G mutually measure the same characteristic (type C problem) would result in modificational action, but a loglinear Rasch model with high uniform LD between items F and G and the subject matter analysis showing that items F and G individually measure the same characteristics as other items in the scale, but *not mutually* would *not* result in modificational action.

⁷ In this connection it should be noted that the loglinear Rasch analysis and the subject matter analysis were performed independent of each other.

Table 4.1. *Choice of modificational action based on analyses results.*

Loglinear Rasch analyses results	Subject matter analyses results				
	No problem	Type B wording similar	Type C measure same characteristic	Type D unclear measure	Type A wording error
RM	No change	No change	No change	Exchange single item	Single item modification
LLRM with <i>low</i> uniform LD	No change	No change	No change	Exchange single item *****	Single item modification
LLRM with <i>moderate</i> uniform LD	No change	Mutual modification of items	Exchange single item	Exchange single item *****	Single item modification
LLRM with <i>high</i> uniform LD	No change	Mutual modification of items	Exchange single item	Exchange single item *****	Single item modification
LLRM with uniform DIF	No change	Exchange single item **	Exchange single item **	Exchange single item	Single item modification
Modeling not possible	No change	Exchange single item (consider additional problematic items for modification)	Exchange single item (consider additional problematic items for modification)	Exchange single item	Single item modification

Note. The occurrence of combinations of results in this study can be seen in table A.2 in the appendix.

As illustrated in Table 4.1, the choice of modificational action depends on combinations of results of the loglinear Rasch analyses and the subject matter analyses in a general pattern where increasing severity of the problems detected in the two types of analysis warrant increasing degrees of changes made to items. However, due to the difference in nature of the two types of analyses, some combinations are logically incompatible, which is why the choice of modificational action in these cases must depend on the results of a single analysis. Furthermore, it is possible to get result combinations, which are incompatible in the sense that the nature of one problem is such that it warrants a certain type of modificational action regardless of the nature of the result of the other analysis, why the choice of modificational action in these cases also must depend on the results of a single analysis.

The cases where the choice of modificational action depends solely on results of one analysis are:

- When the subject matter analysis reveals the presence of less severe (type A) problems with items, this should always lead to a modification of the relevant item(s) to correct the problem(s) regardless of the results of the loglinear Rasch analysis, since these problems effect the reliability of the scale and can contribute to generating spurious problems with other items. These problems should therefore preferably be eliminated before data collection takes place – hence the grey text and the position outside the severity continuum. If these problems occur in combination with the more severe problems (type B and C) that can be found in the subject matter analysis, we recommend that the type A problem is corrected and the other problems are ignored, that is if the potential extra phase of data collection and analyses that could result from any persisting type B and C problems is a possibility.
- When the subject matter analysis does *not* reveal any problems with items, the modificational action *no change* is recommended regardless of the problems pointed out by the loglinear Rasch analysis. The recommendation is made because the problems revealed by the statistical analyses could be caused by statistical type I errors, when they are not reflected in the subject matter analysis. And even if the problems pointed out by the statistical analysis are not due to error, it is not possible to make any changes to the involved items, since we do not have information about the nature of the problems from the subject matter analysis.
- When the subject matter analysis reveals the most severe (type D) problems with an item, we recommend that the item be exchanged with an item measuring another (in the sense precise) characteristic of the style regardless of the results of the loglinear Rasch analysis, due to the effect the inclusion of unclear measures has on the scale's validity. The recommendation includes type D problems discovered in combination with data fitting a pure Rasch model, since the theoretical foundation of the model is then imprecise. The recommendation also extends to the logically incompatible combinations of type D problems and LLRM with any degree of uniform LD (marked ***** in Table 4.1), since we in these

cases consider the LD as being caused by statistical type I error and therefore not contributing information to the decision of modificational action.

- When the loglinear Rasch model includes *low* uniform LD, *no* modificational action is needed. This recommendation is based on the consideration that this type of problem may be due to statistical type I errors that will turn up during any kind of multiple testing situation. The results of the subject matter analysis in this study support this consideration in that we found that a low degree of LD was related with finding *no problems* or *problems of type C* involving only one of the items contributing to the LD – none of which warrant modification of the pair of items involved in the LD.
- When the loglinear Rasch model includes uniform DIF, we recommend that the item be exchanged with an item measuring another characteristic to avoid the inconveniences of dealing with biased items during calculation of scores. The recommendation also extends to the logically incompatible combinations of LLRM with uniform DIF and type B or C problems (marked ** in Table 4.1), since the subject matter problems cannot contribute any information as to the nature of the changes needed to eliminate DIF. Other options outside the framework of this modification strategy are, of course, available: To perform additional loglinear Rasch analysis and thereby choose another model, which is then sought improved with the modification strategy. To perform more advanced qualitative investigations into the nature of the DIF problem (Snider & Styles, 2004). Or to retain the item without modifying it with subsequent test equating to determine the correct scale score for the different groups defined by the exogenous variable involved in the DIF.

