Modeling typical performance measures

Anke M. Weekers

Samenstelling promotiecommissie

Voorzitter/Secretaris	Prof. Dr. H.W.A.M. Coonen
Promotoren	Prof. Dr. C.A.W. Glas Prof. Dr. R.R. Meijer
Assistent-promotor	Dr. Ir. B.P. Veldkamp
Leden	Prof. Dr. C.W.A.M. Aarts Prof. Dr. Ir. T.J.H.M. Eggen Prof. Dr. K. Sanders Prof. Dr. K. Sijtsma

ISBN 978-90-365-2913-6
Druk PrintPartners Ipskamp B.V., Enschede
Cover designed by Suzanne Luikinga

MODELING TYPICAL PERFORMANCE MEASURES

PROEFSCHRIFT

ter verkrijging van

de graad van doctor aan de Universiteit Twente,

op gezag van de rector magnificus,

prof. dr. H. Brinksma,

volgens besluit van het College voor Promoties

in het openbaar te verdedigen

op woensdag 16 december 2009 om 16.45 uur

door

Anke Martine Weekers

geboren op 18 september 1979

te Weert

Dit proefschrift is goedgekeurd door

de promotoren Prof. Dr. C.A.W. Glas en Prof. Dr. R.R. Meijer
en de assistent-promotor Dr. Ir. B.P. Veldkamp

Contents

Chapter 1

Introduction

Attitude and personality measures are getting more and more attention in the educational, employment and clinical context. Analogous to ability measures, these typical performance measures (Cronbach, 1984) are important predictors for outcomes such as performance and satisfaction across situations (Meyer et al., 2001). Attitude and personality traits cannot be observed directly. Therefore, observed responses have to be gathered by means of inventories (or scales), observations, and interviews. To translate the responses to questions into a latent (unobservable) value for the attitude or personality trait, statistical techniques are being used. Although all data collection techniques obtain similar but partially overlapping information, and a combination of the techniques is necessary in practice, in this thesis the focus is on inventories or scales. In the remainder of this thesis the term typical performance measures will be used.

1.1 Typical performance measurement

Observed responses to measure personality traits and attitudes are often collected using inventories. Inventories consist of a number of statements that are supposed to measure the intended traits. Persons have to respond to the statements on dichotomous (2 answer options) or polytomous (3 or more answer options) Likert scales. An example of statements about the personality trait Conscientiousness with response categories on a 4-point Likert scale is given in Figure 1.1. Some typical performance measures are designed to measure just one trait (as in the figure), whereas others are designed to measure a wide range of traits, each consisting of a set of related statements.

Statements	Strongly Disagree	Disagree	Agree	Strongly Agree
I am always prepared	☐	☐	☐	☐
I make a mess of things	☐	☐	☐	☐

Figure 1.1. Example of Conscientiousness statements with 4-point Likert scale response categories.

Statistical procedures can be followed to obtain latent trait estimates about the personality or attitude constructs. Most of these procedures heavily rely on classical test theory (CTT) and factor analytical methods, but recently dominance item response theory (IRT) models have also been used to analyze typical performance data. The models used to construct and analyze typical performance measures are often copied from the area of maximum performance measurement (i.e. from the area of educational and cognitive measurement). However, there are several systematic features of typical performance measures that warrant attention. In this thesis two features are discussed; 1. the multitude of factors in typical performance measures, and 2. response processes on typical performance measures.

The first important difference between maximum performance assessment and typical performance assessment is that personality and attitude scales have more complexity in their factor structures than cognitive ability tests. As noted by many psychological theorists (e.g., Funder, 1997) attitude and personality are usually determined by a multitude of factors. Through this additional complexity, more complex test models, such as multidimensional models, might be much more eminent in typical performance measurement than in educational measurement.

A second difference between maximum and typical performance is that in cognitive assessment it is often useful to think of a domain, where the test items are a sample from the domain. In general, these cognitive tests need to be long to be reliable and, because the domain is so large researchers need a large sample of items to accurately assess the domain. In typical performance assessment, many of the domains are quite restricted. One cannot repeat statements (e.g., asking respondents how depressed they are) over and over again. Thus in typical performance assessment large item pools do not exist or item pools consist of very similar statements (Reise & Henson, 2003). The consequence is that it is difficult to create long inventories for these constructs, because researchers simply run out of non-redundant questions.

Yet, given the relatively long history of typical performance

measurement and the fact that typical performance research has historically been at the forefront of statistical and methodological innovations, it is surprising that recent IRT analyses have shown that the structure of many well-known and often used personality measures is not well understood (e.g., Chernyshenko et al., 2001; Reise & Waller, 2003; Meijer & Baneke, 2004). Personality measures might follow a different response process than implied under the dominance IRT models and factor analytic models used. For example, unfolding response models, which have already been proposed by Coombs (1964) for the attitude domain, may provide a better description of the responses to personality items as well. An advantage of the unfolding models is that items can be written over a broader range of the trait continuum (see Chernyshenko et al., 2001), so more items can be written and larger item pools can be constructed. In the example in Figure 1.1 a neutrally formulated item like "Half of the time I do things according to a plan" could be included in the inventory, whereas this is not an option when using factor analytical or dominance IRT methods.

Because of these differences between maximum performance and typical performance it is important to not simply apply statistical procedures applied in maximum performance testing to typical performance testing. The applicability of the procedures has to be investigated first. This thesis focuses on the usefulness of various models to investigate dimensionality and response behavior on typical performance measures. Multidimensional models will be discussed in Chapter 2 and 3, and unfolding IRT models and statistical testing of these unfolding IRT models will be discussed in Chapter 4, 5 and 6. First some brief descriptions of the structural equation modeling (SEM) and item response theory (IRT) frameworks will be given in the next two paragraphs. SEM (Paragraph 1.2) can be used to investigate the dimensionality of typical performance measures, while IRT (Paragraph 1.3) is used to investigate response processes to typical performance measures. After the explanations of the modeling frameworks and the usefulnes of these models for modeling typical performance measures, an overview of the thesis will be given in Paragraph 1.4.

1.2 Structural equation modeling

The structural equation modeling (SEM or covariance structure analysis; Bollen, 1989; Kline, 2005) framework refers to a family of statistical procedures. Most statistical techniques in SEM make a distinction between observed and latent variables. Relations between observed

variables, between observed variables and latent variables and between latent variables are studied. In general, SEM uses measurement parts, which describe the relations between observed variables (the statements) and latent variables (the constructs measured), and structural parts, that model (causal) relations between constructs. In general, in SEM the major statistics to analyze relations between variables are covariance and correlation.

The structural equation models used in this thesis are mainly based on confirmatory factor analysis, which is a technique for the estimation of measurement models. Covariances and correlations between many observed variables are explained by means of one or more underlying latent variables. The observed statements are considered to be from interval level and to be linearly associated with one another and the underlying construct. Even though the observed statements are considered to be of interval level, the responses to typical performance statements are on the ordinal level as is shown in the example in Figure 1.1. The solution to this difference is that responses to statements are assumed to represent a truncation of a hypothetical underlying continuous normally distributed response process. Thresholds represent the shift from one categorical response to the other. Dependent on the position of a person on the continuous response continuum of a statement, the person will respond in the category that covers the persons' position. The hypothetical continuous responses to statements are used to calculate the relations between the variables.

Typical performance measures consist of a number of statements measuring one or more constructs. Constructs might be (strongly) related facets (subconstructs) of a more general construct or several major constructs. Individual constructs or facets are often described by unidimensional models. In case of more than one (sub)construct, the constructs might be correlated (non-hierarchical multidimensional models) or might measure a more general factor that describes the relationship between the constructs (i.e. second-order model; DeYoung, 2006; Digman, 1997; Gustafsson, 1984). However, other types of multidimensional models (i.e. the bifactor model, that uses both general and domain-specific constructs) might explain the relations between typical performance constructs and/or facets more precisely. Structural equation modeling can help to evaluate inventories containing a multitude of factors. In this thesis the applicability of the non-hierarchical multidimensional model, the second-order model and the advanced bifactor model to typical performance measures will be investigated.

1.3 Item response theory models

Takane and de Leeuw (1987) show that item response theory (IRT) models can be viewed as an extension of the more commonly used factor-analytic models. IRT modeling is a statistical technique that focuses on individual observations. This technique has rapidly become the theoretical basis for maximum performance assessment, and recently is also applied for typical performance measures as well.

Dominance IRT models explain the performance of a person on a test item by latent factors. The influence of respondents and test items is explicitly modeled by different sets of parameters. Categorical observed responses to statements are used directly, and the relationship between a person's item response and the latent factor underlying this response can be described by an item characteristic curve (ICC). An ICC gives the response probability as a function of the latent variable by nonlinear functions. Both dominance IRT models for dichotomous items (i.e. Rasch model, 2-parameter logistic model, 3-parameter logistic model) and for polytomous items (i.e. generalized partial credit model, graded response model, sequential model) exist (Embretson & Reise, 2000; Hambleton, Swaminatan, & Rogers, 1991; Van der Linden & Hambleton, 1997).

In this thesis, IRT models for dichotomous items are discussed. Dominance IRT models assume that the ICCs are monotone increasing or monotone decreasing functions. These functions are modeled by (highly) restricted parametric models (Lord, 1980) or more general non-parametric models (Sijtsma & Molenaar, 2001). The idea behind these models is that the higher a person is located on the latent trait the more statements the person will endorse. A person is likely to endorse all statements that have an item location below the person location. Although research on applications of IRT models to typical performance measurement is increasing (Reise & Waller, 2003), first attempts to model typical performance data by these models found contradictory results. Several researchers reported reasonable fit of 2-parameter logistic models (i.e. models of which the ICC is described by two parameters, a location and a discrimination parameter), but recent studies showed that more general models might be needed to describe typical performance data. One possibility is to use unfolding IRT models, which will be extensively studied in this thesis. Under unfolding IRT models the probability of endorsement of a dichotomous statement is described by a single-peaked ICC. The idea behind these models is that persons only endorse statements if their person location is close to the item location.

Persons located at the higher end of the trait range are not supposed to endorse a high number of statements as is the case for dominance IRT models, but only endorse the statements that represent his/her location, the positively formulated statements. Persons located at the lower end of the trait range are only supposed to endorse statements on the lower end of the continuum, the negatively formulated statements, and persons located in the middle of the trait continuum are supposed to only endorse statements located in the middle of the continuum, the neutrally formulated statements. The applicability of unfolding IRT models will be investigated in this thesis.

1.4 Overview of the thesis

Two main topics will be addressed in this thesis, modeling of the multitude of factors in typical performance measurement, and modeling of the response processes on typical performance measures. Chapters 2 and 3 investigate the dimensionality structure of personality and attitude inventories. In Chapter 2 the differences in appropriateness of a number of factor analytical models (i.e. non-hierarchical multidimensional model, second-order model, and bifactor model) are investigated. Different models are applied using empirical data of a dichotomously scored personality inventory. Using different models, the dimensionality structure of the instrument, the dimensionality of items, the interpretability of scales for practical implications, and the scoring of individuals on constructs will be discussed. Chapter 3 discusses only a selection of these models (non-hierarchical multidimensional model and bifactor model), which are applied to investigate the dimensionality structure of a polytomous attitude inventory, and to investigate the dimensionality of the items, and the interpretability of these scales.

Research on response processes and statistical fit of the associated models are discussed in the Chapters 4, 5, and 6. To obtain more insight into the response processes on typical performance data, in Chapter 4 it is investigated whether dominance or unfolding IRT models give a better description of the response processes on personality trait inventories. In this chapter, both dominance response processes and ideal-point response processes are discussed, and parametric and non-parametric dominance IRT models, and parametric and non-parametric unfolding IRT models are applied. Chapter 5 and 6 move on to investigate statistical fit on unfolding models for ideal-point response processes. An existing model, the

generalized graded unfolding model (GGUM), and three newly developed models, the collapsed generalized partial credit model (CGPCM), the collapsed graded response model (CGRM) and the quadratic logistic regression model (QLOG) are studied. From the beginning of typical performance assessment, authors of inventories are seriously concerned with both measuring and correcting for respondents' tendencies to deceive themselves or others in responding to statements. Therefore two person fit statistics are developed and investigated for unfolding models in Chapter 5. The newly developed person fit statistics are applied in a simulation study on a real attitude data set. On the other hand, item fit is important because instruments are developed that are used in a population of persons. Item fit can help the test constructor to develop an instrument that fits an IRT model in that particular situation. In Chapter 6 two item fit statistics are developed and tested in a simulation study and in a real data example.

After the studies on response processes and unfolding models, conclusions and suggestions for further research will be given in Chapter 7. The chapters in this thesis are self-contained, hence they can be read separately. Therefore, some overlap could not be avoided and the notations, the symbols and the indices may slightly vary across chapters.

Chapter 2

A comparison of factorial models in personality measurement

2.1 Introduction

Psychological tests and questionnaires often measure a number of related constructs. Two examples are intelligence test batteries that include both general and domain-specific intelligence factors such as verbal intelligence and spatial ability, and personality measures such as depression questionnaires that include multiple indicators of, for example, negative mood, suicidal ideation, and social withdrawal. The analysis of the dimensionality structure of these measurement instruments relies heavily on confirmatory factor analysis. The dimensionality structure is often explored using non-hierarchical multidimensional models or higher-order models such as second-order models (e.g., DeYoung, 2006; Digman, 1997; Gustafsson, 1984).

Bifactor models have a rich history in the intelligence domain (e.g., Rindskopf & Rose, 1988; Luo, Petrill, & Thompson, 1994), and the ability and achievement domain (e.g., Gustafsson & Balke, 1993). Rindskopf and Rose (1988), Gustafsson and Balke (1993), Chen, West, and Sousa (2006) and Reise, Morizot, and Hays (2007) discussed statistical and conceptual similarities and differences between non-hierarchical multidimensional models, second-order models, and bifactor models.

Although the use of bifactor models is increasing there is not much experience with these models to analyze personality and health domain

data. Exceptions are Brouwer, Meijer, Weekers, and Baneke (2008), Chen, West, and Sousa (2006), Patrick, Hicks, Nichol, and Krueger (2007) and Reise, Morizot, and Hays (2007). Reise, Morizot, and Hays (2007) and Brouwer, et al. (2008) show that bifactor models are excellent tools to investigate whether multidimensionality of an instrument interferes with the scaling of individuals on unidimensional domain-specific constructs. Any scale which is not simply the repeating of the same item over and over again has some multidimensionality, and on a theoretical and conceptual level the constructs might be described as relatively distinct, but on a measurement level participants might not perceive measures of the domain-specific constructs in this way. Therefore, it is important to investigate if a more general construct is viable, if domain-specific factors make a contribution over and above the general factor, and how the results can be used in practice.

In the present study, we extend the Chen, West, and Sousa (2006) study, and the Reise, Morizot, and Hays (2007) study by analyzing a personality inventory, the Dutch Personality Inventory for Adolescents (Dutch: Junior Nederlandse Persoonlijkheidsvragenlijst; NPV-J; Luteijn, van Dijk, & Barelds, 2005), using the non-hierarchical multidimensional model, the second-order model and the bifactor model, and by discussing the practical implications for interpretation and for scoring of individuals when using the different models. First, we explain the non-hierarchical multidimensional model, the second-order model, and the bifactor model, and discuss the statistical and conceptual similarities and differences between the models. Second, we apply these models to empirical data. Finally, recommendations about the appropriateness of the models in practice are given.

2.2 Factorial models

In Figure 2.1, a graphical representation of the three models used in this study is given. A common representation was used, in which squares represent the observed item responses, circles represent the latent factors, straight arrows represent item factor loadings, and curved double-headed lines represent correlations.

The factorial structure of a particular measure is modeled in three ways. In the non-hierarchical multidimensional model (Figure 2.1a), there is more than one common factor among the items, and the factors are correlated. Each item in a multifactor measure loads on one factor only. When each factor is hypothesized to have a non-zero correlation with every other factor,

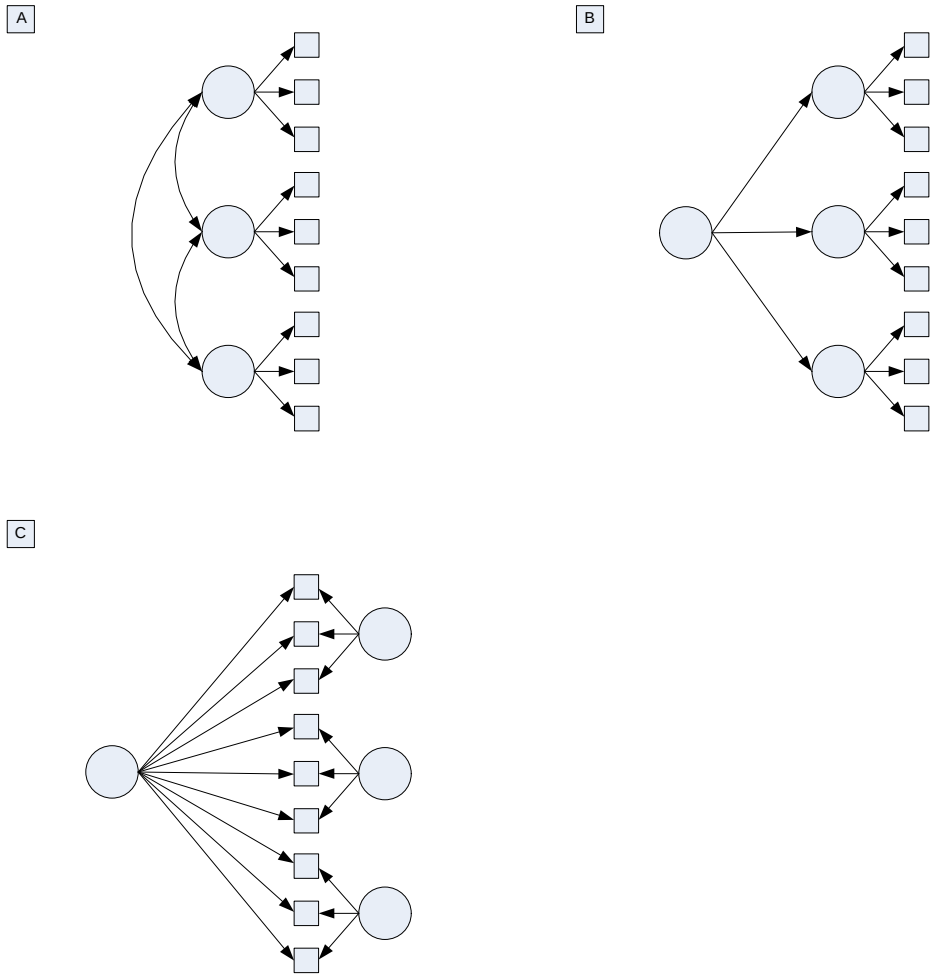

Figure 2.1. Graphical representation of a) the full non-hierarchical multidimensional model, b) the second-order model, and c) the bifactor model.

the model is a full non-hierarchical multidimensional model. However, it is possible that some factors have zero correlations with some other factors and are non-zero correlated with other factors.

In a second-order model (Figure 2.1b), items are loading on first-order factors and first-order factors are loading on second-order factors. The second-order factor is a conceptually different type of dimension, a superordinate dimension, which represents a single broad, coherent construct. So first-order factors account for correlations between items, and second-order factors account for the communality among latent first-order factors.