The cases where the choice of modificational action depends on combinations of the results of the two analyses are the cases where the subject matter analysis reveals problems of type B or C, and the loglinear Rasch analysis reveals no problems (that is when data fits a pure Rasch model), modeling problems or moderate or high uniform LD. The different modificational actions are recommended as follows:

- The modificational action *no change* is recommended when the loglinear Rasch analysis reveals that data fits a pure Rasch model combined with subject matter analysis revealing a type B or type C problem with items included in the Rasch model. The recommendation is based on the finding that no problems revealed in the statistical analysis is reflected in the subject matter analysis as no problems in the majority of cases in this study – in a few cases the subject matter analysis revealed type A problems and in one case a type D problem – leading us to the conclusion that type B and C problems can be so slight that they will not turn up as problems (of LD) in the statistical analysis.
- The modificational action *mutual modification of involved items* is recommended when the subject matter analysis reveals a type B problem (similar wording as other items) with the items pointed out as contributing to moderate or high uniform LD by the loglinear Rasch analysis. It is then, of course, up to the researcher to decide the exact modifications to be made to the items, based

on a thorough analysis of the wording of the items mutually and in relation to the remaining items in the scale.

- The modificational action *exchange of a single item* is recommended when the subject matter analysis reveals a type C problem (measures same characteristic as other items) with the items pointed out as contributing to moderate or high uniform LD by the loglinear Rasch analysis. It is then up to the researcher to decide which of the two items is to be exchanged, based on a thorough analysis of their relationships to the other items in the scale and their precision of measurement.

The recommendations of *mutual modification* respectively *exchange of a single item*, when the loglinear Rasch analysis shows moderate or high uniform LD combined with the subject matter results of type B (similar wording) respectively type C (measure same characteristic) problems, are based on the finding that high LD in the majority of cases in this study is reflected in the subject matter analysis as either type B or type C problems or a combination of both – and vice versa – the analyses results thereby verifying each other. We then extend the recommendation to include moderate LD, since we only consider moderate LD to be problematic if the problem is indeed reflected in the subject matter analysis.

- The modificational actions *exchange of a single item* with an item measuring a different characteristic of the style with *possible modification of other involved items* are recommended when the subject matter analysis reveals a type C problem (measures same characteristic as other items) with the item(s) pointed out as possibly causing modeling problems by the loglinear Rasch analysis, because specific departures from Rasch models generate spurious evidence of other types of problems.

Finally it should be noted that if, in the cases described above, type B and type C problems involving identical items are present, we recommend that modificational action is always chosen based on the type C problem (and the specific problem pointed out by the loglinear Rasch analysis), since type C problems are considered the most severe of the two types of problems.

5. An example

The example application of the modification strategy we present is on the Judicial style scale – the items of the American version of this scale are shown in the first column of table A.1. The loglinear Rasch analyses performed on the Judicial pilot scale resulted in the “best” model being a 7-item loglinear Rasch model (item C excluded) with high uniform LD between items F and G ($\gamma = 0,20$) and items G and H ($\gamma = 0,24$), as illustrated in Figure 5.1.

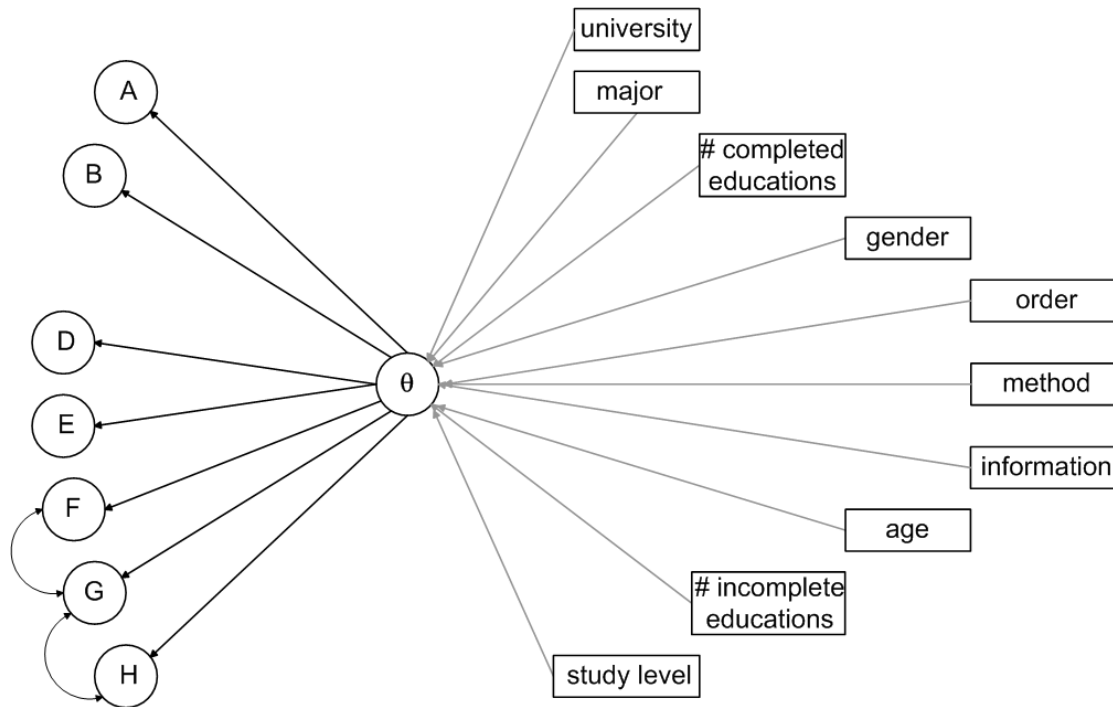


Figure 5.1. The loglinear Rasch model of the Judicial pilot scale⁸.

The loglinear Rasch model in Figure 5.1 defines the starting point for application of the modification strategy in the sense that the scale represented by the model is the one we are aiming to improve by applying the modification strategy. The specific example application of the modification strategy is illustrated in Figure 5.2. The black components show the relevant analyses results and modificational actions for the Judicial scale and the “greyed out” components represent the un-observed results for this scale.

⁸ In Figure 5.1 and Figure 6.1 the recursive structures of the variables in the loglinear Rasch analysis are illustrated by the vertical positions of the exogenous variables, in the sense that equal vertical position refers to the same recursive level in the analysis. The lack of association between exogenous variables illustrates the point that the associations among exogenous variables have *not* been analyzed as part of the loglinear Rasch analysis. The grey associations between exogenous variables and the latent variable are included to represent *potential* associations.

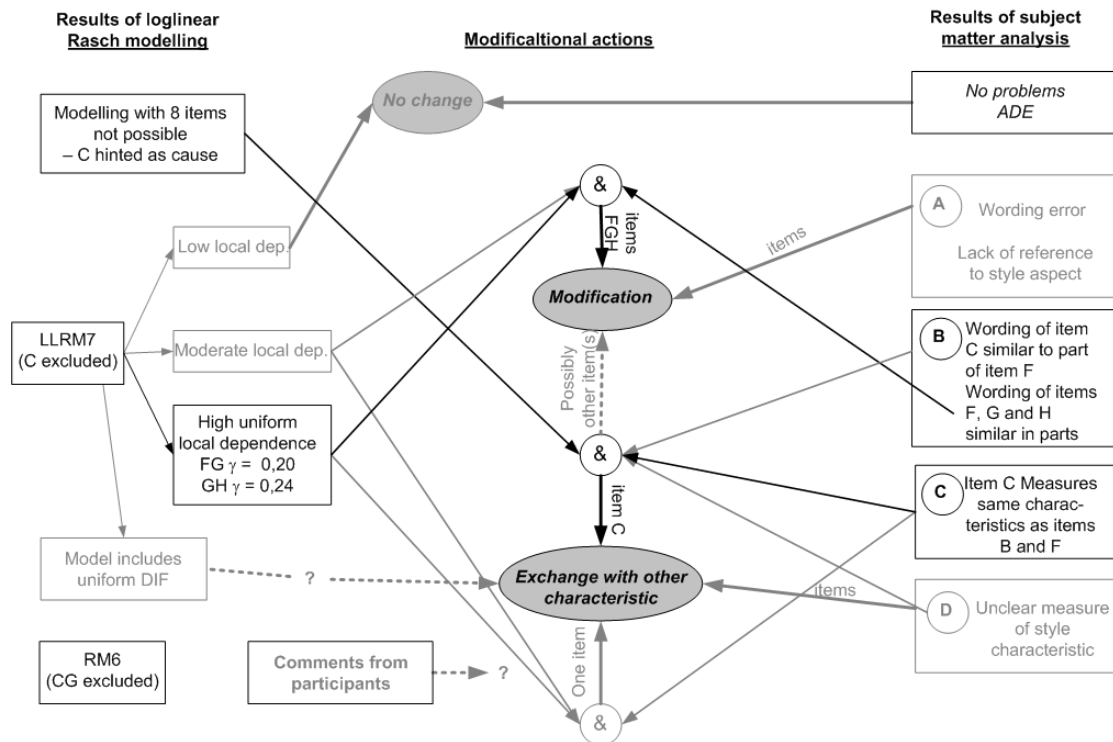


Figure 5.2. Example application of modification strategy on the Judicial pilot scale.