Under this model items are not directly influenced by the general second-order factor.

The bifactor model (also called group-factor model; Figure 2.1c) specifies one general factor, and two or more group factors. In most applications, items load on both the general factor and one of the group factors. In this study, a bifactor model is considered in which general and group factors are assumed to be orthogonal (correlation is zero) to each other. The general factor then explains the item intercorrelations, but in addition there are group factors that attempt to capture the item covariation that is independent of the covariation due to the general factor. Items in the same scale in an inventory are related because they share both general and subscale variance.

2.2.1 Similarities and differences between the factorial models

In the non-hierarchical multidimensional model, the correlations between dimensions are estimated based on the hypothesis that items are influenced by multiple correlated domain-specific factors. However, the higher the correlation among domain-specific factors the more likely a general factor is dominating the item responses, and interpretation of the subscale scores can be confounded by an overall factor. The second-order model and bifactor model provide an overall factor that explains common variance in items of different scales. In the second-order model this is the second-order factor, and in the bifactor model this is the general factor. The overall factors of both the second-order model and the bifactor model correspond to each other, and have similar interpretations (Chen, West, & Sousa, 2006; Gustafsson & Balke, 1993, Rindskopf & Rose, 1988; Yung, Thissen, & McLeod, 1999). The only difference is that the second-order model specifies the common variance via first-order factors, whereas under the bifactor model items directly load on the general factor.

When there are three domain-specific factors the non-hierarchical multidimensional model and second-order model are equivalent (see also Rindskopf & Rose, 1988). The non-hierarchical multidimensional model and the second-order model have the same number of parameters, which makes them statistically undiscriminable, and both models have equal goodness-of-fit statistics and standardized factor loadings on the first-order factors. Model fit alone cannot tell us which model is more appropriate. The difference between the models is that the non-hierarchical multidimensional model estimates correlations between first-order factors,

whereas the second-order model estimates factor loadings of first-order factors on the second-order factor. Both models can only be distinguished in terms of interpretability of parameter estimates and meaningfulness of the model (for further explanation on equivalent models see, Bollen, 1989, and MacCallum, Wegener, Uchino, & Fabrigar, 1993). The difference in interpretation is that the second-order model puts a structure on the pattern of non-restricted correlations among the first-order factors as modeled in the non-hierarchical multidimensional model.

In the bifactor model, item variance can be partitioned into item variance due to general and group factors. The group factors are directly specified in the bifactor model, and explain the item intercorrelations that capture the residual variation due to secondary dimensions. In the second-order model, they are modeled in the disturbances of the first-order factors, and are not directly visible. The disturbances of the first-order factors in the second-order model have the same interpretation as the group factors under the bifactor model. Both explain common variance between items after partialing out the general factor. Only when there is a weak general factor and relevant domain-specific factors, interpretation of the domain-specific scores under the non-hierarchical multidimensional model is less confounded by the general factor, and this model might be a viable alternative.

The differences between the models become more important when researchers are interested in the contribution of one or more of the domain-specific factors over and above the general factor, and in the prediction of external variables. Using a non-hierarchical multidimensional model it is difficult to predict outcome variables of interest, because of substantial overlap in variability. The second-order model separates general and domain-specific variance. However, domain-specific variance is modeled in the disturbances, and as a consequence it is difficult to predict external variables by domain-specific factors. For the bifactor model it is easy to estimate latent domain-specific factors over and above the general factor, and to predict external criteria by these domain-specific factors. Since group factors are identified in the bifactor model only if residual variance is left after the general factor is identified, an empirically informed judgment regarding the utility of creating and scoring domain-specific factors, and the dimensionality of items, unidimensional or multidimensional, can be made. However, the estimation of bifactor models in which one subscale does not exist may cause computational problems. In this case, the second-order model should find few residual variance, and loadings of the first-

order factors on the second-order factor close to unity. Although under the second-order model the correct interpretation would also be that the domain-specific factors do not exist as residual factors (no significant disturbances), this is often overlooked.

The differences between the models are also important when the objective is to investigate whether multidimensionality of the instrument interferes with scaling of individuals on unidimensional constructs. Dependent on the model, individuals will be scored on domain-specific factors (non-hierarchical multidimensional model), on domain-specific factors and general factors as indicators of domain-specific factors (second-order model), or on separate domain-specific and general factors (bifactor model). General factor scores are estimated, via first-order factors (second-order model) or directly from the items (bifactor model), and domain-specific factor scores have a different interpretation under the bifactor model compared to the non-hierarchical multidimensional model and the second-order model. Misspecification of the model may have serious consequences for the scoring of individuals on general and domain-specific latent constructs.

2.3 Aim of this study

This chapter discusses an application of the non-hierarchical multidimensional model, the second-order model, and the bifactor model in the personality domain. These three models, and the uncorrelated unidimensional model as a baseline model (to be discussed below), were used to analyze data of the Dutch Personality Inventory for Adolescents (Dutch: Junior Nederlandse Persoonlijkheidsvragenlijst; NPV-J; Luteijn, van Dijk & Barelds, 2005). The aim of this study was to investigate the relevance of these models to enhance the understanding of the content of the NPV-J, and the scaling of individuals on it. The appropriateness of the models was checked to investigate the dimensionality structure of the NPV-J and the dimensionality of the items, the interpretation of subscale scores for practical implications, and the scoring of individuals on constructs.

2.4 Method

2.4.1 Instrument

The NPV-J is a general personality inventory for selection of adolescents for different types of education, and for diagnostic purposes. The NPV-J consists of five scales; Inadequacy, Persistence, Social Inadequacy, Recalcitrance, and Dominance. The five scales are used as unidimensional scales and individuals are scaled on the personality characteristics using simple sum scores. The simple sum scores are often combined in profile scores.

Barelds and Luteijn (2002) investigated the relation between the Dutch Personality Questionnaire (Dutch: Nederlandse Persoonlijkheids-vragenlijst, NPV, the adult version of the NPV-J; Luteijn, Starren, & van Dijk, 1985), and the Five Factor Personality Inventory (FFPI; Hendriks, Hofstee, & de Raad, 1999). They found that Inadequacy was related to Emotional Stability ($r = -.65$), Social Inadequacy and Dominance were related to Extraversion ($r = -.74$ and $r = .48$, respectively), and Persistence was related to Conscientiousness ($r = .57$). The content of the NPV-J was compared with the content of five factor model questionnaires. Based on independent content sorting, relations between NPV-J scales and five factor model subdomains were found (see Table 2.1). The relations to the five factor model subdomains will be used for scale interpretation under the different models.

2.4.2 Participants and procedure

Data were collected from 609 primary and secondary school pupils, 331 mostly White girls, and 278 mostly White boys. They attended primary and secondary schools in the East of the Netherlands. All participants were between 9 and 15 years of age, with a mean age of 12.7 ($SD = 2.1$).

The participants filled out the inventory, which consisted of 105 statements about themselves. Statements were unequally divided over the five scales. In the NPV-J, statements are administered using a three-point scale (*Agree, ?, Disgree*), but because the instructions of the NPV-J discourage the use of the *?* response, and because I was afraid that many adolescents would choose the *?* category, a two-point-scale, *Agree* versus *Disagree,* was used.

Table 2.1
Relation between NPV-J items and subdomains of the Five Factor Model

NPV-J scale	Subdomains Five Factor Model	Itemnumber NPV-J
Inadequacy	Anxiety	04, 13, 19, 32, 48, 57, 70, 75, 91, 96, 98
	Depression	01, 06, 08, 14, 28, 34, 36, 38, 50, 52, 54, 59, 66, 72, 93, 100, 102
Persistence	Orderliness	33, 104
	Achievement-striving	31, 39, 43, 45, 95
	Dutifulness/Self-discipline	02, 10, 12, 30, 41, 53, 63, 68, 69, 71, 73, 77, 78, 84, 88, 94, 101, 103
Social Inadequacy	Sociability	21, 23, 26, 44, 51, 62, 79, 80, 85, 89, 105
	Introversion	22, 25
Recalcitrance	Trust	05, 18, 24, 29, 37, 40, 49, 55, 61, 74, 82, 83, 87
	Altruism	11, 15, 16, 20, 35, 42, 46, 47, 65, 86, 92
Dominance	Assertiveness	03, 07, 09, 27, 56, 58, 60, 64, 67, 76, 81, 90, 97
	Activity Level	17, 99

2.4.3 Analyses

The quality of the items of the NPV-J was investigated in an earlier study on scaling of response processes on the NPV-J (see Chapter 4). Number of items (k), scale means, Cronbach's α, skewness, kurtosis, and mean item-test correlations (ρ_{iT}) for the original five scales are given in Table 2.2. Reliability ranged from .62 to .87, and mean item-test correlations from .25 to .42.

The correlations between the sum scores on all five scales are shown in Table 2.3. These values are similar to the values found by Luteijn, van Dijk, and Barelds (2005, p.16). Because of the moderate to high correlations, to compare the different models, we selected the three scales, Inadequacy, Social Inadequacy, and Recalcitrance. Luteijn, van Dijk, and Barelds (2005) already mentioned the strong relations between Inadequacy,

Table 2.2

Descriptive statistics NPV-J data

Scales	k	M	SD	α	Skewness	Kurtosis	ρ_{iT}
Inadequacy	28	6.36	5.37	.87	1.13	0.91	.42
Persistence	25	18.27	3.80	.73	-0.63	0.00	.27
Social Inadequacy	13	5.28	3.09	.78	0.23	-0.81	.40
Recalcitrance	24	8.45	3.68	.72	0.69	0.47	.27
Dominance	15	5.10	2.48	.62	0.73	0.63	.25

Table 2.3

Correlations of sumscore between NPV-J scales

Scale	Inadequacy	Persistence	Social Inadequacy	Recalcitrance
Persistence	-.004			
Social Inadequacy	.523	.111		
Recalcitrance	.475	.079	.361	
Dominance	.088	.075	-.108	.206

Social Inadequacy, and Recalcitrance. These moderate correlations may indicate the presence of a higher-order factor (Chen, West, & Sousa, 2006; Reise, Morizot, & Hays, 2007).

Also theoretically the three scales might measure one general construct. Inadequacy and Social Inadequacy both measure insecure and anxious behavior, whereas Recalcitrance measures distrust and non-cooperative behavior. Together, these three scales can be seen as a measure of Inadequate Behavior. From this point of view Inadequate Behavior is the general factor consisting of 65 items, of which 28 items measure Inadequacy, 13 items measure Social Inadequacy, and 24 items measure Recalcitrance. The question is whether the general factor is strong enough to be measured as a separate construct or whether a multidimensional representation has to be preferred.

Because subscales of the NPV-J are used as independent scales, the uncorrelated unidimensional model consisting of three uncorrelated unidimensional scales was estimated besides the non-hierarchical multidimensional model, the second-order model, and the bifactor model. The expectation is that the uncorrelated unidimensional model does not fit the data well, because of moderate sumscore correlations between the three scales. In this study, the uncorrelated unidimensional model will be used as a baseline model.

MPLUS 4.1 (Muthén & Muthén, 1998-2006) was used to estimate

Table 2.4
Fit statistics NPV-J data

Model	χ^2 (df) p-value	CFI	TLI	RMSEA	SRMR
Unidimensional model	16558.44 (2015) <.01	.55	.53	.11	.19
Non-hierarchical multidimensional model	4498.81 (2012) <.01	.92	.92	.05	.10
Second-order model	4498.81 (2012) <.01	.92	.92	.05	.10
Bifactor model	3678.64 (1950) <.01	.95	.94	.04	.08

the four models. The Weighted Least Squares Mean Adjusted (WLSM) estimation option was used for all calibrations. For model evaluation the WLSM estimation option provides a Chi-square statistic (χ^2), Comparative Fit Index (CFI), Tucker-Lewis Index (TLI), Root Mean Squared Error of Approximation (RMSEA), and Standardized Root Mean Squared Residual (SRMR). Criteria for the fit statistics were set at values of .95 or higher for CFI and TLI, a value of .08 or lower for SRMR, and a value of .06 or lower for RMSEA. These values constitute good fit as was suggested by Hu and Bentler (1999). Furthermore, factor loadings, correlations, residual variance, and factor scores were studied. Items are interpreted to load on a factor if the factor loading is at least .35 (Stevens, 2002). Items with loadings greater than or equal to .35 on more than one factor are interpreted as multidimensional. Individuals' simple sum scores on the factors (conform scoring as described in the manual) were computed to compare them to the individuals' weighted factor scores estimated under MPLUS. Factor score values range from negative to positive, with a mean value of (about) zero, and an estimated standard deviation.

2.5 Results

2.5.1 Dimensionality structure and interpretation

Fit statistics for all models are shown in Table 2.4, and item factor loadings under the models are shown in Table 2.5.

As expected, the uncorrelated unidimensional model showed no acceptable fit. Values for all fit statistics were below the cutoff criteria

(CFI and TLI) or above the cutoff criteria (RMSEA and SRMR). As expected, the non-hierarchical multidimensional model and the second-order model showed equally reasonable fit, whereas the bifactor model showed acceptable fit. For all three models the RMSEA statistic was below the cutoff criterion. The CFI and SRMR statistic were above and below the cutoff criteria for the bifactor model only, and the TLI statistic was below the cutoff criterion for the three models, but almost equal to the cutoff criterion for the bifactor model.

Table 2.5 shows that under the uncorrelated unidimensional model all Inadequacy items, most Social Inadequacy items and seventeen out of 24 Recalcitrance items had loadings of .35 or higher on their constructs. The two Sociability items of the Social Inadequacy scale and the four Trust items and three Altruism items of the Recalcitrance scale with loadings below .35 were items that also in a former study by Weekers and Meijer (2008; described in Chapter 4) were found to be of low quality or had single-peaked response curves.

For the non-hierarchical multidimensional model and the second-order model, besides equal fit, the factor loadings of the items on the first-order factors were equal as well. Equal loadings on first-order factors were the result of equivalence of both models, because there were only three first-order factors. However, correlations (non-hierarchical multidimensional model) and factor loadings of the first-order factors on the second-order factor (second-order model) differed. The analyses showed that all Inadequacy, most Social Inadequacy, and fifteen out of 24 Recalcitrance items had loadings of .35 or higher on their constructs. One Sociability item of the Social Inadequacy scale, and five Altruism items and four Trust items of the Recalcitrance scale had loadings below .35. Not all Trust and Altruism items were similar to the items with loadings below .35 under the uncorrelated unidimensional model.

Correlations between the three constructs under the non-hierarchical multidimensional model were moderate to high; $r = .66$ between Inadequacy and Social Inadequacy, $r = .74$ between Inadequacy and Recalcitrance, and $r = .54$ between Social Inadequacy and Recalcitrance. Under the second-order model loadings of the first-order factors on the second-order factor were high also; $\lambda = .95$ for Inadequacy, $\lambda = .70$ for Social Inadequacy, and $\lambda = .78$ for Recalcitrance, and residual variance was low; $\zeta = .10$ for Inadequacy; $\zeta = .51$ for Social Inadequacy, and $\zeta = .39$ for Recalcitrance. Under the second-order model, Inadequacy was an almost perfect indicator of the general factor.

Table 2.5

Item factor loadings for NPV-J data

Item	Subdomain	UUM			NHMM/SOM			BFM			
		IN	SI	RE	IN	SI	RE	GE	IN	SI	RE
IN1	De	.508			.459			.340	.446		
IN2	An	.402			.386			.325	.251		
IN3	De	.476			.449			.341	.405		
IN4	De	.790			.766			.654	.444		
IN5	An	.378			.403			.384	.103		
IN6	De	.778			.771			.683	.380		
IN7	An	.384			.454			.518	-.175		
IN8	De	.867			.844			.701	.539		
IN9	An	.522			.461			.310	.555		
IN10	De	.578			.546			.426	.466		
IN11	De	.638			.674			.632	.224		
IN12	De	.635			.615			.517	.389		
IN13	An	.557			.594			.589	.076		
IN14	De	.584			.524			.366	.591		
IN15	De	.504			.518			.474	.211		
IN16	De	.810			.790			.673	.469		
IN17	An	.625			.676			.658	.138		
IN18	De	.578			.610			.634	-.018		
IN19	De	.810			.813			.752	.296		
IN20	An	.732			.761			.769	.067		
IN21	De	.514			.533			.516	.115		
IN22	An	.764			.757			.704	.261		
IN23	An	.587			.605			.590	.117		
IN24	De	.848			.844			.775	.329		
IN25	An	.729			.727			.660	.310		
IN26	An	.455			.455			.397	.259		
IN27	De	.520			.491			.410	.338		
IN28	De	.773			.789			.780	.127		
SI1	So		.544			.429		.220		.546	
SI2	In		.489			.603		.490		.226	
SI3	So		.748			.677		.426		.631	
SI4	In		.635			.616		.422		.471	
SI5	So		.721			.709		.493		.532	
SI6	So		.245			.204		.130		.186	
SI7	So		.593			.593		.416		.432	
SI8	So		.769			.693		.430		.656	
SI9	So		.666			.785		.623		.332	
SI10	So		.307			.408		.357		.083	
SI11	So		.770			.766		.535		.547	
SI12	So		.744			.702		.470		.572	
SI13	So		.622			.727		.578		.302	

* UUM = uncorrelated unidimensional model, NHMM = non-hierarchical multidimensional model, BFM = bifactor model, IN = inadequacy, SI = social inadequacy, RE = recalcitrance De = depression, An = anxiety, In = introversion, So = sociability

Table 2.5 (continued)

Item factor loadings for NPV-J data

Item	Subdomain	UUM			NHMM/SOM			BFM			
		IN	SI	RE	IN	SI	RE	GE	IN	SI	RE
RE1	Tr			.292			.366	.329			.104
RE2	Al			.479			.367	.249			.537
RE3	Al			.393			.175	.036			.602
RE4	Al			.318			.247	.170			.360
RE5	Tr			.474			.457	.384			.140
RE6	Al			.168			-.079	-.152			.416
RE7	Tr			.372			.286	.207			.337
RE8	Tr			.571			.554	.455			.327
RE9	Al			.455			.476	.399			.200
RE10	Tr			.485			.510	.447			.093
RE11	Tr			.272			.129	.025			.545
RE12	Al			.497			.492	.397			.305
RE13	Al			.368			.335	.267			.221
RE14	Al			.317			.161	.071			.403
RE15	Tr			.577			.596	.517			.138
RE16	Tr			.293			.154	.081			.338
RE17	Tr			.516			.472	.381			.289
RE18	Al			.659			.547	.404			.559
RE19	Tr			.734			.785	.674			.257
RE20	Tr			.369			.419	.373			.018
RE21	Tr			.676			.816	.706			.203
RE22	Al			.663			.754	.644			.260
RE23	Tr			.293			.175	.077			.535
RE24	Al			.627			.636	.513			.435

* UUM = uncorrelated unidimensional model, NHMM = non-hierarchical multidimensional model, BFM = bifactor model, IN = inadequacy, SI = social inadequacy, RE = recalcitrance, Tr = trust, Al = altruism

Both the non-hierarchical multidimensional model and the second-order model indicated that the original scales shared much variance, and thus a theoretically based general factor might be valid. However, a number of items had loadings below .35, which might indicate that they have higher loadings on an additional second dimension. As Table 2.5 shows, for the bifactor model, out of the 28 Inadequacy items there were only five items, mostly measuring Depression, which had higher loadings on the domain-specific Inadequacy scale than on the general factor. On the other hand there were 22 items, thirteen measuring Depression and nine measuring Anxiety, with a higher loading on the general factor than on the domain-specific Inadequacy factor. Only one item (Anxiety) had loadings below .35 on both general factor and domain-specific Inadequacy factor. Almost

80% of the Inadequacy items loaded on the general factor, which supports the high loadings of the first-order Inadequacy factor on the second-order factor under the second-order model. The items with both loadings on general factor and domain-specific Inadequacy factor or only loadings on the domain-specific Inadequacy factor were mostly items measuring Depression.