Starting with the results of the loglinear Rasch modeling for the Judicial pilot scale, we see that modeling was not possible with the full 8-item scale and that the analysis hinted that C possibly was the problematic item in the scale – F, G and H were also potentially contributing to the modeling problem, but this was not supported in the subsequent analyses. That the problematic item indeed was item C is supported by the fact that it was possible to model a LLRM when excluding item C from the scale. This 7-item LLRM includes high uniform LD between items F and G ($\gamma = 0,20$) and items G and H ($\gamma = 0,24$). The last result of the loglinear Rasch analyses needed with regard to modification is the 6-item Rasch model accomplished by the exclusion of items C and G – the items that are included in both the local dependencies in the 7-item LLRM. In conclusion, the results of the loglinear Rasch analyses suggests that it is to items C, F, G and H that special attention should be paid in the subject matter analysis – although all items were subjected to this analysis.

The results of the subject matter analysis presented on the right side of the figure support the results of the loglinear Rasch analyses with regard to which of the items in the scale are the problematic ones – the ones in need of modification of some kind. With regard to item C we find: 1) That item C measures the same characteristic as items B and F, and 2) that parts of the wording of item C and item F are similar. With regard to items F, G and H we find that the wording of the three items are similar in parts. No problems with other items were found in the subject matter analysis.

The results of the loglinear Rasch analyses and the subject matter analysis were then combined to determine which modifications should be performed on the problematic items, in the following way.

Item C

With regard to *item C* the loglinear Rasch analysis on the full 8-item scale revealed that inclusion of item C made modeling impossible. The subject matter analysis revealed that item C measured the same characteristic of the Judicial style as items B and F, and that the item wording was similar in parts to the wording of item F. The solution to this problem then was the exclusion of item C. The modeling problem might have its cause in the overlap of measurement with two other items and wording similarity with one of these items. In order to deal with this problem, the modificational action should in this case depend on the most severe of the two problems pointed out by the subject matter analysis (type C) in combination with the modeling problem pointed out by the loglinear Rasch analysis. The appropriate action then was to exchange the item with a new one measuring a different characteristic of the Judicial style, ensuring that there would not be wording similarities with any other items, thereby hoping to solve any conflict with items B and F and preventing new problems to arise in relation to the remaining items. Item C was replaced with an item measuring a characteristic that was only measured by one other item (G).

The finding in the subject matter analysis that items F and B measure the same characteristic is not reflected in the results of the loglinear Rasch analysis – this does *not* reveal any problems concerning these two items. This type of finding, where a potentially serious problem is identified in the subject matter analysis, but no problem is found in the loglinear Rasch analyses, is a rare one in this study. The general finding is that either problems of similar severity are identified in both analyses, or problems which are less severe are found in only one of the analyses – for example LD with $|\gamma\text{-values}| = < 0,1$ in the loglinear Rasch analysis or wording errors and lack of reference to style aspect in the subject matter analysis. For this reason, we decided that no changes were warranted in this case.

Items F, G and H

With regard to *items F, G and H*, the loglinear Rasch analysis of the 7-item scale with item C excluded revealed high uniform LD between item F and G and items G and H. The subject matter analysis revealed that the wording of the three items were mutually similar in parts. The changes should in this case be aimed at removing the wording similarities by differentiating the items mutually and thereby removing the local dependence. The appropriate modificational action was therefore to modify the texts of the three items to be mutually more different in wording without changing the measured characteristics of any of them – of course trying to ensure that the modifications did not cause any new problems to arise in relation with the remaining items. The revised items of the Judicial scale (translated into English) are shown in the second column of table A.1 – It should be noted that it is not possible to see the modification done to item F, since this modification concerned the translation of “points of view”, which is why the two English versions of this item are identical.

6. Results

The fit of the 13 revised scales to loglinear Rasch models was analyzed by data from the follow up sample.

The example scale

The application of the modification strategy to the Judicial pilot scale resulted in a revised Judicial scale with 8 items which fit a loglinear Rasch model with high uniform LD between items B and C ($\gamma = 0,46$), as illustrated in Figure 6.1.

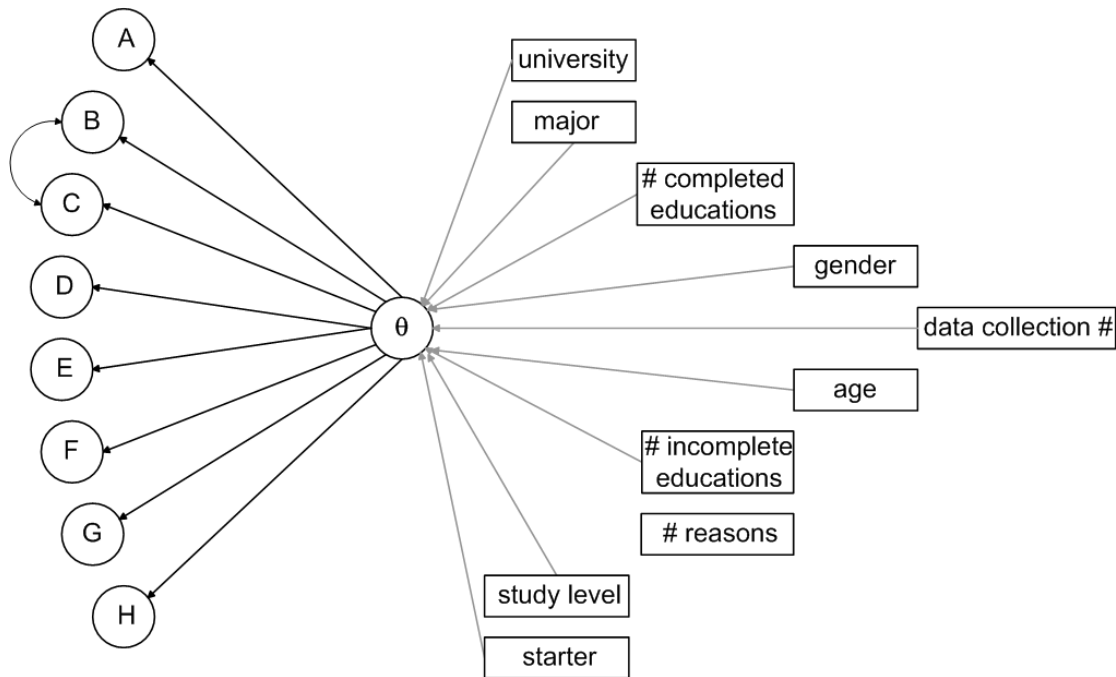


Figure 6.1. Loglinear Rasch model of the revised Judicial scale⁸.

The loglinear Rasch model for the revised Judicial scale in figure 6.1 is not a pure Rasch model, but nevertheless represents a considerable improvement compared to the loglinear Rasch model (figure 5.1) for the pilot study Judicial scale. To recap briefly: We started out with a scale with 7 items (item C was excluded) that could be modeled with a loglinear Rasch model. The model included uniform LD between items F and G and items G and H. The scale was then modified by exchanging item C with a new item measuring a different characteristic of the style and modifying the three items F, G and H. The resulting 8-item Judicial scale, which can also be modeled by a loglinear Rasch model, no longer includes the LD between items F, G and H, it now includes item C and high uniform LD between items B and C. The improvements and the resulting loglinear Rasch model correspond well with the changes made to the scale: The modifications of items F, G and H results in the disappearance of the uniform LD between items F, G and H. The exchange of item C results in item C being incorporated in the scale, and also results in the appearance of a new uniform LD between items B and C. The appearance of the LD between items B and C was potentially expectable from the results of the subject matter analysis on the pilot scale items. However, item C was exchanged with an item measuring a different characteristic, why this problem should not

have arisen. A renewed subject matter analysis of the revised scale does, however, show that there is now a minor wording similarity between items B and C, which could be contributing to the LD. In hindsight, more effort to ensure that such a problem did not arise should have been applied.

The 13 scales of the Danish edition of the MSG Thinking Styles Inventory

The results of applying the modification strategy to the 13 scales of the first Danish edition of the MSG Thinking Styles Inventory show overall improvement of the instrument with regard to the number of items in the subscales fitting loglinear Rasch models and with regard to the type and complexity of the model attained by loglinear Rasch analyses. Scale by scale results are summarized in Table 6.1. No revised subscale have less than seven usable items, while the “best” scale models achieved with the pilot study scales has a larger variation in the number of items included in the fitted models. Four revised subscales fitted pure Rasch models compared to two of the original pilot scales.