Of the thirteen Social Inadequacy items, eight items, mostly measuring Sociability, had higher loadings on the domain-specific Social Inadequacy factor than on the general factor. Four items, three measuring Sociability and one measuring Introversion, had higher loading on the general factor than on the domain-specific factor, and only one item had no loading of .35 or higher on the general or domain-specific Social Inadequacy factor. Although most items had higher loadings on the domain-specific Social Inadequacy factor than on the general factor, seven out of eight had a slightly lower loading, but still loadings $>$ .35 on the general factor as well. This is in accordance with the high loading of the first-order Social Inadequacy factor on the second-order factor under the second-order model. Because items of both the Inadequacy scale and the Social Inadequacy scale have acceptable loadings on the general factor, this may explain the high correlation between both scales under the non-hierarchical multidimensional model.

Twelve out of 24 Recalcitrance items, eight measuring Trust and four measuring Altruism, had higher loadings on the general factor than on the domain-specific Recalcitrance factor. Furthermore there were eight items with higher loadings on the domain-specific Recalcitrance factor than on the general factor, of which two items were Trust items and the other six items were Altruism items. Four items, three measuring Trust and one measuring Altruism, had loadings below .35 on both general factor and domain-specific Recalcitrance factor. Whereas both Altruism and Trust items had high loadings on the general factor, mainly Altruism items had high loadings on the domain-specific Recalcitrance factor. About 50% of the items of the Recalcitrance scale had loadings above .35 on the general factor. This explains the high loading of the first-order Recalcitrance factor on the second-order factor under the second-order model, but also the moderate correlation between the Recalcitrance factor and the Inadequacy factor, and the Recalcitrance factor and the Social Inadequacy factor under the non-hierarchical multidimensional model.

The general factor consists of items measuring Anxiety, Depression, Sociability, Introversion, Trust, and Altruism. This indicates there is a general factor that may measure Inadequate behavior. The Anxiety

and Depression items had higher loadings on the general factor than the Sociability, Introversion, Trust and Altruism items. The Inadequacy domain-specific group construct showed some additional information on Depression, the Social Inadequacy domain-specific group factor measured additional information on both Sociability and Introversion, and the Recalcitrance domain-specific group factor measured mainly Altruism. Furthermore, the bifactor solution found four Recalcitrance items, one Social Inadequacy item, and one Inadequacy item which had no acceptable loadings on the general or domain-specific group factor.

2.5.2 Scoring of persons on constructs

For all models, individuals' weighted factor scores (mean around zero and varying standard deviation) were estimated using MPlus. The factor scores were compared to simple sum scores on the scales. In both the second-order model and the bifactor model, a general factor score with a similar interpretation was estimated. Furthermore, the simple sum score over all 65 items was calculated for each individual in the sample. The simple sum score and the factor scores under the second-order model and the bifactor model were compared to check whether the ordering of persons is the same for the different techniques and models. The factor scores on the general factor for the second-order model and bifactor model correlated highly with the simple sum scores; correlations were equal to $r = .95$ between the sum score and the second-order model factor score, $r = .93$ between the sum score and the bifactor model factor score, and $r = .97$ between the second-order model factor score and the bifactor model factor score. Plotting the relations showed that the relation between simple sum score and factor scores of both models was slightly scattered around the diagonal line for persons scoring around the mean and below the mean (see Figure 2.2a upper and lower left panel). However, relations between the two factor scores were clustered along a diagonal line over the whole continuum (lower right panel). This indicates that the two factor scores led to the same ordering of individuals, whereas the simple sum score led to a different ordering for the low performing individuals, but not for the high performing individuals.

Furthermore, weighted factor scores and simple sum scores were determined for the domain-specific Inadequacy, Social Inadequacy, and Recalcitrance factors. The uncorrelated unidimensional model, the non-hierarchical multidimensional model and the second-order model investigated domain-specific factors, and the bifactor model investigated

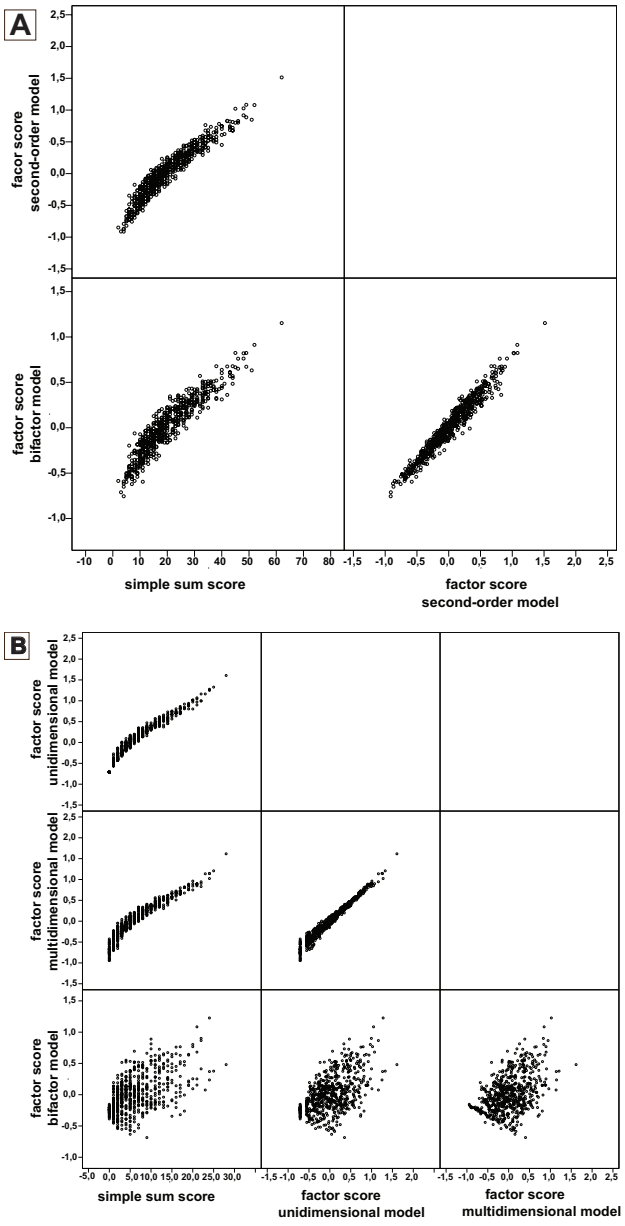

Figure 2.2. Scatter plot comparison of latent trait estimates a) between sum score estimates, second-order model factor score estimates and bifactor model score estimates for the general construct, b) between sum score estimates, uncorrelated unidimensional model factor score estimates, non-hierarchical multidimensional model/second-order model factor score estimates and bifactor model factor score estimates for the Inadequacy construct.

domain-specific factors after partialing out the general factor. When the general factor explains common variance, this might result in domain-specific factors that measure slightly different constructs as under the other models, whereas when the general factor explains almost no common variance the domain-specific factors will measure the same construct as under the other models. The factor scores on the domain-specific factor under the non-hierarchical multidimensional model, and the second-order model were equal resulting from equivalence of the models. Correlations between the uncorrelated unidimensional model factor scores, the non-hierarchical multidimensional model/second-order model factor scores, the bifactor model factor scores, and the simple sum scores on the domain-specific Inadequacy, Social Inadequacy and Recalcitrance scales are shown in Table 2.6. For all three constructs, Inadequacy, Social Inadequacy and Recalcitrance, estimated factor scores under the uncorrelated unidimensional model, estimated factor scores under the non-hierarchical multidimensional/second-order model and estimated simple sum scores were highly correlated. Plotting the relations between the three scores showed clustering of estimated scores along a diagonal line for the Social Inadequacy and Recalcitrance scales. For the Inadequacy scale this was only partly the case, as is shown in Figure 2.2b in the upper and middle row. For persons around and below the mean score value, estimates scattered around the diagonal line, which indicated a different ordering of persons when using different models or scoring techniques. Correlations between the estimated bifactor model factor scores and the simple sum scores, factor scores under the uncorrelated unidimensional model, and factor scores under the non-hierarchical multidimensional/second-order model were moderate. Although plots of the Social Inadequacy and Recalcitrance scales form a broad-banded line, for the Inadequacy scale this was not the case (see Figure 2.2b; lower row); estimates scattered around the diagonal line. This indicates that group factors under the bifactor model did measure a slightly different construct than domain-specific factors under an uncorrelated unidimensional model, or a non-hierarchical multidimensional or second-order model.

2.6 Discussion

The appropriateness of the non-hierarchical multidimensional model, the second-order model, and the bifactor model was investigated. With respect to the NPV-J, the bifactor model fitted best, and a multidimensional factor

Table 2.6

Correlations between simple sum scores and weighted factor scores on domain-specific factors under all models

Inadequacy scale	Simple sum score	Factor scores		
		UUM	NHMM/ SOM	BFM
Simple sum score	1.000	0.965	0.946	0.553
Factor score UUM		1.000	0.984	0.541
Factor score NHMM/SOM			1.000	0.449
Factor score BFM				1.000

Social Inadequacy scale	Simple sum score	Factor scores		
		UUM	NHMM/ SOM	BFM
Simple sum score	1.000	0.980	0.966	0.761
Factor score UUM		1.000	0.977	0.807
Factor score NHMM/SOM			1.000	0.673
Factor score BFM				1.000

Recalcitrance scale	Simple sum score	Factor scores		
		UUM	NHMM/ SOM	BFM
Simple sum score	1.000	0.967	0.874	0.747
Factor score UUM		1.000	0.940	0.653
Factor score NHMM/SOM			1.000	0.415
Factor score BFM				1.000

* UUM = uncorrelated unidimensional model, NHMM = non-hierarchical multi-dimensional model, SOM = second-order model, BFM = bifactor model

structure with both general and domain-specific constructs was found. The general Inadequate Behavior factor was strong, and, as expected, consisted of Inadequacy items, Social Inadequacy items and Recalcitrance items. Loadings of Inadequacy items were stronger on the general factor than loadings of Social Inadequacy items and Recalcitrance items. Under the non-hierarchical multidimensional model and the second-order model a lot of shared variance between the scales was found, which also indicated the relevance of a general factor. Further, the bifactor model showed

that some Depression items, most Social Inadequacy items, and most Altruism items shared additional variance over and above the general factor, and formed three domain-specific factors. It can be concluded that part of the Inadequacy items were multidimensional, while others were unidimensional measuring the general factor. Social Inadequacy items were mostly multidimensional and Recalcitrance items were mostly unidimensional, measuring the general factor or the domain-specific group factor.

Finding the best fitting and most appropriate model for the data was not only important for decisions on the dimensionality structure of the model and its interpretation. Sum scores and factor scores under different models might not always lead to the same ordering of individuals on constructs, as was found for the general factor, and the domain-specific Inadequacy factor. Misspecification of the model may have serious consequences for the ordering of persons. As a consequence it may effect conclusions in a diagnostic, classification or selection context.

Finally, the bifactor model is the most general model for analyzing and constructing psychological instruments consisting of two or more related constructs that might measure a more general and theoretically interpretable construct, and some additional domain-specific constructs. The bifactor model gives clear results on the dimensionality structure of the instrument, the dimensionality of items, the interpretation of both general and domain-specific factors, and the scoring of individuals. When there is only a strong general factor and there are less important domain-specific factors the second-order model provides similar information. When there is no significant general factor, but only significant domain-specific factors the non-hierarchical multidimensional model is a good alternative. The bifactor model gives a statistically based conclusion about the appropriateness of the non-hierarchical multidimensional model and the second-order model, even in case of two or three domain-specific factors.

Chapter 3

Analyzing the dimensionality of the Students' Conceptions of Assessment inventory

3.1 Introduction

Assessment plays an important role in contemporary life, having meaningful consequences in education and employment contexts. Scores, as derived from tests, assessments, and evaluations, influence a person's future. While intelligence, socio-economic status, and cultural factors are known to contribute to such scores, the role of personal beliefs and attitudes in determining test scores is less well understood. Ajzen's (1991) theory of planned behavior claims that behavioral outcomes are predicted by individuals' intentions, beliefs about likely consequences, normative expectations of others, and by perceptions of behavioral control (self-efficacy beliefs). Thus, the reasons, opinions, attitudes, beliefs, and intentions people have influence their behavioral achievement.

In education, academic achievement is influenced by students' learning and study behavior. In line with the theory of planned behavior, Entwistle (1991) discussed that learning, studying, and academic achievement are influenced by both external factors (e.g., the learning environment and context) and interactions between students and their context. Students'

This chapter will be published as Weekers, A. M., Brown, G. T. L., & Veldkamp, B. P. (2009). In D. M. McInerney, G. T. L. Brown, & G. A. D. Liem (Eds.), *Student perspectives on assessment: What students can tell us about assessment for learning.* Charlotte, NC: Information Age Publishing.

perceptions of the learning environment and context and their intentions when approaching a task have additional value to outcomes as will be discussed below.

Since educational assessment has significant consequences for learners (i.e., it can be used to monitor, motivate, and certify learning), students' perceptions of assessment seem to matter. Research has shown that assessment influences students' behaviors, learning, studying, and achievement (Entwistle, 1991; Peterson & Irving, 2008; Struyven, Dochy, & Janssens, 2005). Variation in how students perceive, understand, and evaluate assessment has been investigated internationally.

Struyven, Dochy, and Janssens (2005) reported that university-level students had multiple perceptions of assessment (i.e., it was inaccurate, inappropriate, arbitrary, unfair, and irrelevant; enjoyable and beneficial; a way to improve learning; a way to demonstrate personal growth; and a way to achieve greater quality in learning). To ascertain important aspects of high school students' attitudes and beliefs about assessment, a series of inventory survey studies have been conducted in New Zealand (Brown, 2006; Brown & Hirschfeld, 2005, 2007, 2008; Hirschfeld & Brown, 2009).

The fifth version of the Students' Conceptions of Assessment (SCoA-V) inventory (Brown, Irving, Peterson, & Hirschfeld, 2009) is studied in this chapter because this version was validated on a large representative sample of New Zealand high school students and used in a subsequent study that related student conceptions of assessment to academic outcomes (Brown, Irving, & Peterson, 2008). The SCoA-V inventory measures four major inter-correlated constructs. Brown, Irving, and Peterson (2008) suggested that the inter-correlations between the major conceptions might indicate a more general student conception of assessment and that alternative models of how students' conceptions of assessment are structured needed to be investigated. The dimensionality structure of the Students' Conceptions of Assessment (SCoA-V) instrument is studied in this chapter.

The chapter is organized into three main sections. First, the background information of the inventory is addressed. Second, different models to investigate the dimensionality of students' perceptions of assessment are offered. Third, the dimensionality structure of the SCoA-V is evaluated, using two alternative measurement model structures (i.e., non-hierarchical multidimensional and bifactoral), which results in recommendations concerning students' conceptions of assessment.

3.2 Students' Conceptions of Assessment

In order to situate this research, it is necessary to review briefly both the international literature on students' conceptions of assessment and the development of the Student Conceptions' of Assessment inventory. In a sense, this requires us to go back in time to what we knew about students' thinking about assessment before the New Zealand series of survey studies was conducted. Hence, the section draws heavily on previous reviews of the literature, reported in Brown (2008) and Brown and Hirschfeld (2007, 2008). At the same time, however, this review is able to bring in findings from the New Zealand survey studies. The motivation for this research, as touched on in Brown and Hirschfeld (2008), was the suggestion that teachers' conceptions of assessment may have their origins in the belief systems of students (Pajares, 1992). Hence, it seemed logical to consider the possibility that secondary school students would have similar ways of conceiving of assessment as teachers. Thus, much of Brown's research has been guided by the idea that there would be some similarity between how students and teachers conceived of assessment.

The research literature on students' conceptions of assessment is not vast (e.g., Harlen (2007) devotes$1\frac{1}{2}$ pages to the topic) and is largely focused on tertiary or higher education students (see Struyven, Dochy, & Janssens, 2005 for a review). Our analysis of the previously reported empirical studies into how students understand the purposes of assessment has identified, four major purposes, some of which are similar to teachers' conceptions of assessment (Brown, 2004a). First and foremost, students are aware that assessment exists in order to improve learning and teaching. Second, students are aware that assessment is used to evaluate external factors outside their own control such as the quality of their schools, their intelligence, and their future. Thirdly, the literature clearly indicates that students are aware of an affective purpose for assessment- assessment impacts on their emotional well-being and the quality of relationships they have with other students. Finally, students are aware that assessment can be an unfair, negative, or irrelevant process in their lives. To summarize, these purposes can be expressed as simply as (a) improvement, (b) externality, (c) affect, and (d) irrelevance.

3.2.1 Improvement

Students in the compulsory school sector (K-12) want assessment to lead to improved learning (Peterson & Irving, 2008). Good teachers regularly

test and provide feedback to students about learning (Olsen & Moore, 1984) and do not hide from students uncomfortable messages about the need to improve or the processes of improvement (Pajares & Graham, 1998). Information that increases students' sense of personal agency, in terms of knowing how to earn grades that accurately describe their abilities, is sought (Stralberg, 2006). It would appear that students do not make an artificial distinction between summative and formative assessments; rather, it would appear that many see all tests and evaluations as a source of information about how they can improve (Peterson & Irving, 2008). Indeed, improvement goes beyond informing the student; it also involves teachers use of assessment so that students can improve (Peterson & Irving, 2008).

While some readers may be disturbed or concerned that students see assessment in instrumental terms of higher assessment grades or scores, it is an unavoidable fact of society that we assess students in order to discover how much or how well they have learnt. And students are clearly aware of this process. A number of studies have reported that students in Israel (Zeidner, 1992), the United States (Brookhart & Bronowicz, 2003), and the UK (Reay & Wiliam, 1999) are aware that assessment is used to judge or evaluate student learning. Harlen (2007) suggests that higher-attaining students tend to associate assessment with improvement.

The New Zealand survey studies have found that the use of assessment to hold students accountable is linked to the notion of improvement. Using version 1 of the SCoA inventory, Brown and Hirschfeld (2007) found that a group of items related to assessment as a self-regulatory feedback and motivational process predicted greater performance in mathematics. Further, Brown and Hirschfeld (2008), using 11 items from version 2 of the SCoA inventory, showed that the conception of student accountability predicted positively students scores on a reading comprehension test. Later versions of the inventory, as the number of items and factors was increased, clearly embedded the notion that assessment makes students accountable with the notion of student self-regulation (Brown, Irving, & Peterson, 2008). This self-regulation conception of improvement was also linked strongly to the idea teachers use assessment to improve their teaching of students (Brown, Irving, Peterson, & Hirschfeld, 2009) and together these improvement oriented conceptions of assessment predicted higher test scores (Brown, Irving, & Peterson, 2008).