Table 6.1. *Scale by scale summary of model fit across the pilot and follow up samples.*

Scale	pilot sample model fit	→	follow up sample model fit
Legislative	7-item LLRM	→	8-item LLRM
Executive	8-item LLRM	→	8-item LLRM
Judicial	7-item LLRM	→	8-item LLRM
Monarchic	5-item RM	→	8-item LLRM
Hierarchic	6-item LLRM	→	7-item LLRM
Oligarchic	7-item LLRM	→	7-item RM
Anarchic	8-item LLRM	→	8-item RM
Local	8-item LLRM	→	8-item LLRM
Global	6-item LLRM	→	8-item LLRM
Internal	6-item LLRM	→	7-item RM
External	6-item RM	→	7-item LLRM
Liberal	8-item LLRM	→	8-item RM
Conservative	6-item LLRM	→	7-item LLRM

Minor problems with DIF are present in the loglinear Rasch models fitting the revised scales. Minor in the sense that of the five single item cases of differential item functioning detected: Two cases are of negligible size ($|\gamma| < .09$). One case is considered spurious due to the fact that the DIF problem is not present for the pilot scale and the involved item has not been modified. The last case is considered genuine bias relative to gender introduced with the construction of a new item C on the revised external style scale ($\gamma = 0,59$). Equating women’s and men’s test results will subtract from 0 to 1.3 points on the external style from women’s scores. The external style scale has a range from 0 to 48. Correcting for DIF therefore has very little practical implications. No serious problems with DIF were present in the fitted models for the pilot sample. That is, two loglinear Rasch models include uniform differential item functioning between a single item and the design variable “order” (stating whether the items appeared in the inventory ordered by style or randomly). However, since the DIF in both cases are uniform, it tells us that the inventory fits the Rasch model in both the random and the ordered-by-style design separately and thereby that as long as items appear in the same

order for all respondents the results indicate that measurements are both valid and objective.

We find that 10 scales have improved considerably. Considerable improvements are considered to be either 1) A higher number of items are included in the “best” revised scale model than was the case for the “best” pilot scale model, while maintaining simplicity of the model – this is the case for 7 scales; the Legislative, Judicial, Monarchic, Hierarchic, Global, Internal, and Conservative scales. Or 2) The same number of items are included, but the “best” revised scale model has improved from being a loglinear Rasch model to being a pure Rasch model – this is the case for 3 scales; the Oligarchic, Anarchic and Liberal scales.

One scale (the Executive) has not changed in any noteworthy way with regard to the “best” scale model (LLRM), however data for the revised scale fitted a pure Rasch model including one item more than was the case for the pilot scale data. The Executive scale remains fitting an 8-item loglinear Rasch model with one uniform LD for the revised scale, however the LD involves a different item and is of a more acceptable size in the revised scale model (revised scale, DH $\gamma = -0,20$ respectively pilot scale, GH $\gamma = 0,57$). The pure Rasch model for the Executive scale has improved from a 6-item Rasch model (GH excluded) for the pilot scale to a 7-item Rasch model (H excluded) for the revised scale.

Two scales declined slightly in quality: The local style scale declined in quality in the sense that the “best” revised scale model scale remains a loglinear Rasch model with 8 items included, but with an additional uniform LD. Also the pure Rasch model for the local style scale declines from including 6 items in the pilot scale to including only 5 items in the revised scale. We have not been able to pinpoint the cause of the decline of the scale. The external style scale declined in quality in the sense that item C in the revised scale turned out to be biased relative to gender as commented on earlier in this section.

7. Discussion

The suggested general modification strategy has emerged from our work with improvement of the 13 scales of the first Danish edition of the MSG Thinking Styles Inventory. The present study does not include examples of all the combinations of results from the two different types of analyses presented in Table 4.1 as can be seen in table A.2 in the appendix. Some of the combinations *not occurring* are considered unimportant due to their logical incompatibility. With regard to the remaining non-occurring combinations we have earlier argued their relationship to modificational action, however we cannot document the effect of the recommended modificational action for these combinations of analyses results. We hope to be able to further explore this part of the strategy in future research.

Another crucial point in this study is the general lack of DIF problems discovered across the scales, which makes our recommendations for item improvement in such cases of a more general methodological and less empirical nature. The area of dealing

with items showing DIF within this strategy (or others) for scale improvement is therefore another area where additional research is needed.

On a different note, we have in one case based our choice of modificational action on wrongful reasoning, thereby diverting from the strategy by mistake. The exception from the recommended modificational action according to the strategy appeared in the analyses of the external scale when the loglinear Rasch analysis revealed that modeling was not possible when including items D and E and the subject matter analysis revealed that items C and E measure the same characteristic of the style, and that items C, D and E have mutual similar wording (type B problem with same items as the modeling problem). These results should, according to the strategy, lead to mutual modification of items D and E. However, we chose to modify item E and exchange item C, based – to put it bluntly – on wrongful reasoning. We simply reasoned as if the two items, C and E, which measured the same characteristic, where the items also involved in the modeling problem. The changes resulted in the inclusion of item E, but not of item D, in the revised external scale, but also in the appearance of a LD problem and the DIF problem reported earlier. Of course we can only guess what the result would have been, had we not made this mistake.