Hence, it seems feasible to conclude that students are aware that a major purpose of assessment is to lead to improved teaching and improved learning, which in turn leads to improved assessment

scores. Notwithstanding the instrumental nature of this relationship, it seems logical that students should understand improvement in terms of assessments that are used to make decisions such as certification, promotion, retention, awards, and so on.

3.2.2 Externality

The beliefs students have about where control lies have an important relationship to assessment. Students who attribute academic consequences (i.e., assessment outcomes) to external (e.g., my teacher or my school), unstable (e.g., luck or teacher whimsy), or uncontrollable (e.g., my parent's wealth or my intelligence) causes consistently do worse (Schunk & Zimmerman, 2006). Likewise, students who believe that the locus of control lies outside their personal control do worse academically (Rotter, 1982). Thus, it seems logical to infer that, if the purpose of assessment is focused on an attribute external to the student (e.g., evaluation of the school), student performance will be negatively impacted.

Thus, the question arises as to whether students are aware that assessments have a strong external component. Peterson and Irving (2008) conducted a series of focus group studies with New Zealand high school students and found considerable evidence for an external component in students' understandings of assessment. For example, they reported (p. 244) that "several students ascribed their poor grades to the teacher "being mean" or the teacher "doesn't like me". In the same study, students in middle and higher socio-economic communities indicated that assessment was primarily for their parents who may punish them for unacceptable grades. Similarly, the New Zealand high school students believed assessment cast a shadow over their personal futures; grades are used by future employers, and may help them avoid bad jobs. This study specifically inspired, the development of items around the externality in the SCoA inventory.

The national survey of secondary school students (Brown, Irving, Peterson, & Hirschfeld, 2009) found that assessment as a measure of school quality and of external factors such as intelligence, parents, and jobs were highly related. Furthermore, the students gave nearly moderate levels of agreement towards these two factors. The subsequent study (Brown, Irving, & Peterson, 2008) reported a very similar structure of beliefs concerning externality and, more importantly, showed that the joint external factors had negative impact on academic performance in mathematics. This result is consistent with the research on control beliefs, such as attribution

theory (Weiner, 1985) and the self-determination theory (Ryan, Connell, & Deci). It would appear that the more students believe that the purpose of assessment is related to external factors outside their control the worse they do in school.

3.2.3 Affect

A matter of great concern to educators has been the emotional impact of assessment on students (Linn & Gronlund, 2000). This is motivated partly by concern for the well-being of young people (e.g., Weeden, Winter, & Broadfoot, 2002) and the validity of test scores-too much anxiety or distress is bad for children and invalidates interpretations based on sub-optimal student performance. Younger students appear to enjoy a wide variety of assessment methods (Atkinson, 2003). Even relatively formal, though low-stakes, paper-and-pencil tests have been seen by high school students as being enjoyable (Hattie, Brown, Ward, Irving, & Keegan, 2006). Harlen (2007, p. 42) suggests that students positively evaluate tests because tests give "clear-cut measures of progress based on 'right or wrong"'. Perhaps students prefer the system of assessment that they experience, regardless of the merits or deficiencies of that system (Blaikie, Schönau, & Steers, 2004). Nonetheless, there is some consensus that as students progress through schooling they become increasingly disaffected by assessment (Harlen, 2007; Moni, van Kraayenoord, & Baker, 2002).

Brown and Hirschfeld (2007) reported that a small sample of New Zealand high school students had a moderate level of agreement that assessment could be enjoyable and improve the social climate of the class. Brown and Hirschfeld's (2008) survey with version 2 of the SCoA found that a large ($N = 3504$) sample of students had less than slight agreement towards the same factor. A national survey of 700 New Zealand secondary school students using version 5 of the SCoA inventory with an extended number of items for both classroom benefit and personal enjoyment, found that the two factors were strongly inter-correlated and elicited slight to somewhat negative agreement (Brown, Irving, Peterson, & Hirschfeld, 2009). In all the New Zealand studies, the affect factors had negative relations towards scores of mathematics or reading comprehension. In other words, the more students enjoyed assessment or the more they believed assessment improved classroom relations, the worse they did academically. Hence, it could be argued increased positive emotion towards assessment is a counter-productive purpose for assessment.

3.2.4 Irrelevance

Assessment may be considered irrelevant by students if they think of it as being bad or unfair or tainted with teacher subjectivity. Furthermore, many students, especially lower-performing ones, disregard or ignore assessment results. A small group of Australian high school students were negative towards assessment because of the volume of assessment and because they perceived teacher's decisions as subjective (Moni, van Kraayenoord, & Baker, 2002). A group of urban African American and Latino high school seniors perceived the high-stakes university entrance tests they were about to undertake as unfair because of their impact upon student life chances (Walpole, McDonough, Bauer, Gibson, Kanyi, & Toliver, 2005). Peterson and Irving (2008) reported that the New Zealand high school students saw assessment as irrelevant if their career aspirations did not require academic success. The same students also insisted that assessments were unfair unless "you had to complete it by yourself, under controlled conditions, without any assistance or second chance" (Peterson & Irving, 2008, p. 245). A consequence of this last belief was that peer assessment, specifically, was considered irrelevant since interpersonal factors contaminated the validity of peer feedback.

The SCoA inventory studies have consistently identified factors related to the negative nature of assessment. In version 1, the conception that assessment interfered with learning was not agreed with; however, increased agreement had a negative impact on mathematics performance (Brown & Hirschfeld, 2007). In version 2, students again rejected the notion that they ignored assessment and increased agreement had a negative impact on reading performance (Brown & Hirschfeld, 2008). In version 5, the national sample rejected the conception of ignoring assessment and treating it as irrelevant (Brown, Irving, Peterson, & Hirschfeld, 2009). Interestingly, the more students agreed with ignoring assessment, the more likely they were to think of informal-interactive assessment practices (i.e., self assessment, peer assessment, portfolios, etc.); a result which was replicated in the follow-up study (Brown, Irving, & Peterson, 2008).

Thus, it would appear that students are quite sensitive to assessments which they perceive to be unfair, bad, or irrelevant to them. The irrelevance of assessment does not appear to be a permanent attitude; rather it appears to be a response to the appearance of subjectivity, disparity, and inequity. The New Zealand survey studies consistently report that students reject this conception and it has no direct relationship to academic performance. Nonetheless, students appear to consider some kinds of assessment as

worthy of being ignored; a result which should be of some concern to those advocating a reduced emphasis on examinations or tests in educational assessment.

3.3 Background to the Students' Conceptions of Assessment Inventory

The Students' Conceptions of Assessment (SCoA) inventory is a self-rating instrument in which high school students indicate the extent to which they agree or disagree with statements about the purposes of assessment. Brown and Hirschfeld (2007) trialed the first version of the inventory (SCoA-I) in 2003 as four independent parts to mitigate potential participant fatigue. Four purposes of assessment were identified (i.e., "assessment makes schools and students accountable", "assessment improves teaching and learning", "assessment is negative or bad", and "assessment is useful").

With the second version of the inventory (SCoA-II), four conceptions (i.e., "assessment makes schools accountable", "assessment makes students accountable", "assessment is fun", and "assessment is ignored") were estimated simultaneously in a survey conducted in 2004 with nearly 3500 high school students (Brown & Hirschfeld, 2008). Three conceptions were correlated highly with each other, while the "assessment is ignored" conception was weakly and negatively correlated with the same three conceptions. In an invariance study of the SCoA-II inventory, Hirschfeld and Brown (2009) concluded that the instrument had invariant measurement properties across sex, year level, and ethnicity.

In further extending the meaning of students' conceptions of assessment with two progressively more complete inventories (SCoA-III: Brown & Hirschfeld, 2005; SCoA-IV: Brown, 2006) students were asked to also indicate what types of assessment practices they associated with the term 'assessment'. Two major classes of assessment types were found (i.e., teacher-controlled test-like assessments and informal-interactive assessments). In the SCoA-IV, six inter-correlated conceptions of assessment were found (i.e., "assessment makes students accountable", "I use assessment", "teachers use assessment", "the public uses assessment", "assessment is fun", and "assessment is irrelevant). All conceptions, except "assessment is irrelevant", had weak correlations with the interactive assessment type. Furthermore, five of the conceptions were positively inter-correlated, while the "assessment is irrelevant" conception was weakly and negatively correlated with all the other conceptions.

In a national survey of high school students conducted in 2006, the fifth version of the inventory (SCoA-V) was used to establish the structure of student conceptions of assessment and their relations to assessment type (Brown, Irving, Peterson, & Hirschfeld, 2009). Four major 2^{nd}-order conceptions were found (i.e., "assessment improves learning", "assessment makes students and schools accountable", "assessment is beneficial", and "assessment is irrelevant"). Three of the four major conceptions were strongly and positively inter-correlated, while "assessment is irrelevant" was again weakly (indeed, not statistically different to zero for path to "assessment is beneficial") and negatively related to those three conceptions. The conception "accountability/external factors" measures lack of personal autonomy or control, divided into the degree to which assessment measures a fixed personal future or it measures school quality. The conception "affect/benefit" measures the affective or emotional impact of assessment and consists of assessment as a personally enjoyable experience and assessment as a benefit to the class environment. The conception "improvement" indicates that the goal of assessment is to improve students' own use of assessment to improve learning and teachers' use to improve teaching. The conception "irrelevance" measures a negative evaluation of assessment because it is seen as bad, subjective, or unfair and whether it is tolerated but ignored. Each conception was divided into two 1^{st}-order sub-conceptions which were used in a structural model to establish relations to assessment types. There was one additional pathway from "assessment is irrelevant" to the 2^{nd}-order conception "personal enjoyment".

Most recently, the SCoA-V inventory was used in 2007 to investigate the beliefs of three cohorts of high school students in relation to their definitions of assessment and their performance in mathematics (Brown, Irving, & Peterson, 2008). This study developed a sixth version of the inventory (SCoA-VI) by revising the measurement model only. In the SCoA-VI the items are identical to SCoA-V but all four 2^{nd}-order conceptions were inter-correlated and the pathways from the 2^{nd}-order conceptions to the 1^{st}-order conceptions were simplified. The pathway from "assessment is ignored" to "personal enjoyment" was removed to attain structural simplicity and, as a consequence, the 1^{st}-order sub-conception "assessment is ignored" generated negative error variance. Hence, all the items were given paths directly to the 2^{nd}-order conception "assessment is irrelevant". This revised solution was configurally invariant for both SCoA-V and SCoA-VI samples. Again the inter-correlations were strongly positive for three of the 2^{nd}-order conceptions, while the arcs from "assessment is irrelevant" were negative

and weak to moderate.

3.4 Dimensionality of the SCoA inventory

Research into students' conceptions of assessment has identified four major categories of students' thinking concerning the purposes of assessment. There is also strong evidence that students' conceptions influence considerably their academic performance (explained variance in test scores ranges from 20 to 25%). Further, there is evidence that these conceptions are meaningfully aligned with theories of self-regulation, self-determination (Ryan, Connell, & Deci, 1985), and attribution (Weiner, 1985) in explaining how the beliefs translate into outcomes. Furthermore, research with the SCoA inventory has reported stable results with multiple samples of New Zealand high school students.

The SCoA inventory, in its current form, consists of multiple inter-correlated factors with hierarchical structure containing eight 1^{st}-order factors, and four correlated 2^{nd}-order factors. This structure is a consequence of the developmental process-measures for each sub-conception were developed within a conceptual framework that the sub-conceptions were members of four higher-order structures. Further, one conception (i.e., "assessment is irrelevant") is always weakly and negatively correlated with the other conceptions. However, independent examination of the dimensionality structure of the SCoA inventory has not been carried out.

The highly positive inter-correlations between the major conceptions "external factors", "positive affect", and "improvement", and the more variable, but in general moderate negative inter-correlations with "irrelevance" suggested some overlap between the conceptions (Brown, Irving, Peterson, & Hirschfeld, 2009). Conceptually, the three conceptions "external factors", "positive affect", and "improvement" all suggest a positive and important influence of assessment on the studying and learning context of students. Students might experience that assessment in general leads to positive feelings and a positive environment. The conception "irrelevance" might suggest unimportant, invaluable or no impact on the study and learning context. As such, "Irrelevance" measures a negative contrary conception of the other three conceptions. Students might interpret the four conceptions as four separate constructs that share some information, however another option is that students see just the positive impact on their studying and learning context in general, which might be expressed in a general conception. Besides the in general positive impact it

might be possible that some items give specific information about reasons or situations when assessment has positive influences, like that assessment "improves learning", "provides positive external factors", "is beneficial" or "is not irrelevant".

The estimated model in Brown et al. (2009), shows reasonable to high correlations between the four 2^{nd}-order conceptions. This might methodologically suggest a general students' conception of assessment. When there are high correlations among constructs it is possible that a general factor is dominating item responses (Chen, West, & Sousa, 2006; Reise, Morizot, & Hays, 2007; Yung, Thissen, & McLeod, 1999). Although coherence between the four major conceptions and the multiple studies existed, no further research has been done to investigate the relations between the conceptions. It is important to investigate which model is most appropriate to describe the inventory, because different models might result in different interpretations of the inventory, and persons might be ordered on the continuum differently when using other models or scoring techniques (see Chapter 2).

Relations between constructs can be modelled in two ways; as non-hierarchical structures, and as hierarchical structures. Non-hierarchical structures describe constructs that are on the same level, and of the same order. The constructs might be correlated. The most common non-hierarchical structures are measurement models, which describe different constructs each measuring a separate concept. These constructs might be correlated (non-hierarchical multidimensional model) or uncorrelated (uncorrelated unidimensional model). For example consider an instrument that measures the five factors of the Big Five. There are five separate constructs of the same order and these are on the same level, and can be modelled as correlated or not correlated. Hierarchical structures consist of both constructs that are more general, and constructs that are less general. The general and specific constructs might be on the same level, but can be on different levels as well. The most commonly used hierarchical models are the higher-order model and the bifactor model. The higher-order model is a structural model that describes constructs on different measurement levels, namely domain-specific constructs and general constructs. The domain-specific constructs are on the same level and predict items. The higher-order general construct is on a different, more general, level and does not predict items directly, but does predict the domain-specific constructs, via which items are predicted. For example an intelligence test that measures verbal intelligence and spatial ability, but also measures the higher-order

construct general intelligence that predicts both verbal intelligence and spatial ability. The bifactor model is a measurement model consisting of both general and domain-specific constructs, which are on the same level. Under this model items are multidimensional, because they are predicted by both domain-specific and general factors. The general construct of the bifactor model has a similar interpretation as the general construct of the higher-order model. An example is an inventory that consists of items about motivation in different contexts (i.e. school, sports). The items measure both motivation in general (general construct) and motivation in the domain-specific context (e.g., school and sports).

In this chapter two models will be used to analyze the SCoA-V data; the non-hierarchical multidimensional model and the bifactor model. There are four reasons for choosing these two models. First, both types of modeling relations, non-hierarchical and hierarchical modeling, are investigated. Second, both models are measurement models, and thus the different conceptions predict the items directly, and not via other constructs as is the case in for example higher-order models. Third, there is sufficient information that the conceptions are related; therefore, the non-hierarchical multidimensional model is estimated instead of the uncorrelated unidimensional model. And fourth, the bifactor model is more general than the higher-order model, in the sense that the higher-order model only estimates a general conception, via domain-specific conceptions, whereas the bifactor model both estimates a general conception, and domain-specific conceptions that provide extra information besides the general factor. The models used, the interpretation of the models, and scoring of persons on the models' conceptions, will be explained in more detail below.

3.5 Method

3.5.1 Instrument

The fifth version of the Students' Conceptions of Assessment inventory (SCoA-V) consists of 33 items (for item content, see Appendix) selected on the basis of content and factor analytic studies. The items measure four major conceptions, "external factors" [6 items], "affect/benefit" [8 items], "improvement" [11 items], and "irrelevance" [8 items].

Students indicated how much they agreed with the 33 statements on a six-point positively-packed agreement rating scale, with two negative

responses (strongly disagree, mostly disagree), and four positive responses (slightly agree, moderately agree, mostly agree, strongly agree). This response format was chosen, because students were inclined to respond positively to all items, and positively-packed rating scales generate greater variance and precision (Brown, 2004b; Klockars & Yamagishi, 1988; Lam & Klockars, 1982).

3.5.2 Participants and Procedure

Respondents were 705 students from 31 secondary schools in New Zealand, enrolled in Years 9 and 10 (the first two years) of secondary school. About half of them were boys (n = 342; 48.5%) and students were between 13 and 15 years of age, with a mean age of 14.14 (SD = .96). The students filled out the inventory about their conceptions of assessment during a single lesson supervised by a teacher.

3.5.3 Analyses

This study investigates the dimensionality structure of the SCoA by comparing two measurement models (i.e., non-hierarchical multi-dimensional model, and bifactor model). The uncorrelated unidimensional model (Figure 3.1a) is used as a baseline against which the non-hierarchical multidimensional model (Figure 3.1b) and the bifactor model (Figure 3.1c) are compared. The expectation is that the uncorrelated unidimensional model does not fit the data. Because some complex models cannot be estimated when the number of items loading on a factor is low (i.e., sets of items smaller than 3), and because the 1^{st}-order factors of each conception were highly correlated, all analyses were conducted without the 1^{st}-order factors of the model by Brown, Irving, Peterson, and Hirschfeld (2009). In other words, all items were treated as being directly predicted by the four major conceptions of assessment reported by the SCoA inventory.

The uncorrelated unidimensional model consists of four separate unidimensional latent factors. Each latent factor in the model measures one major conception; "external factors", "affect", "improvement", or "irrelevance" and each item is predicted by one conception only. Correlations between the conceptions were set to zero, but the latent factors were simultaneously estimated. The interpretation of this model is that students' conceptions of assessment can be described best by four separate conceptions, which do not have any shared variance or overlap. The persons can be ordered on each of the four conceptions based on a

weighted sum score, in which the factor loadings could be used as weights. This model is not likely to fit well because there is considerable evidence in the development of the SCoA that the conceptions are inter-correlated.

Second, the non-hierarchical multidimensional model was investigated. The four major conceptions "external factors", "affect", "improvement", and "irrelevance" were measured by four latent factors, which were allowed to correlate with each other. All items measure one conception only. Although it might be reasonable to expect that some factors have zero correlations with some other factors, a full non-hierarchical multidimensional model was investigated. In a full non-hierarchical multidimensional model it is hypothesized that each factor is non-zero correlated with every other factor in that measure. This model is similar to the models reported on the SCoA-V and SCoA-VI, except for that the eight sub-conceptions are not in the model, for reasons outlined earlier. The interpretation of this model would be that students' conceptions can be described best by four separate conceptions, which are related to each other. The different conceptions share information. For scoring the persons on one of the factors, both the weights of the items and the correlations between the factors have to be taken into account.

Third, the bifactor model (Figure 2c) was investigated. The bifactor model specifies both general and (domain-specific) group factors. The general factor is an overall measure of students' conceptions of assessment. All 33 items are predicted by this general factor. The general factor explains the common variance between items of different conceptions, and explains the item inter-correlations of all items. The four group factors (i.e., "external factors", "affect", "improvement", and "irrelevance") are additional to the general factor, and measure the shared variance between items of the same conception after partialing out the general factor. The four group factors, thus, measure what is left of the four different conceptions, after controlling for the general factor. All items have two loadings; one loading on the general factor and one loading on the group factor. Because items are predicted by two latent factors they are assumed to be multidimensional. Correlations between the general factor and the group factors were fixed at zero, so that the five factors were independent of each other. If there are meaningful item factor loadings on both general factor and group factors, the interpretation of the model would be that students' conceptions can be described best by five conceptions, one general student conception and four domain-specific conceptions. Scoring of persons is done on both general and group factors by factor scores, which

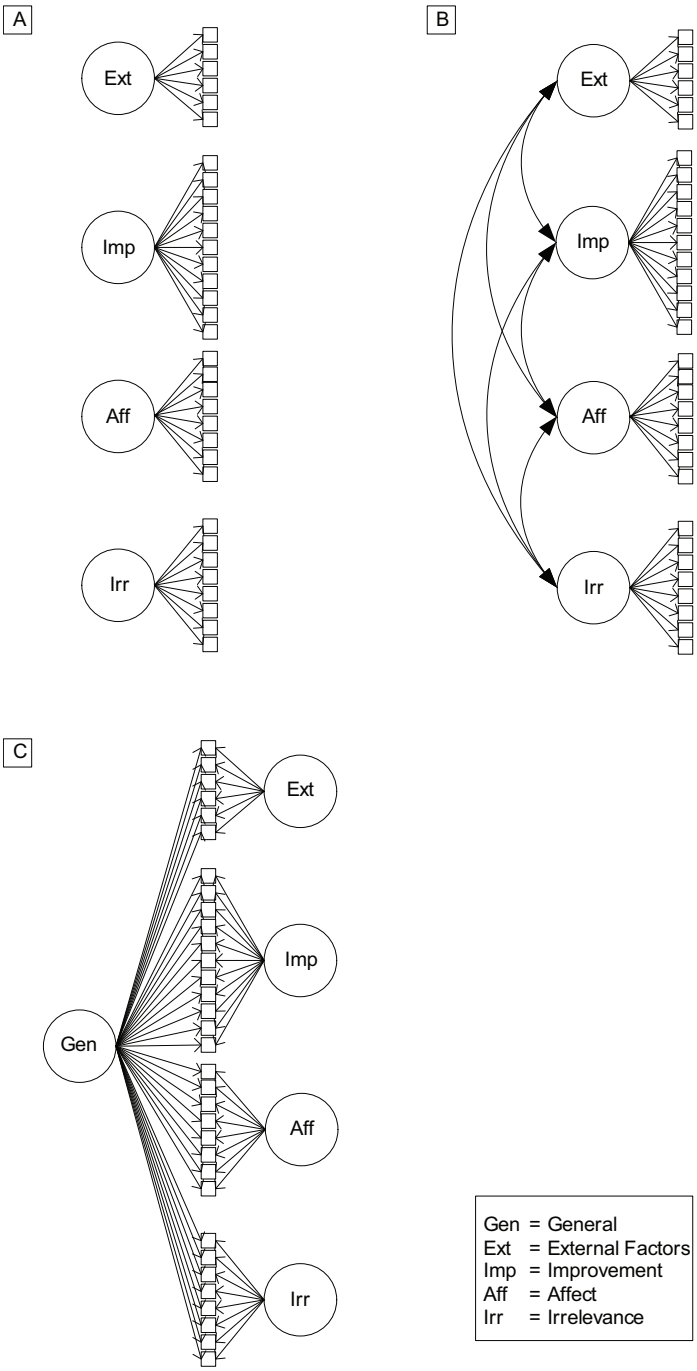

Figure 3.1. Models analyzed in this study: a) uncorrelated unidimensional model, b) non-hierarchical multidimensional model, and c) bifactor model.

indicate the position of a person on the general conception, and a position on the four domain-specific conceptions. The interpretation of the four domain-specific conceptions is different from the interpretation of the four conceptions under the non-hierarchical multidimensional model. The four conceptions under the bifactor model describe the position of a person on a scale that is more restricted than the scales under the non-hierarchical multidimensional model. In a sense the scores are dependent on the general factor, because they measure what is explained besides the variance on the general factor. If the loadings on the general or the group factors are low, the interpretation will be different. If only significant loadings on the general factor are found, only the general factor describes the data, and persons can be ordered on the general factor only. If only significant loadings on domain-specific group factors are found, the interpretation will be similar to the interpretation under the non-hierarchical multidimensional model.

MPlus (Muthén & Muthén, 1998-2006) was used to estimate all models using Weighted Least Squares Mean Adjusted (WLSM) estimation derived from polychoric correlations. This is different from previous studies which were conducted using AMOS (Arbuckle, 2007) using maximum likelihood estimation derived from Pearson product moment correlations. The general recommendation analyzing categorical responses is to use polychoric correlations (Jöreskog, 2007). However, it is not necessary to assume that the ordered response options (i.e., strongly disagree, mostly disagree, slightly agree, moderately agree, mostly agree, and strongly agree) used in this inventory represent discrete categories of response. Lam and Klockars (1982) showed through a scaling study on the desirability of terms that the terms *fair, good, very good*, and *excellent* were equally spaced from each other and that respondents used the rating anchors in responding to items. Klockars and Yamagishi (1988) reported that the scale distance between *fair* and *good* is constant whether the words are used in balanced or packed rating scales. Hattie (personal communication, February, 1999) reported unpublished research (similar in method to that of Lam & Klockars, 1982) which indicated that the following adverbs would provide nearly-equal intervals on an underlying scale of agreement (i.e., *strongly, mostly, moderately*, and *slightly*). Respondents tend to treat adverbs as symmetrical when used in positive and negative sides of neutral; hence, strongly and mostly would have equivalent values when applied to agree and disagree (Smith, Mohler, Harkness, & Onodera, 2005). Thus, there is support for the analysis of the six points used in this rating scale

as points on a continuous, rather than ordinal scale. If this assertion is accepted, then the use of the Pearson product moment correlations is legitimate. Nonetheless, in this study, the more conservative approach (polychoric correlations) is taken, and it should be noted that the quality of fit is likely to be negatively impacted by this approach.

Given the large number of participants and the complexity of models being evaluated, it is important to select appropriate fit statistics and cut-off values. Fan and Sivo (2007) have demonstrated that the standardized root mean squared residual (SRMR) and gamma hat statistics are most resistant to sample size, model complexity, and model misspecification (an issue being directly tested here). Nonetheless, multiple statistics (i.e., χ^2, comparative fit index (CFI), Tucker-Lewis index (TLI), and root mean squared error of approximation (RMSEA)) are reported in accordance with best practice (Fan & Sivo, 2005). Cutoff criteria are conventionally set at .95 for CFI, TLI and gamma hat, .08 for SRMR, and .06 for RMSEA (Hu & Bentler, 1999); though Marsh, Hau, and Wen (2004) have argued that goodness-of-fit values $>$ .90 indicate adequate model fit. Furthermore, factor loadings, correlations and residual variance were studied. Standardized regression weights $\lambda \geq .35$ are considered an adequate indicator that the item is well predicted by the latent trait (Stevens, 1992). When items have a loading of $\geq .35$ on only one factor they are considered unidimensional, and when item loadings are $\geq .35$ on two factors they are treated as multidimensional.

3.6 Results

Measurement model fit statistics are shown in Table 3.1 and item factor loadings are shown in Table 3.2.

The uncorrelated unidimensional model did not fit the data, since values for fit statistics did not approach the cutoff criteria. The non-hierarchical multidimensional model and the bifactor model had mixed-message fit values. The CFI and TLI values were $>$ .90 and SRMR values were $<$ .08. However, the more robust gamma hat values were considerably below the cutoff criteria and the RMSEA values were above even Steiger's (2000) generous cutoff criterion of .10. Of these two models, the non-hierarchical multidimensional model fit statistics were closest to the cutoff criteria.

Table 3.1
Fit statistics Students' Conceptions of Assessment data

Model	χ^2(df)	χ^2/df p-value	CFI	TLI	RMSEA	SRMR	$\hat{\gamma}$
UUM	27750.53 (495)	56.06 <.01	.60	.57	.28	.22	.30
NHMM	4592.01 (489)	9.39 <.01	.94	.94	.11	.07	.74
BFM	5167.21 (462)	11.18 <.01	.93	.92	.12	.07	.71

* UUM = uncorrelated unidimensional model, NHMM = non-hierarchical multidimensional model, BFM = bifactor model

3.6.1 Baseline Uncorrelated Unidimensional Model

Under the uncorrelated unidimensional model conceptions strongly predicted items (i.e., $\lambda \geq .60$ on external factors, $\lambda \geq .61$ on affect, $\lambda \geq .57$ on improvement, and $\lambda \geq .54$ on irrelevance; λ is a factor loading). Thus, items were adequately predicted by the latent constructs they were conceptually related to and the constructs had strong measurement scales.

3.6.2 Non-Hierarchical Multidimensional Model

Results under the non-hierarchical multidimensional model showed factor loadings of $\lambda \geq .53$ for external factors, $\lambda \geq .62$ for affect, $\lambda \geq .55$ for improvement, and $\lambda \geq .54$ for irrelevance. Although, some loadings increased compared to the regressions under the uncorrelated unidimensional model, other loadings decreased. The correlations between the three positive students' conceptions of assessment constructs were high (i.e., $r = .56$ between external factors and affect, $r = .71$ between external factors and improvement, and $r = .56$ between affect and improvement). Relations between irrelevance and the other three constructs were all negative, as was expected. The correlation with improvement was highly negative ($r = -.58$), the correlations with external factors and affect were weakly negative ($r = -.19$, and $r = -.13$ respectively).

3.6.3 Bifactor Model

The bifactor model resulted in estimates of factor loadings for both the general factor and the group factors. The general factor is a plausible

Table 3.2

Model results Students' conceptions of assessment data

Item	Sub-	UUM				NHMM				BFM				
	scale	Ext	Aff	Imp	Irr	Ext	Aff	Imp	Irr	G	Ext	Aff	Imp	Irr
Ext4	PF	.60				.54				.38	.46			
Ext11	EQ	.66				.69				.52	.43			
Ext16	PF	.65				.60				.40	.53			
Ext20	PF	.71				.71				.54	.44			
Ext24	EQ	.64				.54				*.33*	.61			
Ext33	PF	.61				.75				.61	*.26*			
Aff2	CA		.74				.73			.40		.63		
Aff6	SA		.61				.68			.50		.38		
Aff12	CA		.81				.83			.50		.65		
Aff17	CA		.82				.79			.39		.73		
Aff21	CA		.63				.63			.36		.52		
Aff25	CA		.81				.76			*.31*		.78		
Aff28	CA		.73				.73			.43		.59		
Aff31	SA		.64				.69			.50		.41		

* Ext=external factors, Aff=affect, Imp=improvement, Irr=irrelevance, PF=personal future, EQ=external quality,SA=self affect, CA=class affect, SI=self improvement, TI=teacher improvement, BA=bad, IG=ignore

Table 3.2 (continued)

Model results Students' conceptions of assessment data

Item	Sub-	UUM				NHMM				BFM				
	scale	Ext	Aff	Imp	Irr	Ext	Aff	Imp	Irr	G	Ext	Aff	Imp	Irr
Imp1	SI			.73				.72		.68			.42	
Imp5	TI			.63				.63		.65			*−.18*	
Imp8	TI			.61				.63		.65			*−.21*	
Imp9	TI			.57				.56		.57			*−.08*	
Imp10	SI			.78				.77		.74			*.31*	
Imp14	SI			.80				.78		.75			*.35*	
Imp15	SI			.83				.82		.81			*.23*	
Imp19	SI			.77				.78		.77			*.16*	
Imp23	TI			.74				.77		.78			*−.15*	
Imp27	TI			.61				.61		.63			*−.18*	
Imp30	TI			.60				.64		.66			*−.10*	
Irr3	BA				.68				.68	*−.33*				.60
Irr7	IG				.68				.67	*−.32*				.59
Irr13	BA				.68				.68	*−.30*				.61
Irr18	BA				.66				.64	*−.30*				.59
Irr22	BA				.55				.54	*−.27*				.48
Irr26	BA				.71				.70	*−.33*				.63
Irr29	IG				.71				.78	−.43				.58
Irr32	IG				.62				.59	*−.26*				.56

* Ext=external factors, Aff=affect, Imp=improvement, Irr=irrelevance, PF=personal future, EQ=external quality, SA=self affect, CA=class affect, SI=self improvement, TI=teacher improvement, BA=bad, IG=ignore

factor since loadingss were $\geq .35$ for all improvement items, most external and affect items, but only one irrelevance item. The factor loadings were highest among the improvement items, suggesting this conception reflects the dominant dimension in the general factor.

Most improvement and irrelevance items measured one conception only; the general conception or the group conception respectively. However, although the loadings were $< .35$, all the items measuring "student self improvement" had positive values on the improvement group factor; whereas the items measuring "teacher improvement" had negative loadings from the improvement group factor. This indicated that possibly, after taking into account their common variance, the improvement group was internally multidimensional.

Within the dominant pattern that the items within the improvement and irrelevant conceptions were primarily related to one construct, there were two exceptions. One improvement item had a strong loading from the improvement group factor (i.e., an item measuring the student "self improvement" sub-conception). This item (Imp1) indirectly focused on learning and studying (i.e., I pay attention to my assessment results in order to focus on what I could do better next time), in contrast to items that more directly focus on learning and studying (e.g., Imp10: I make use of the feedback I get to improve my learning; and Imp19: I use assessment to identify what I need to study next). Among the irrelevance items, one item (i.e., Irr29: I ignore or throw away my assessment results) had a factor loading $> .35$ for both the general and the group factor. Interestingly, this item has a double action embedded in it-both ignoring and actively throwing away information. This is in contrast to other items in the irrelevance group (e.g., Irr32: Assessment has little impact on my learning or Irr7: Assessment is ignored) which have only one concept. This additional verb–being an active rejection of the improvement process– may partially explain why the item is predicted by the general factor, as well as the irrelevance group factor. Hence, content analysis of these two items provided some insight into their discrepant behavior.

In contrast to the improvement and irrelevance items, most external factors items and most affect items had factor loadings $\geq .35$ on both general and group factor and are, thus, clearly multidimensional. Most external factors items had similar loadings for the general and group factors. Most affect items, were strongly predicted by both the general and group factors. The six "class affect" items were more strongly predicted by the group factor than the general factor, whereas the "personal enjoyment"

items were more strongly predicted by the general factor than the group factor. The factor loadings of affect items for the general and group factor were most alike when loadings on the general factor were somewhat higher, while the values were reasonably different when loadings on the group factor were higher.

The bifactor analysis showed that the improvement items were mainly predicted by the general factor and the "irrelevance" group factor predicted the irrelevance items. The affect group factor strongly predicted the affect items which were simultaneously predicted by the general factor. Likewise, the external factor items were predicted by both the general and group factors. Items that were predicted by both general and group factors had similar factor loadings or had far higher values for the group factor than the general factor. Improvement items were predicted more strongly by the general factor than affect and external factor items. Hence, the bifactor model solution showed that there are four different factors rather than five factors.

3.7 Discussion

Three models were used to evaluate the dimensionality of the SCoA-V inventory. The bifactor and non-hierarchical multidimensional models showed better fit than an uncorrelated unidimensional baseline model. Although, both the bifactor and non-hierarchical multidimensional models modeled relations between items of different conceptions, the non-hierarchical multidimensional model fitted better. Indeed, the bifactor analysis largely corroborated the non-hierarchical multidimensional model results, in that the general factor consisted mainly of improvement items and was, thus, not truly representative of a general conception of assessment. Although the external factor and affect items were predicted by both the group and general factors, it was clear they were more strongly predicted by their respective group factors than the general factor. Hence, it is defensible to conclude that the two group factors explained these items. This was even clearer for the irrelevance items, which were strongly predicted by only the group factor.

Based on these results, we conclude that the SCoA items consist of four correlated but distinct dimensions. The dimensionality structure of the Students' Conceptions of Assessment inventory can be best described by a non-hierarchical multidimensional model. Furthermore, it seems that the improvement factor is a more general construct than the other constructs

which seem more specific. This might be the reason for the shared variance of the improvement items with the affect and external factors items.

When predicting other external factors (e.g., academic performance or achievement) or when scoring persons on the conceptions, the non-hierarchical multidimensional model can be used, instead of a model requiring a general factor. When scoring individuals it is important to use estimated factor scores based on the non-hierarchical multidimensional model. As Chapter 2 showed, using simple sum scores on the domain-specific constructs might result in different ordering of persons, especially at the extremes of the continuum, which might lead to misspecification in a selection or classification context.

Note that in this study only the four major conceptions are used, and that the eight sub-conceptions were not taken into account. The model with four major conceptions and eight sub-conceptions as found by Brown, Irving, Peterson, and Hirschfeld (2009) has been estimated using MPlus (Muthén & Muthén, 1998-2006), resulting in values for fit statistics of $\chi^2(df) = 3797.585(481)$, and $\chi^2/df = 7.90$, with a p-value of $< .01$, CFI= .95, TLI= .95, RMSEA= .10, SRMR= 0.07, and gamma hat = .77. This multidimensional second-order model fitted slightly better than the non-hierarchical multidimensional model. One potential reason for the better fit of the multidimensional second-order model is that the model was more complex and permitted the sub-factors to behave more independently than the non-hierarchical multidimensional model in this paper, which forced the sub-factors to act together. However, to take into account the sub-conceptions when analyzing the full dimensionality structure of the instrument, prediction of external factors, and scoring of persons, it is important that each (sub)-conception contains enough items. It might be relevant to extend the number of items especially for the sub-conceptions containing four or less items.

While not able to fully test the sub-structure under all models, there are indications in the bifactor analysis that suggest item multidimensionality within the improvement and affect conceptions. Nevertheless, this study has shown that the current modeling of the SCoA as four inter-correlated conceptions is supported. This implies that the independent effects of the various conceptions upon academic performance are not artefactual but rather real structural relations. Researchers should have confidence in using the SCoA-V inventory in studies of student thinking about assessment.

Appendix

The Students' Conceptions of Assessment inventory

Item		Conception	Sub-conception
01	I pay attention to my assessment results in order to focus on what I could do better next time	Improvement	Self
02	Assessment encourages my class to work together and help each other	Affect	Class
03	Assessement is unfair to students	Irrelevant	Bad
04	Assessment results show how intelligent I am	External Factors	Personal Future
05	Assessment helps teachers track my progress	Improvement	Teacher
06	Assessment is an engaging and enjoyable experience for me	Affect	Self
07	I ignore assessment information	Irrelevant	Ignore
08	Assessment is a way to determine how much I have learned from teaching	Improvement	Teacher
09	Assessment is checking off my progress against achievement objectives and standards	Improvement	Teacher
10	I make use of the feedback I get to improve my learning	Improvement	Self
11	Assessment provides information on how well schools are doing	External Factors	External Quality
12	Assessment motivates me and my classmates to help each other	Affect	Class
13	Assessment interferes with my learning	Irrelevant	Bad
14	I look at what I got wrong or did poorly on to guide what I should learn next	Improvement	Self
15	I use assessments to take responsibility for my next learning steps	Improvement	Self
16	Assessment results predict my future performance	External Factors	Personal Future

The Students' Conceptions of Assessment inventory (continued)

Item		Conception	Sub-conception
17	Our class becomes more supportive when we are assessed	Affect	Class
18	Teachers are over-assessing	Irrelevant	Bad
19	I use assessment to identify what I need to study next	Improvement	Self
20	Assessment is important for my future career or job	External Factors	Personal Future
21	When we do assessments, there is good atmosphere in our class	Affect	Class
22	Assessment results are not very accurate	Irrelevant	Bad
23	My teachers use assessment to help me improve	Improvement	Teacher
24	Assessment measures the worth or quality of schools	External Factors	External Quality
25	Assessment makes our class cooperate more with eachother	Affect	Class
26	Assessment is valueless	Irrelevant	Bad
27	Teachers use my assessment results to see what they need to teach me next	Improvement	Teacher
28	When we are assessed, our class becomes more motivated to learn	Affect	Class
29	I ignore or throw away my assessment results	Irrelevant	Ignore
30	Assessment shows whether I can analyse and think critically about a topic	Improvement	Teacher
31	I find myself really enjoying learning when I am assessed	Affect	Self
32	Assessment has little impact on my learning	Irrelevant	Ignore
33	Assessment tells my parents how much I've learnt	External Factors	Personal Future

Chapter 4

Scaling Response Processes on Personality Items using Unfolding and Dominance Models

Personality trait assessment plays an important role in several fields of applied psychology and it has important diagnostic, classification, and selection consequences. For example, besides cognitive tests and job knowledge questionnaires, personality inventories are used to predict future job performance and future job satisfaction, and several studies showed substantial validities for personality variables across occupations (e.g., Ozer & Benet-Martinez, 2006).