Predictability

A subject of importance is the predictability with regard to the degree of improvement achieved with application of the modification strategy. A good strategy for item modification should of course have a high degree of predictability with regard to the improvements achieved by application of the strategy – as was the case with the Judicial scale in the example earlier. This, however, was not to be the case with all scales. We therefore investigated the details of the relationship between the results of the loglinear Rasch analyses, the results of the subject matter analyses and the actual changes made to each scale on one hand and the resulting scale models for the revised scales on the other. As a result of this investigation, we have to distinguish between three types of results: One with high predictability, one with moderate predictability, and one with low predictability.

The high predictability group: This group is characterized by a good basis on which the modifications are made. That is the “best” pilot scale model is a very simple loglinear Rasch model with few local dependencies and a high number of items included, and the subject matter analysis reveals only a few problems with the scale. For scales belonging to this group the relationship between pilot analysis results and modifications on one hand and the “best” revised scale model on the other is explainable and the resulting models thereby predictable. Belonging to this group are the Legislative, Liberal, and Judicial scales (see earlier for exemplification).

The moderate predictability group: This group is characterized by a moderate basis on which the modifications are made. That is the “best” pilot scale model includes a high number of items, but is a more complicated loglinear Rasch model with more local dependencies, or the subject matter analysis revealed more problems for these scales, than was the case for the high predictability group. For scales belonging to this group the

relationship between pilot analysis results and modifications on the one hand and the “best” revised scale model on the other is partly explainable and the resulting models thereby moderately predictable. Belonging to this group are the Executive, Oligarchic, Anarchic, Local and Conservative scales.

The low predictability group: This group is characterized by a poor basis on which the modifications are made. That is the “best” pilot scale model includes a low number of items and/or is a more complicated loglinear Rasch model with more local dependencies, and the subject matter analysis reveals more problems for these scales, than was the case for the high predictability group. For scales belonging to this group the relationship between pilot analysis results and modifications on the one hand and the “best” revised scale model on the other is partly explainable, but in a less systematic way than was the case in the moderate predictability group and the resulting models thereby less predictable. Belonging to this group are the Monarchic, Hierarchic, Global, Internal and External scales.

A tentative conclusion is that, the better the basis for making item modifications, the more predictable the results, and vice versa. The issue of predictability of degree of improvement is, however, an area of research, which should be developed further in connection with future research into item modification strategies, preferably within different subject areas.

References

- Agresti, A. (1985): *Analysis of Ordinal Categorical Data*. New York: John Wiley and Sons.
- Andersen, E. B. (1970): Asymptotic properties of conditional maximum likelihood estimators. *Journal of Royal Statistical Society, Series B*, 32, 283-301
- Andersen, E. B. (1971): Asymptotic properties of conditional likelihood ratio tests. *Journal of American Statistical Association*, 66, 630-633.
- Andersen, E. B. (1972): The numerical solution of a set of conditional estimation equations. *Journal of Royal Statistical Society, Series B*, 34, 42-54.
- Andersen, E. B. (1973a): A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123-140.
- Andersen, E. B. (1973b): *Conditional Inference and Models for Measuring*. Copenhagen, Mentalhygiejnisk forlag.
- Andersen, E. B. (1977): Sufficient statistics and latent trait models. *Psychometrika*, 42, 357-374.
- Andersen, E. B. (1994): *The Statistical Analysis of categorical Data. Third Edition*. Berlin, Springer-Verlag.
- Christensen, K. B. et al. (2004). Latent regression in loglinear Rasch models. *Communications in Statistics. Theory and Methods*, 33 (6), 1295-1314.
- Costa, P. T. Jr. & McCrae, R. R. (1992). The revised NEO Personality Inventory (NEO-PI-R) and NEO five-factor-inventory (NEO-FFI). Professional manual. Psychological Assessment Resources. Odessa, FL.