Self-report inventories are often used to assess personality traits and item analysis is an essential part in the construction of these inventories. In the past, construction of self-report inventories mainly relied on optimizing internal consistency reliability by selecting items with high item-test correlations and by using factor analysis. Recently, however, item response theory (IRT) models have been increasingly used to analyze self-report personality data. The advantages of IRT models to analyze test data compared to classical approaches have been described in many sources (e.g., Embretson & Reise, 2000).

This chapter has been published as Weekers, A. M. & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment, 24*, 65-77.

In IRT models the probability of endorsing an item (i) is specified by the item response function (IRF) or item characteristic curve (ICC), which relates the probability of endorsing an item to a person's latent trait level (denoted as θ). Almost all studies that apply IRT models in the personality domain (e.g., Meijer & Baneke, 2004; Reise & Waller, 2003) use models that assume that a dominance process underlies item responding. That is, it is assumed that the higher someone's score on the latent trait, the higher the probability of endorsing an item. For example, when measuring depression by means of items like "I am often down in the dumps" it seems reasonable to assume that the more depressed someone is, the higher the probability that he or she will endorse this item. Recently, however, Chernyshenko, Stark, Chan, Drasgow, and Williams (2001) and Stark, Chernyshenko, Drasgow, and Williams (2006) showed that the IRFs of some items of the 16 Personality Factor Questionnaire (16PF, Conn & Rieke, 1994) cannot be described by monotonically increasing IRFs. Meijer and Baneke (2004) reported similar findings for the depression content scale of the Minnesota Multiphasic Personality Inventory-2 (MMPI-2, Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). Chernyshenko et al. (2001) and Meijer and Baneke (2004) found that the probability of endorsing an item sometimes decreased at the higher end of the trait continuum, whereas Stark et al. (2006) found that the IRFs for some items were single-peaked. These results indicate that response processes on self-report inventories may be different from what is expected under dominance models. As an alternative Stark et al. (2006) proposed ideal-point response processes and unfolding models to describe item responding (see also Roberts, 2001; Roberts, Laughlin, & Wedell, 1999). In an ideal point response process both persons and items are located on a continuum representing the trait of interest, and a person endorses an item only if the persons' latent trait value is located near the item location on the latent continuum. This assumption leads to nonmonotonic, single-peaked IRFs.

Stark et al. (2006) discussed that fitting the correct model to empirical data has important consequences in a personnel selection context. They showed that misspecification of the item response process for only a few items in the scale had serious consequences for the ordering of persons according to their latent trait scores. If, say, the top 10% or 20% highest or lowest scoring applicants are selected, this greatly affects who is being selected.

Besides the studies cited above, there has been very little experience with the use of other models than dominance models. Because inadequately

describing the response process on self-report inventories may result in wrong decisions, it is important to investigate which models best describe self-report inventory data. In the present study we extended the study by Stark et al. (2006). More specifically, the aim of this study was to (1) explore the usefulness of different IRT models to describe self-report personality data and (2) compare results obtained from dominance and unfolding IRT models.

4.1 Dominance and Unfolding IRT Models

Most IRT models for dichotomously scored items (0/1-scores) assume unidimensionality, local independence, and a specific form of the IRF. Unidimensionality means that test responses are assumed to depend on only one latent trait. Local independence holds when the response on an item in a test given theta is not influenced by the responses on other items in the test given theta. Furthermore, IRT models make assumptions about the shape of the IRF. An important goal of fitting IRT models is to identify the IRF that best describes the relation between the trait level and the probability of item endorsement.

4.1.1 Dominance IRT Models

In dominance IRT models it is assumed that the probability of item endorsement should increase as the trait level increases, thus, IRFs are monotonically increasing functions. Two types of dominance IRT models are characterized: parametric models and nonparametric models. The parametric IRT models (e.g., Embretson & Reise, 2000) describe the shape of the IRFs by parameters for items and persons. Parametric dominance models are characterized by s-shaped IRFs. As an alternative to parametric models, nonparametric models (e.g., Sijtsma & Molenaar, 2002) only assume monotone increasing IRFs. No parameters are estimated and an exact shape of the IRF is not specified. Therefore, nonparametric models have the advantage of being more flexible than parametric models. However, nonparametric models only allow the ordering of persons with respect to θ using the unweighted sum of item scores. One advantage of parametric IRT models compared to nonparametric IRT models is that they allow for the computation of item and scale information functions. The scale information function indicates where on the latent trait continuum measurement precision is high or low because it is inversely related to the standard error of measurement (e.g., Embretson & Reise, 2000). The

amount of information an item provides with respect to θ is determined by the discrimination parameter (α_i), and the location on the θ-scale where the information is maximized is determined by the item location (β_i).

4.1.2 Unfolding IRT Models

In IRT, both persons and items are located on a continuum representing the attribute of interest. In unfolding IRT models, persons only endorse items if the person's location on the trait continuum (the ideal point) and the item's location are close to each other. If not, a person will disagree with the item. A person does not endorse an item for one out of two reasons. Either a person is located too far above the item location, or a person is located too far below the item location. If the distance between person and item location increases, a person's probability of endorsing the item decreases. Unfolding models are, thus, characterized by single-peaked IRFs.

As in dominance IRT models, both parametric unfolding IRT models and nonparametric unfolding IRT models exist. The parametric model (e.g., Roberts, Donohogue, & Laughlin, 2000) describes the shape of the IRF by parameters for items and persons, which results in a specific bell-shaped form of the IRF. The IRF is symmetric around the item location and has an increasing s-shaped form on the lower side of the item location and a decreasing z-shaped form on the higher side of the item location. Item information and scale information statistics can be computed based on discrimination parameter values.

Nonparametric unfolding IRT models (e.g., Post, 1992) are more general models than parametric unfolding IRT models. They only assume single-peaked items. No parameters are estimated and no specific bell-shaped form is expected. The only assumption is that the probability of endorsement increases, reaches a maximum, and then decreases.

4.1.3 Differences between Dominance and Unfolding IRT Models

Stark et al. (2006; see also Post, van Duijn, & van Baarsen, 2001) give several arguments for considering unfolding models to analyze personality trait data.

First, scale construction under the dominance approach is based on searching scales with high item-total correlations, high internal consistency reliability, and a single dominant factor with high item factor loadings.

Using this approach, constructed scales are built up of mostly positively and negatively worded items with item locations that range from slightly positive or negative to extremely positive or negative and there is a tendency not to use neutrally worded items in the scales. As a consequence these scales have high measurement precision at one of the extreme regions of the trait continuum, but not in the middle of the trait range. In contrast, Stark et al. (2006) showed that, using unfolding models, neutrally as well as positively and negatively worded items - items located at any point on the trait continuum - can be selected. This expands the pool of usable items, may increase overall measurement precision of personality scales (see also Roberts et al., 1999), and may add to the construct validity of the scale.

When responding to personality inventories, which ask respondents to select statements or options that describe them best, we assume that there is a continuous scale where persons and items (statements) are located. Each statement or option describes a situation and has a threshold for the trait. The better the match between the statements formulated in the items and a person's self-perception, the higher the probability of endorsement. In most currently used personality scales, items are formulated relatively positive or negative (e.g., "I often crave excitement", or "I am not a cheerful optimist") and are endorsed by persons located at the higher and lower trait extreme, respectively. These items tend to have monotonic IRFs and can be described by a dominance model. On the other hand, items that describe behaviors tending toward neutrality and that describe average situations (e.g., "My ability to plan is about average" or, "I have sometimes done things just for 'kicks' or 'thrills"') can be described by nonmonotonic IRFs because persons in the middle of the continuum (i.e., persons with an average trait value in between the negative and positive extreme) have the highest probability of endorsing these types of items. Thus, we assume that each situation described in the statement has a threshold for the trait being measured and neutrally worded items describe behavior that is endorsed by persons in the middle of the continuum.

So, neutrally worded items might fill an item gap on an interval of the continuum where a substantial proportion of respondents may be located. Figure 4.4 (to be further discussed in the Discussion section) shows the trait estimates for persons on two personality inventories under two models.

Note that most persons have a person-location on the continuum between -1 and 1, which is in the middle range of the continuum and the area where neutrally worded items are located.

Information of neutrally worded items under unfolding models is double-

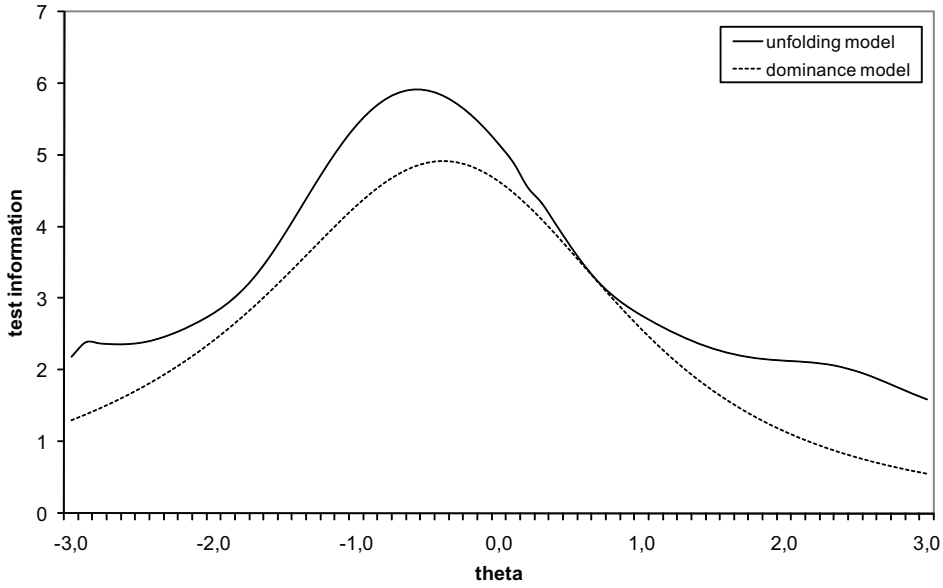

Figure 4.1. Example of difference in test information between dominance and unfolding model for an inventory with some neutrally worded items.

peaked as opposed to single-peaked under a dominance model. Item information under an unfolding model is spread over the whole continuum and results in a higher level of test information across the scale as compared to dominance models when neutrally worded items are included in the inventory. This is shown in Figure 4.1 for a scale with a few neutrally formulated items. Thus, neutrally worded items may increase measurement precision of an instrument.

Second, unfolding IRT models are more general models than dominance IRT models; that is, the IRF of a dominance IRT model can be considered a special case of the IRF of an unfolding IRT model, namely a single-peaked model with its peak at plus or minus infinity. Using a more general model may prevent misspecification of the response process. Misspecification of the response process may affect decision making in test applications like equating, the study of differential item functioning, and computerized adaptive testing.

Third, self-report personality inventories often consist of a mix of positively and negatively worded items. Researchers often implicitly assume that positively worded items measure the same latent trait as negatively worded items. This is not necessarily the case. Endorsing a

positively formulated item does not necessarily imply not endorsing an associated negatively formulated item and vice versa. When reverse scoring such items interpretation problems may result. An exception is when the IRF of the reversed scored item has the same form as the IRF of the non-reverse-scored item. Post et al. (2001) conclude that reverse-scored items can have consequences for the reliability and validity of an item set. For unfolding models it is not necessary to use reverse-scored items.

4.2 Aim of the Present Study

Although recent results by Stark et al. (2006), Meijer and Baneke (2004) and Chernyshenko et al. (2001) shed a new light on personality inventory analysis and personality inventory construction, it is unclear how well these results generalize to other personality inventories measuring different personality constructs and whether similar results can be found using different IRT models. Therefore, in the present study we extend these studies through analyzing two inventories, which measure partly different personality constructs than in the Stark et al. (2006), Meijer and Baneke (2004) and Chernyshenko et al. (2007) studies. These personality inventories differ in the way they are constructed. One inventory is constructed based on dominance models, whereas the other is constructed based on unfolding models.

4.3 Method

4.3.1 Instruments

NPV-J

The Dutch Personality Questionnaire Junior (Dutch: Junior Nederlandse Persoonlijkheidsvragenlijst, NPV-J; Luteijn et al., 2005) consists of 105 mostly positively formulated items and is intended to determine how adolescents between 9 and 15 years of age judge their own behavior on five scales. These scales represent the traits inadequacy (IN), persistence (PE), social inadequacy (SI), recalcitrance (RE) and dominance (DO). The NPV-J was constructed making use of items from the California Personality Inventory (CPI; Gough & Bradley, 1996) and The Dutch Personality Questionnaire (Dutch: Nederlandse Persoonlijkheidsvragenlijst, NPV; Luteijn, 1974), which is the adult version

of the NPV-J. Scales were constructed using a combination of different dominance response strategies: maximizing internal consistency, factor analysis, and empirical keying (Luteijn et.al., 2005, p. 8).

The NPV-J was used because it is a commonly used personality inventory in the Netherlands and it provides measures of several important personality traits. Barelds and Luteijn (2002) compared the factor structure of the adult version of the NPV-J (the NPV) with the Five-Factor Personality Inventory (FFPI; Hendriks, Hofstee, & de Raad, 1999) and the Dutch version of the Eysenck Personality Questionnaire (EPQ; Sanderman, Arrindell, Ranchor, Eysenck, & Eysenck, 1995). Barelds and Luteijn (2002) found that IN (or Neuroticism) correlated highly with Emotional Stability (FFPI; $-.65$) and Neuroticism (EPQ; .78). Furthermore SI (or Social Anxiety) and DO correlated highly with Extraversion (FFPI; $-.74$ and .48 respectively, and EPQ; $-.67$ and .58 respectively) and PE (or Rigidity) correlated highly (.57) with Conscientiousness (FFPI).

Scoring was originally done on a three-point scale (*Agree, ?, Disagree*) but because the instructions of the NPV-J discourage the use of the *?* response, and because we were afraid that many adolescents would choose the *?* category, a two-point scale (*Agree, Disagree*) was used. The answer *Agree* was scored as one and the answer *Disagree* was scored as zero.

Data analysis of the NPV-J items was based on the original scales. To date, the psychometric properties of the NPV-J have mainly been investigated using classical test theory (CTT) and factor analytical approaches. Luteijn et al. (2005) showed a reasonable fit for a five-factor model with Cronbach's α for the subscales between .68 (DO) and .90 (IN). Although the NPV-J was constructed using dominance models, a first analysis of our data using a dominance model (see results below) showed that a substantial part of the items had low item-discrimination parameters. This could be because some items may be better described by an unfolding model. We were curious to know whether this is the case.

Order Scale

Second, we used a Dutch translation of a personality inventory that is intended to determine the self-judgement of adolescents and adults on the order-facet (an important facet of Conscientiousness). The inventory was recently constructed by Chernyshenko et al. (2007) by creating items to represent the full range of behaviors (positive, negative, and moderate/neutral) associated with orderliness. The content of the items was rated on a 7-point scale in terms of their extremity/location. For

example, the item "I spend a lot of time looking for objects I misplaced" received a rating of 1 (most negative pole) and the item "Every item in my room and on my desk has its own designated place" received a rating of 7 (most positive pole). The item "My room neatness is about average" received a rating of 4, which indicates a neutral position. Then, unfolding models were used to analyze the items and construct the final scale.

The final scale consisted of 20 positively, neutrally, and negatively worded items. Scoring was done on a 4-point scale (*Strongly agree, Agree, Disagree, and Strongly disagree*). As in Chernyshenko et al. (2007), in the present study data were dichotomized before they were analyzed. The categories *Strongly agree* and *Agree* were scored as one and the categories *Disagree* and *Strongly disagree* were scored as zero. This was done because there were few persons that endorsed the options *Strongly Agree* and *Strongly Disagree*, and as a result the item parameters of these categories would have been estimated very inaccurately.

As far as we know, the Order scale is the only personality scale that is developed using ideal-point modeling. We were curious to know how this scale behaved in a replication study using partly different models and a different population.

4.3.2 Participants and Procedure

NPV-J

Data were collected from 866 persons who attended primary and secondary education in the east of the Netherlands. Participants were 492 girls and 374 boys; the majority were white. Mean age of the participants was 13.8 years of age ($SD = 2.7$). 70.7% were between 9 and 15 years of age; 2.1% were (a few months) younger, and 27.2% were between 15 and 18 years of age. Although the NPV-J is originally intended for children between $9 - 15$ years of age, it is our experience that the inventory is also very useful for persons who are somewhat younger or older. Because we were only interested in studying the response processes (and did not need norm tables, which are only available for children between 9 to 15 years of age) we included the younger and older persons in the sample.

Order Scale

Data on the Order scale were collected from 704 persons who attended secondary and college education. Participants were 397 girls and 299 boys;

the majority were white. Mean age of the participants was 16.0 years ($SD = 2.6$).

Note that the samples contained somewhat different populations because the NPV-J and the Order scale are intended for different populations (adolescents, and adolescents and adults, respectively). Before the adolescents filled out the inventories, standardized oral instructions were provided by the authors. During test administration, the authors were available for further clarification. No time limits were set.

4.3.3 Analyses

Both dominance IRT models and unfolding IRT models were used to analyze the NPV-J and Order scale data. For each type of model we used a parametric and a nonparametric model. We evaluated the robustness of our results by applying different scaling methods and IRT models to the same dataset.

Dominance models

Mokken's model of monotone homogeneity Mokken's nonparametric monotone homogeneity model (MHM; e.g., Sijtsma & Molenaar, 2002) assumes unidimensionality, local independence, and monotonically increasing IRFs. Particular shapes of the IRFs are not specified. The MHM imposes the following restriction: $P(\theta_a) < P(\theta_b)$ for all $\theta_a \leq \theta_b$, in which a and b denote persons. This restriction is not restrictive enough to allow the estimation of an individual's score on a latent variable, but rather the MHM uses a person's raw scale-score to order persons on a construct.

To check whether the IRFs are monotonically increasing we used the computer program Mokken Scaling for Polytomous Items, version 5.0 (MSP5.0, Molenaar & Sijtsma, 2000). One way to check monotonicity involves the computation of H-coefficients that indicate the scalability of items (H_i), of item pairs (H_{ij}), and of the total scales (H). The coefficients are expressed by a ratio of the observed covariance and the maximum value of this covariance. The program calculates item, item pair, and scale coefficients for the total scales. Values of these scalability coefficients range from zero to one. Increasing values between .30 and 1.00 indicate evidence for monotone increasing IRFs, whereas values below .30 indicate violations of increasing IRFs.

One-parameter logistic model The one-parameter logistic model (OPLM, Verhelst & Glas, 1995) is a parametric dominance IRT model. Similar to the MHM, unidimensionality and local independence are assumed. The shape of the IRF is a logistic function. OPLM estimates the shape of the IRF by estimating the location parameter after choosing a user-specified integer slope, which is the item discrimination. The IRF is defined as

$$P(X_i = 1 \mid \theta) = \frac{\exp A_i(\theta - \beta_i)}{1 + \exp A_i(\theta - \beta_i)} \tag{4.1}$$

where $X_i = x_i$ is the observed item response ($x_i = 1$ for the agree response, and $x_i = 0$ for the disagree response), A_i is the user-specified integer slope ($A_i > 0$), and β_i is the item location parameter. When integer values are chosen for the slopes, the statistical properties are analogous to the properties of the Rasch model.

The program OPLM (Verhelst, Glas & Verstralen, 1995) first estimates the location parameters after choosing a user-specified integer slope, which is equal for each item. When the model with equal user-specified slopes does not fit the data, values for the integer slopes can be changed to higher and lower values. These integer slope values differ over items. If the slope parameters are fixed at integer values, the only parameters to be estimated are the location parameters. This is repeated until a fitting model is found.

The program OPLM was used to check the fit of the logistic IRT model. The null hypothesis of monotonicity and sufficiency of the total score was investigated by means of four item statistics (one χ^2 for each item and three M-statistics) and a global asymptotic χ^2 statistic, denoted by R_{1c} (Glas, 1988). χ^2 statistics were calculated for each item. The χ^2 statistic for the items is sensitive to overestimation or underestimation of the item slope indices. M-statistics give information about when the integer slopes have to be adapted and in what direction. The R_{1c}-statistic is the sum of the item χ^2 statistics and gives information about model fit.

Unfolding Models

Multiple unidimensional unfolding model The nonparametric multiple unidimensional unfolding model (MUDFOLD; Van Schuur and Post, 1998; Post et al., 2001) assumes unidimensionality, local independence, and single-peaked IRFs. The model does not further define the shape of the IRFs. The IRF of an item gives the probability of endorsement given a persons' location in relation to the item. Persons located below or above the item location will have lower probability of

endorsing the item than persons who have a location similar to the item. This assumption is not restrictive enough to allow the estimation of an individual's score on a latent variable but, based on scores on items, persons are ordered on an ordinal bipolar scale.

H-coefficients are used to check the scalability of the items. Scalability coefficients are used for item triples, H_{ijk} , for an item, H_i, and for the total scale, H. They are expressed by the ratio between observed and expected triple values. The MUDFOLD program was used to calculate the H-coefficients. The program first searches the optimal smallest unfolding scale, that is, a first subset of items that together have a relatively high H-value and, thus, form a strong scale. The selection of the first triple of items is based on selecting the three items with the highest positive scalability value H_{ijk} in one triple order, while the other scalability values of that triple are negative. In addition, H_{ijk} must be higher than a user-specified value c (default is $c = 0.30$). Then items are added one by one, constantly checking if H_{ijk}, H_i, and H are equal to or larger than c. Second, c-values were lowered, all items of the initial scale are added, and H-coefficients for the total scale, its items, and item pairs were calculated. Increasing values between .30 and 1.00 indicate more convincing evidence for single-peaked IRFs, whereas values below .30 indicate violations of single-peakedness.

An important diagnostic in MUDFOLD is the conditional adjacency matrix (CAM; Post & Snijders, 1993) and its fit statistics, the ISO-, UNI-, and MAX-statistics. The CAM matrix can be used to check the IRFs. The ijth entry in the matrix represents the conditional probability of choosing row item I_i, given that column item I_j is endorsed. The different values in the rows indicate the shape of the IRF. If the scale is described by an unfolding model, the maxima of the rows shift from the top left column to bottom right column, except for inversions around the diagonal. The ISO-statistic gives the degree of violation of the unimodality in a row, whereas the UNI-statistic shows which items form disturbances of unimodality. The MAX-statistic controls for the order of shifting tops, and shows which items have a disturbance.

Generalized graded unfolding model The generalized graded unfolding model (GGUM, Roberts, Donoghue, & Laughlin, 2000; Roberts, Fang, Cui & Wang, 2004) assumes unidimensionality and local independence. The IRF is a bell-shaped function. The GGUM IRF has three item parameters, the location parameter β_i, the discrimination parameter α_i, and the subjective response category threshold τ_i. The IRF

for the dichotomous case is given by

$$P(X_i = 1 \mid \theta) = \frac{\exp(f) + \exp(g)}{1 + \exp(f) + \exp(g) + \exp(h)}. \tag{4.2}$$

in which

$$f = \alpha_i(1(\theta - \beta_i) - \tau_i),$$
$$g = \alpha_i(2(\theta - \beta_i) - \tau_i),$$
$$h = \alpha_i(3(\theta - \beta_i)).$$

In Figure 4.2 we show an example of an IRF with parameters: $\beta_i = 1.0$, $\alpha_i = 1.5$ and $\tau_i = -0.5$. As can be seen the IRF attains its maximum value in β_i and has a single-peaked response function. When the distance between item location β_i and the person location θ increases, the probability of endorsing the item decreases. The discrimination parameter α_i expresses how well we can discriminate between persons ($\alpha_i > 0$). The τ_i parameter ($\tau_i < 0$) expresses the distance on the θ continuum from the item location β_i to the location where endorsing becomes more likely than not endorsing the item (for a more detailed description see Roberts et al., 2000).

The GGUM2004 program (Roberts, et al., 2000, 2004) was used to estimate the item and person parameters and to inspect graphs of the IRFs. Although GGUM2004 also contains item and model fit statistics, such as infit and outfit statistics, according to the manual these statistics are generalized from cumulative IRT applications and are not mathematically deduced for GGUM. Little is known about their distribution, their power and their Type I error rate. Therefore the MODFIT computer program (Stark, 2001) was used to compute χ^2 statistics and fit plots for GGUM. Means and standard deviations of the adjusted χ^2/df ratios were computed to summarize the results for each scale. Values above 3.00 may indicate model misfit.

4.4 Results

4.4.1 Dominance Models

Number of items (K), scale means (M), standard deviations (SD), skewness and kurtosis, Cronbach's α, and average item-test correlations (ρ_{iT}) are given in Table 4.1. Scale distribution and values of Cronbach's α for the NPV-J sample were comparable to the values found in earlier

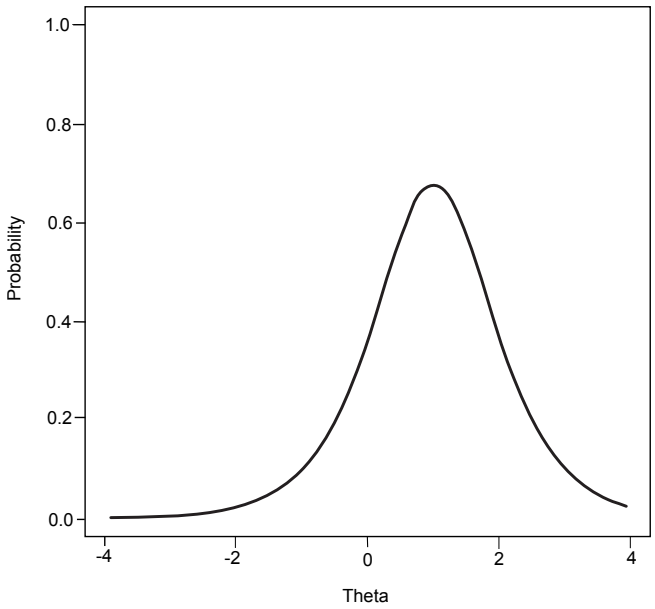

Figure 4.2. Example of an IRF under GGUM model.

studies (Luteijn et al., 2005). Average item-test correlations were higher for the IN scale and the SI scale than for the PE, RE, and DO scales. The average item-test correlation for the Order scale was .30 and Cronbach's α was equal to .74.

Results with respect to the RE scale are not discussed in detail in the following paragraphs. This scale formed a weak overall scale under all models. Items were monotonically increasing, but most items showed low discrimination values. Although this scale is not discussed, information on this scale is displayed in the tables.

Table 4.1
Descriptive statistics and MSP H-coefficients

Scales	K	M	SD	α	Skewness	Kurtosis	ρ_{iT}	H	H_i-range
Inadequacy	28	6.23	5.19	.87	1.18	1.14	.414	.28	.15-.47
Persistence	25	17.93	3.96	.74	-0.64	0.14	.278	.16	.05-.27
Social Inadequacy	13	5.01	3.17	.79	0.27	-0.86	.419	.34	.18-.43
Recalcitrance	24	8.25	3.68	.72	0.64	0.26	.266	.17	.09-.26
Dominance	15	5.46	2.65	.66	0.74	0.42	.278	.20	.10-.31
Order	20	12.19	3.74	.74	-0.18	-0.66	.301	.19	.04-.28

Table 4.2
Fit statistics and A_i-values OPLM

Scale	K	R_{1c}-Rasch (p-value)	A_i	R_{1c}-OPLM (p-value)	nonfitting items
Inadequacy	28	313.06 (.00*)	22252 52722 33232 63364 24364 224	76.31 (.60)	1
Persistence	25	246.23 (.00*)	13342 36232 44552 24243 44425	71.12 (.51)	0
Social Inadequacy	13	256.24 (.00*)	32534 13541 533	46.48 (.11)	1
Recalcitrance	24	160.78 (.00*) (.00*)	23323 13433 23224 34463 5524	65.58 (.59)	2
Dominance	15	133.32 (.00*)	36322 22442 36741	32.14 (.84)	1
Order	20	302.97 (.00*)	36523 62x52 23243 21325	65.84 (.13)	0

Monotone Homogeneity Model

NPV-J H- and the range of H_i values are shown in Table 4.1. H-values varied between .16 and .34. For the IN and SI scales most items had H_i-values larger than .20. The PE and DO scales consisted of items with low H_i-values. Note that low H_i values point at low item discrimination but may also indicate the presence of single-peaked IRFs.

Order scale As expected, the Order scale formed a weak overall scale with an H-coefficient of .19. This low value was not surprising because the scale was constructed based on ideal-point response processes and thus single-peaked IRFs were expected. The range of H_i-values of the items (Table 4.1) showed that many items had low discriminating power.

OPLM

NPV-J We first fitted the Rasch model for each scale. Table 4.2 (Column 3, R_{1c}-Rasch) shows that for each scale the Rasch model did not fit the data. All R_{1c}-values for model fit were high and had p values below .05. The slope values were adapted and the R_{1c} was computed. Table 4.2 (Column 5, R_{1c}-OPLM) shows the fit of this model. In all cases R_{1c} was improved substantially compared to the Rasch model.

The interpretation of A_i-values and their variation is relative to their

scale: If all A_i-values were doubled, the fit would remain unchanged. Nevertheless, most items of each scale had slope indices around its mean (mean $A_i = 3$). This indicated that items in the scales had relatively equal slope indices (A_i ranging from 2 to 4). OPLM results and MHM results were similar. The items with relatively high A_i-values equal to 5, 6, or 7 had the highest H_i-values in their scales under the MHM, whereas items with the lowest H_i-values in their scales had relatively low A_i-values equal to 1 or 2.

Order scale Table 4.2 shows a high R_{1c}-value and a p-value smaller than .05 for the Rasch model. The Rasch model did not describe the data well. After the slope values were adapted and fit indices were computed, the model still did not describe the data well. One item (item 8: "my room neatness is about average") had to be deleted to obtain reasonable fit indices (see Table 4.2). Notice that this is a neutrally worded item, which might have a single-peaked IRF. Most items in the remaining scale had A_i-values between 2 and 4. OPLM results were similar to MHM results. The item that had to be deleted had H_i-values of .32 under the MHM.

4.4.2 Unfolding Models

GGUM

NPV-J Mean and standard deviations of the adjusted χ^2/df ratios for singles, doubles and triples were computed (not tabulated). Most scales showed good fit for singles, doubles and triples. Only the IN scale did not fit the data well for both singles and doubles. Mean values were slightly above 3.

In line with the results discussed above all scales consisted of items with monotone increasing IRFs; item location parameters were mostly around 2 or above. Most scales consisted of dominance items with a mix of high discriminating items and low discriminating items. For the IN and SI scales some items showed a trend to single-peakedness at the higher end of the trait continuum. The PE, SI and, DO scales contained items with single-peaked IRFs. In the PE scale the six items with single-peaked IRFs had both highly and low discrimination values, while the two single-peaked items in the SI scale and the two single-peaked items in the DO scale had high discrimination values. An example of a single-peaked item from the PE scale was "When I have done something wrong, I feel terrible". An example of a highly discriminating single-peaked item of the DO scale was

"I like to act the boss".

For most scales, the results under the GGUM model were similar to the results under the dominance models. Only the two single-peaked DO items discussed above were highly discriminating increasing items under the dominance models, and were also highly discriminating under GGUM. The reason is that under both types of models the items are located at the higher extreme of the θ scale (around $\theta = 1.5$).

Order Scale Mean and standard deviation values of the adjusted χ^2/df ratios for singles, doubles and triples were calculated (not tabulated). The Order scale showed good fit under the GGUM model. Eight items of the Order scale showed monotonically increasing IRFs and eight items showed monotonically decreasing IRFs. About half of the monotonically increasing and half of the monotonically decreasing items were highly discriminating items. Some items showed folding at the higher or lower end of the continuum. Four items were found with single-peaked IRFs, two of which had high discrimination values. Two examples of single-peaked items were "Being neat is not exactly my strength" (highly discriminating) and "My ability to plan is about average" (low discriminating). Furthermore, results were mostly similar to results under the dominance models.

MUDFOLD

NPV-J We first used MUDFOLD to identify reasonably strong initial clusters of items for all the scales (see Table 4.3). All first clusters had H-values between .37 and .43, and contained 4 to 10 items. Second, the selected clusters of items were used as a start set to select all items of each scale. Not all a priori scales were scalable under MUDFOLD. For the IN and PE scales the program showed a warning indicating that it could not order all items. A possible explanation is that the items had item locations too close to each other; another explanation is that the item discrimination was low across the trait continuum. For the SI and DO scales items could be ordered. Table 4.3 shows that $H = .35$ for the SI scale, whereas for the DO scale $H = .20$. For the SI scale all items had H_i values between .27 and .53, whereas for the DO scale there was a mix of high and low H_i-values.

CAM-matrices, ISO-, UNI- and MAX-statistics were used to check the shape of the IRFs in the SI and DO scales. ISO-, UNI- and MAX-values for the SI and DO scales showed that there were some violations to the modeled item shapes. Inspection of the item statistics (not tabulated) showed small disturbances in unimodality and shifting tops for only a

Table 4.3
H-coefficients and H_i-coefficients MUDFOLD for both cluster steps

	1^{st} step			2^{nd} Step		
Scale	K	H	H_i-range	K	H	H_i-range
Inadequacy	6	.37	.31-.43	-	-	-
Persistence	10	.37	.33-.41	-	-	-
Social Inadequacy	8	.37	.31-.54	13	.35	.27-.53
Recalcitrance	4	.43	.41-.46	-	-	-
Dominance	6	.38	.32-.47	15	.20	.15-.24
Order	7	.40	.35-.45	20	.23	.15-.30

few items in the SI scale. The CAM matrix showed that most items in the SI scale had monotonically increasing or decreasing IRFs, sometimes with a trend to single-peakedness at the higher end of the trait continuum. Two items showed single-peaked IRFs. In general, the CAM matrix of the DO scale showed monotonically increasing items with weak discriminating values. Three items with (a trend to) single-peaked IRFs were found. Weak discrimination may lead to weak disturbances in unimodality and shifting tops. This was reflected by the item statistics for the DO scale (not tabulated).

Results for the SI scale were similar to the results found under the dominance models and the GGUM model. An example of a folding item under all programs was "I prefer playing games I am familiar with". This item had very flat slopes under the dominance models, whereas it showed single-peaked IRFs under the unfolding models. An explanation is that highly socially adequate respondents do not want to play familiar games because they find it boring, whereas very socially inadequate respondents do not want to play games at all because they do not like to interact with other people in an ill-defined context. In general, results for the DO scale were also in agreement with the results found under the dominance models and the GGUM model.

Order scale A first cluster of seven items was found (Table 4.3). This first cluster was selected as a start set to select all items from the scale. *H* was equal to .23 for the total scale. The ISO-, UNI- and MAX-statistics showed that there were some violations to the model. The CAM matrix showed that the Order scale consisted of a mix of monotonically increasing (10 items), monotonically decreasing (6 items) and single-peaked items (4 items). The item statistics (not tabulated) pointed at small violations

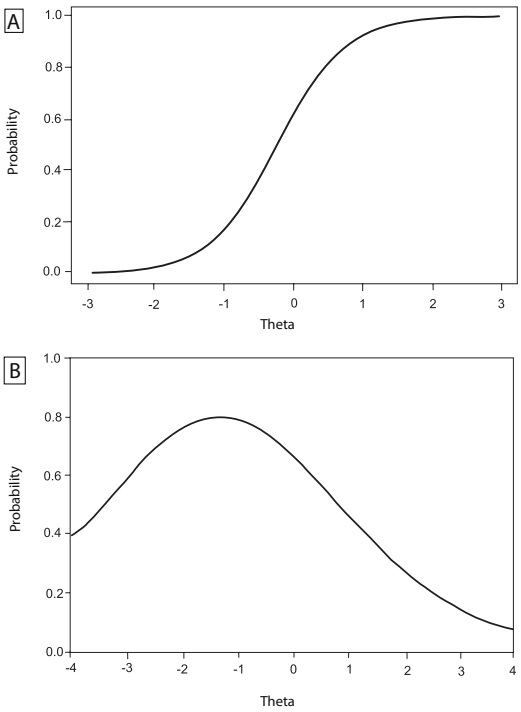

Figure 4.3. IRFs for item "Although I try to keep everything in its place, it does not always work for me" for a) OPLM, and b) GGUM.

against unimodality and shifting tops and some monotonically increasing and decreasing items showed some folding at the higher and lower end of the trait continuum, respectively.

In general, results found under GGUM and MUDFOLD were similar. In Figure 4.3 the IRFs of the item, "Although I try to keep everything in its place, it does not always work for me" are shown. This item showed consistent results under all models. The item had a relatively flat slope ($A_i = 2$) under the dominance models, whereas it showed single-peaked IRFs under the unfolding models. An explanation might be that ordered persons always keep things in their places, while unordered people do not even try, or do not succeed.

4.5 Discussion

In this study, we investigated the fit of both dominance and unfolding IRT models to self-report personality inventories. With respect to the NPV-

J, results confirmed the findings of Stark et al. (2006), Chernyshenko et al. (2001), and Meijer and Baneke (2004) that some items in a dominance response-based personality inventory are single peaked or show a trend to single-peakedness at the higher or lower end of the continuum. Because single-peaked items are often neutrally worded items that are situated in the middle of the latent trait continuum these items may add to the measurement precision in an area where dominance items are difficult to formulate.