- Davis, J. A. (1967): A partial coefficient for Goodman and Kruskal's Gamma. *Journal of American Statistical Association*, 69, 174-180.
- Fischer, G. H. & Molenaar, I. W. (eds.) (1995): *Rasch Models. Foundations, Recent Developments, and Applications*. New York, Springer-Verlag.
- Glas, C. A. W. & Verhelst, N. D. (1994): Testing the Rasch model. In Fischer, G. H. & Molenaar, I. W. (eds.) (1995): *Rasch Models. Foundations, Recent Developments, and Applications* (pp. 69-96). New York, Springer-Verlag.
- Kelderman, H. (1984): Log-linear Rasch model tests. *Psykometrika*, 49, 223-245.
- Kelderman, H. (1989): Item bias detection using loglinear IRT. *Psychometrika*, 54, 681-697.
- Kelderman, H. & Rijkens, C. P. M. (1994): Loglinear multidimensional IRT models for polytomously scored items. *Psykometrika*, 59, 149-176.
- Kelderman, H. & Steen, R. (1988): *LOGIMO I: Loglinear item response theory modeling*. Program manual. Groningen: proGAMMA.
- Kreiner, S. & Christensen, K. B. (2002): Graphical Rasch models. In Mesbah, M., Cole, B. F. & Lee, M. T. (eds.) (2002): *Statistical Methods for Quality of Life Studies* (pp. 187-203). Dordrecht, Kluwer Academic Publishers.
- Kreiner, S. & Bang Christensen, K. (2004): Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. In *Communication in Statistics. Theory and methods*. 33 (6), 1239 – 1276.
- Lauritzen, S. L. (1996): *Graphical Models*. Oxford: Clarendon Press.
- Masters, G. N. (1982): A Rasch model for Partial credit scoring. *Psychometrika*, 47, 149-174.
- Nielsen, T., Kreiner, S. & Styles, I. (2005). Mental Self-Government: Development of the Additional Democratic Learning Style using Rasch Measurement Models. *Paper presented at the 22nd Nordic Congress on Psychology. Copenhagen, Denmark, 18-20th August 2004.* (paper two in thesis).
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark, Denmark's Paedagogiske Institut.
- Rasch, G. (1968). A mathematical theory of objectivity and its consequences for model construction. In *Report from European meeting on Statistics, Economics and Management Sciences*. Amsterdam.
- Rosenbaum, P. R. (1989): Criterion-related construct validity. *Psychometrika*, 54, 625-633.
- Snider, P. & Styles, I. (2004): *Using differential item functioning in the analysis of Triandis' instruments of individualism and collectivism*. Paper presented at the 2nd International Conference on Measurement in Health, Education, Psychology and Marketing: Developments in Rasch and Unfolding Models. Murdoch University, Perth, Western Australia, January 20-22 2004.
- Sternberg, R. J. (1988): Mental self-government: A theory of intellectual styles and their development. *Human Development*, 31, 197-221.
- Sternberg, R. J. (1997): *Thinking Styles*. USA, Cambridge University Press.

Whittaker, J. (1990): *Graphical Models in Applied Multivariate Statistics*. Chichester: John Wiley and Sons.

Appendix. Additional tables

Table A.1. American version of items of the Judicial style scale before and after application of modification strategy.

The original 8 items making up the Judicial scale in the MSG Thinking Styles Inventory (Sternberg, 1997).	The 8 revised items of the Revised Judicial scale.
<p>A) When discussing or writing down ideas, I like criticizing other’s ways of doing things.</p> <p>B) When faced with opposing ideas, I like to decide which is the right way to do something.</p> <p>C) I like to check and rate opposing points of view or conflicting ideas.</p> <p>D) I like projects where I can study and rate different views and ideas.</p> <p>E) I prefer tasks or problems where I can grade the design or methods of others.</p> <p>F) When making a decision, I like to compare the opposing points of view.</p> <p>G) I like situations where I can compare and rate different ways of doings things.</p> <p>H) I enjoy work that involves analyzing, grading, or comparing things.</p>	<p>A) When discussing or writing down ideas, I like criticizing other’s ways of doing things.</p> <p>B) When faced with opposing ideas, I like to decide which is the right way to do something.</p> <p>C) I like to evaluate different methods and procedures in relation to each other.</p> <p>D) I like projects where I can study and rate different views and ideas.</p> <p>E) I prefer tasks or problems where I can grade the design or methods of others.</p> <p>F) When making a decision, I like to compare the opposing points of view.</p> <p>G) I like situations where I can evaluate different ways of doings things.</p> <p>H) I enjoy work that demands analysis, grading, or comparison.</p>

Note. Please note that we have been working with Danish editions of the items in this study.

Table A.2. Occurrence of specific combinations of results of loglinear Rasch analyses and subject matter analyses involving identical items.

Loglinear Rasch analyses results	Subject matter analyses results				
	No problem	Type B wording similar	Type C measure same characteristic	Type D unclear measure	type A wording error
RM	+		+	+	+
LLRM with <i>low</i> uniform LD	+			+ logically incompatible	
LLRM with <i>moderate</i> uniform LD	+		+	logically incompatible	+
LLRM with <i>high</i> uniform LD		+	+	logically incompatible	+
LLRM with uniform DIF	+	logically incompatible	logically incompatible		
Modeling not possible	+	+	+	+	+

+ Signifies that the combination has occurred in this study.

The coloring of the table cells are identical to that of Table 4.1 to facilitate comparison.