In general, results with respect to the Order scale confirmed the result found in the Chernyshenko et al. (2007) study. As they showed, it is possible to construct a scale containing items with monotonically increasing, monotonically decreasing and single-peaked IRFs. However, results on the Dutch translation of the scale showed fewer single-peaked items and items had, in general, lower discrimination power.

Item analysis is an essential part of scale development. It is our belief that nonparametric and parametric dominance and unfolding models are useful models to obtain information about the characteristics of the IRF. In many papers about the fit of different IRT models to personality items, there is no distinction between "general" personality items (like the items in the NEO-PI-R; Costa & McCrea, 1992) and psychopathology items (like the items in the MMPI; Butcher, et al., 1989). There may be, however, an important difference between these two types of items, which may influence the fit of an IRT model to the data. Psychopathology scales usually consist of statements that are rather extreme because clinicians are mainly interested in extreme behavior. For example, consider one of the MMPI scales, the acute anxiety scale. Items in this scale consist of extreme statements like "I sometimes feel that I am about to go to pieces". Even in a psychiatric population, this item is only endorsed by people scoring very high on acute anxiety. As a result most items in this scale are characterized by IRFs that are located at the higher end of the latent trait scale. These items fit a dominance model because there will be very few highly anxious people that will not endorse these extreme items. On the other hand, general personality self-report inventories like the ones that are used for personnel selection purposes consist of scales that consist of items that are more spread across the latent trait continuum. For example, a Conscientiousness scale may consist of items that discriminate across the whole range of the trait because high conscientiousness predicts job performance and middle to low conscientiousness predicts risky behavior in job situations (Ozer & Benet-Martinez, 2006). In this case, neutrally

worded items may be formulated that are best described by an unfolding model.

As Stark et al. (2006) discussed, misspecification of the item response process may have serious consequences for the ordering of persons according to their latent trait scores and may affect the conclusions in a diagnostic, classification, or selection context. To illustrate this, consider Figure 4.4. Here we plotted the estimated θ values ($\hat{\theta}$) of the IN scale and the Order scale under a dominance and an unfolding model. The correlations between the trait scores for both models were high (.988 and .971 for the IN and Order scale, respectively). As Figure 4.4a illustrates, when all the IRFs are characterized by a dominance model (IN scale) the $\hat{\theta}$ values cluster tightly along the diagonal line with a few exceptions, thus ordering the $\hat{\theta}$ values similarly. In contrast, for the Order scale (Figure 4.4b) which is characterized by both single-peaked and monotonically increasing IRFs scattering of the $\hat{\theta}$ values about the diagonal line indicates that the $\hat{\theta}$ values are differently ordered under the two models, especially at the upper extreme.

Finally, the results in this study showed some evidence for the application of unfolding IRT models in personality measurement. Future research in the field of scale analysis and scale construction should point out in which domains of personality measurement unfolding models are especially useful.

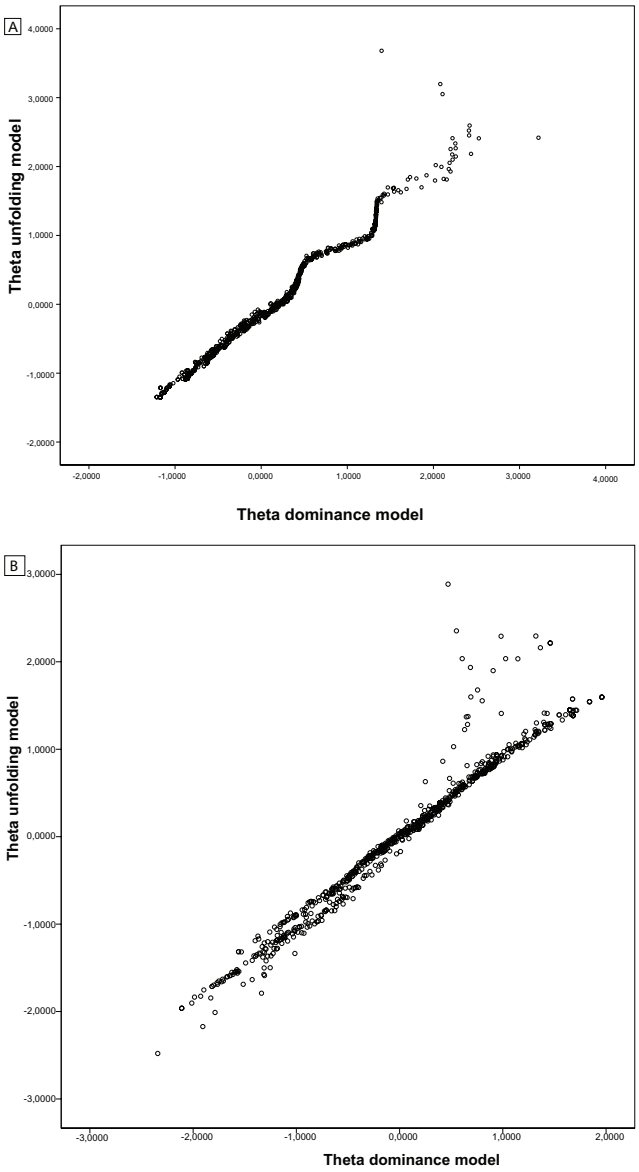

Figure 4.4. Scatter-plot comparisons of traits estimates from dominance and unfolding models for a) Inadequacy scale, and b) Order scale.Every circle represents a person's trait estimates under the two models.

Chapter 5

Person fit tests for unfolding IRT models

5.1 Introduction

Typical performance measures (e.g., attitude and personality inventories) are often analyzed and scored using factor analytic and dominance IRT models. However, there are indications that responses to typical performance measures do not follow a factor analytic or dominance IRT model (Chernyshenko, Stark, Drasgow, & Roberts, 2007; Weekers & Meijer, 2008). Contrary to maximum performance measures (e.g., educational tests) on typical performance measures it is likely to expect that persons only endorse items that are close to their personal location on the continuum; persons located on the higher end of the trait continuum will endorse (extremely) positively formulated statements, persons located on the lower end of the trait continuum will endorse (extremely) negatively formulated statements. Persons in the middle of the trait continuum, the persons with an average location on the trait, will endorse neutrally formulated statements. This indicates that items will not only have monotone increasing tracelines and monotone decreasing tracelines, but single-peaked tracelines as well. Unfolding models are characterized by single-peaked response functions.

Following unfolding models, persons may not endorse a statement for one out of two reasons, as was already stated by Thurstone (1928) and Coombs (1964). A person who has a person location too far above the item location will disagree because (s)he has a more positive opinion than what is stated in the item, and thus *disagrees from above*, whereas a person

who has a person location below the item location will *disagree from below*, because (s)he has a more negative opinion than what was stated in the item. Each item has a trait range in which the item is most likely to be endorsed, the lattitude of acceptance (Coombs, 1964), that is, the range from the threshold between disagree from below and agree to the threshold between agree and disagree from above. Various parametric unfolding models are developed. The models differ in the way they model the lattitude of acceptance (Luo, 1998). The lattitude of acceptance is expressed by one or two parameters for heighth and width of the item response curve. Some models are expressed in terms of polytomous dominance IRT models, such as the cosine hyperbolic model (Andrich, 1996), collapsed the partial credit model (Verhelst & Verstralen, 1993), collapsed graded response model (Korobko, 2007) and the generalized graded response model (Roberts, Donoghue, & Laughlin, 2000, 2002).

Although studies showed that unfolding models fit the data and person scores could be calculated under the model (Chernyshenko, Stark, Drasgow, & Roberts, 2007.; Weekers & Meijer, 2008), no research has been done to detect person misfit. The responses a single person gives to items might not be conform the model, that is, the model might not be valid for each person. Reasons for invalidity of test scores can be due to various factors (see also Reise & Flannery, 1996), think about; faking good (social desirability) or bad (malingering), unmotivated test responding, misalignment, tendency to agree or extreme item responding. The validity of persons' test scores can be investigated using person fit statistics. In the context of maximum performance testing, several person fit statistics for dichotomous and polytomous IRT models have been proposed to identify aberrant item score patterns. These statistics are also proven helpful in typical performance assessment. Meijer and Sijtsma (1995, 2001) give an overview of person fit statistics for dominance IRT models. Although some research is done on person fit of typical performance measures, it is mainly focussed on dominance models, person fit for unfolding models is hardly investigated.

In this study we will develop person fit statistics for unfolding models based on the Lagrange Multiplier-test (LM-test). The LM-test to measure person fit was proposed by Glas and Dagohoy (2007) for polytomous dominance IRT models (e.g., generalized partial credit model (Muraki,1992), graded response model (Samejima, 1969) and sequential model (Tutz, 1990)). These person fit statistics are also useful for the dichotomous dominance IRT models, which are special cases of their

polytomous variants. In this paper the person fit statistic based on the Lagrange Multiplier statistic will be used to detect person misfit to unfolding IRT models for dichotomous items. The statistics can be extended to use for person fit on polytomous unfolding models.

The paper is organized as follows. First four unfolding IRT models are described. Some models are already existing models, other models are developed in this study. Second, the LM-tests for person fit are explained. Two person fit tests will be developed, one test to test constancy of theta, and one test for tendency to agree. Third, simulation studies will be conducted to assess Type I error rate and power of both LM-statistics and to test robustness of the four different models. Finally some recommendations are given.

5.2 Unfolding IRT models

Four models will be introduced. Three models (i.e., generalized graded unfolding model, collapsed partial credit model, collapsed graded response model) are collapsed versions of well-known polytomous IRT models, and one model (i.e., quadratic logistic regression model) is a simple logistic regression model. For all models, let a test consist of a certain number of items, labeled $i = 1, ..., K$, with dichotomous response categories $j = 0, 1$, and let the item responses be denoted by stochastic variable X_i with realization x_i, and $x_i = 1$ if the item is endorsed, and $x_i = 0$ if the items is not endorsed. The probability of scoring in a response category is given by $P(X_i = j \mid \theta)$, in which theta (θ) is a latent ability variable for the person.

5.2.1 The generalized Graded Unfolding Model

The generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000) is a collapsed version of the generalized partial credit model (GPCM; Muraki, 1992). The dichotomous case of GGUM can be seen as a collapsed version of a 4-category GPCM (Figure 5.1). Persons can disagree with a statement for two reasons; they have a location too far above the item location (disagree from above; curve 4 in Figure 5.1) or they have a position too far below the item location (disagree from below; curve 1 in Figure 5.1). Similarly, a person can endorse an item, because the person is located slightly above the item location (agree from above; curve 3 in Figure 5.1) or slightly below the item location (agree from below; curve 2 in Figure 5.1). The two observed category responses $j = 0, 1$ on an item i, are the result of the four latent response categories $z = 1, 2, 3, 4$.

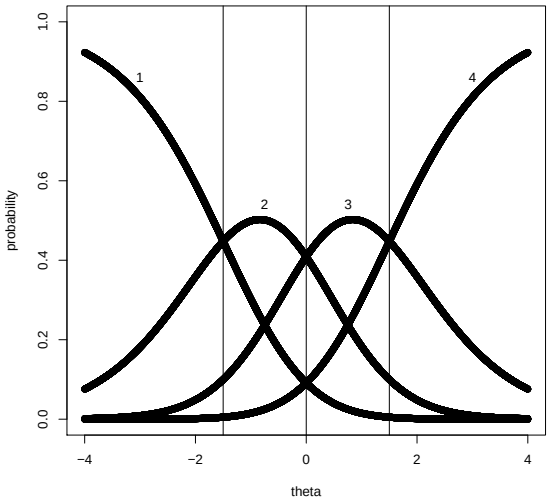

Figure 5.1. 4-category generalized partial credit model with parameter values $\alpha_i = 1.00$, $\beta_i = 0.00$, $\tau_i = -1.50$

The probability of answering in the latent category of an item, is denoted by $P(Y_i = z \mid \theta)$, with the realization of Y_i as the response in the z^{th}-category to the i^{th} item, and $z = 1$ is the latent response in the category disagree from below, $z = 2$ is the latent response agree from below, $z = 3$ is the latent response agree from above and $z = 4$ is the latent response disgree from above. The response functions are given by

$$P(Y_i = 1 \mid \theta) = \frac{1}{1 + \exp(f) + \exp(g) + \exp(h)},$$

$$P(Y_i = 2 \mid \theta) = \frac{\exp(f)}{1 + \exp(f) + \exp(g) + \exp(h)},$$

$$P(Y_i = 3 \mid \theta) = \frac{\exp(g)}{1 + \exp(f) + \exp(g) + \exp(h)},$$

and

$$P(Y_i = 4 \mid \theta) = \frac{\exp(h)}{1 + \exp(f) + \exp(g) + \exp(h)},$$

in which

$$f = \alpha_i(1(\theta - \beta_i) - \tau_i),$$
$$g = \alpha_i(2(\theta - \beta_i) - \tau_i),$$

$$h = \alpha_i(3(\theta - \beta_i)),$$

and where β_i is the location of the ith item on the latent continuum (in Figure 5.1 $\beta_i = 0.00$); α_i is the discrimination of the ith item ($\alpha_i > 0$); τ_i is the relative location of the subjective response category threshold of the ith item ($\tau_i < 0$); in Figure 5.1 the distance between the threshold of curve 2 and 3, and the threshold of curve 3 and 4 (or 1 and 2).

The dichotomous case for GGUM is a collapsed version of this 4-category GPCM. The four latent category responses $z = 1, ..., 4$ (disagree from below, agree from below, agree from above, disagree from above) collapse into the dichotomous observed responses $j = 0, 1$ (disagree, agree). The two latent disagree categories $z = 1$ and $z = 4$ define the observed disagree response $j = 0$ on an item, whereas the two latent agree responses $z = 2$, and $z = 3$ define the observed agree response $j = 1$ on an item. The probability of scoring in a response category $P(X_i = j \mid \theta)$, is then equal to

$$P(X_i = 0 \mid \theta) = \frac{1 + \exp(h)}{1 + \exp(f) + \exp(g) + \exp(h)},$$

and

$$P(X_i = 1 \mid \theta) = \frac{\exp(f) + \exp(g)}{1 + \exp(f) + \exp(g) + \exp(h)}. \tag{5.1}$$

This parameterization is according to Roberts, Donoghue, and Laughlin (2000).

5.2.2 The collapsed Generalized Partial Credit Model

The collapsed generalized partial credit model (CGPCM) is a collapsed version of the 3-category generalized partial credit model (GPCM; Muraki, 1992). Where GGUM assumes that a response on a dichotomous item is the result of a choice in one of four categories, the CGPCM assumes that a response on a dichotomous item is the result of a choice between three categories $z = 1, 2, 3$; disagree from below ($z = 1$), agree ($z = 2$) and disagree from above ($z = 3$, Figure 5.2). Although the model is slightly simpler the model has a similar interpretation as GGUM. For this model, persons located close to the item location will endorse the item without making a specification whether they are located above or below the item location (curve 2 in Figure 5.2), whereas persons may disagree with the item for one out of two reasons, disagree from below (curve 1 in Figure 5.2) or disagree from above (curve 3 in Figure 5.2).

The probability of scoring in the three latent categories $P(Y_i = z \mid \theta)$ according to the GPCM are given by

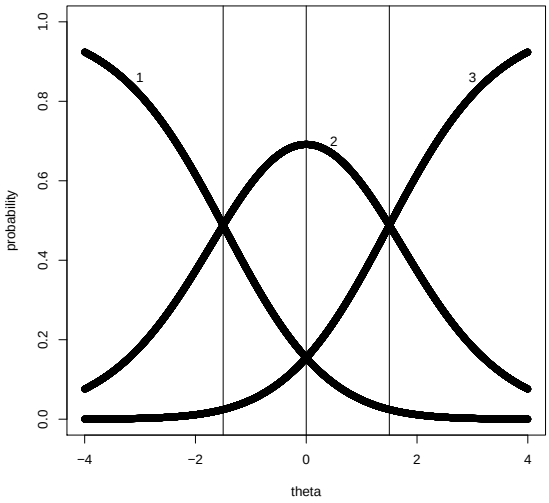

Figure 5.2. 3-category generalized partial credit model with parameter values $\alpha_i = 1.00$, $\beta_i = 0.00$, $\tau_i = -1.50$

$$P(Y_i = 1 \mid \theta) = \frac{1}{1 + \exp(k) + \exp(l)},$$

$$P(Y_i = 2 \mid \theta) = \frac{\exp(k)}{1 + \exp(k) + \exp(l)},$$

and

$$P(Y_i = 3 \mid \theta) = \frac{\exp(l)}{1 + \exp(k) + \exp(l)},$$

in which

$$k = \alpha_i((\theta - \beta_i) - \tau_i),$$

and

$$l = \alpha_i(2(\theta - \beta_i)),$$

and where α_i is the discrimination parameter ($\alpha_i > 0$), β_i is the location parameter, which is equal to the position of the top (in Figure 5.2 $\beta_i = 0.00$), and τ_i is the response category threshold ($\tau_i < 0$), which is equal to the distance between the location parameter and the threshold (in Figure 5.2; $\tau_i = -1.50$).

Under the CGPCM the observed agree response $j = 1$ is not a collapsed response and is equal to the latent agree category response $z = 1$. The

two latent disagree responses, $z = 0$, and $z = 2$ do collapse into one observed disagree response $j = 0$ .The probability of scoring in the observed categories on an item $P(X_i = j \mid \theta)$ is equal to

$$P(X_i = 0 \mid \theta) = \frac{1 + \exp(l)}{1 + \exp(k) + \exp(l)},$$

and

$$P(X_i = 1 \mid \theta) = \frac{\exp(k)}{1 + \exp(k) + \exp(l)}. \tag{5.2}$$

5.2.3 The collapsed Graded Response Model

The collapsed graded response model (CGRM; Korobko, 2007) is the collapsed version of Samejima's (1969) 3-category graded response model (GRM). This model is similar to the CGPCM, in the way that there are three latent response categories $z = 0, 1, 2$ (disagree from below, agree and disagree from above respectively) that collapse into the unfolding dichotomous response model CGRM. The 3-category GRM is shown in Figure 5.3. The two latent disagree responses $z = 0$ (curve 0 in Figure 5.3) and $z = 2$ (curve 2 in Figure 5.3) collapse into an observed disagree response $j = 0$ on the dichotomous item, whereas the observed agree response $j = 1$ is equal to the latent agree response $z = 1$ (curve 1 in Figure 5.3).

Using the abbreviation of the logistic function given by

$$\pi\left(y\right) = \frac{\exp(y)}{1 + \exp(y)},$$

the probability of scoring in the three latent categories, $P(Y_i = z \mid \theta)$ is equal to

$$P(Y_i = 0 \mid \theta) = 1 - \pi_{i1}(m),$$

$$P(Y_i = 1 \mid \theta) = \pi_{i1}(m) - \pi_{i2}(n),$$

and

$$P(Y_i = 2 \mid \theta) = \pi_{i2}(n),$$

in which

$$m = \alpha_i\theta - \beta_{i1},$$

$$n = \alpha_i\theta - \beta_{i2},$$

and α_i is the discrimination parameter ($\alpha_i > 0$), β_{i1} is the location parameter of $1 - \pi_{i1}$ (see Figure 5.3) and β_{i2} is the location parameter

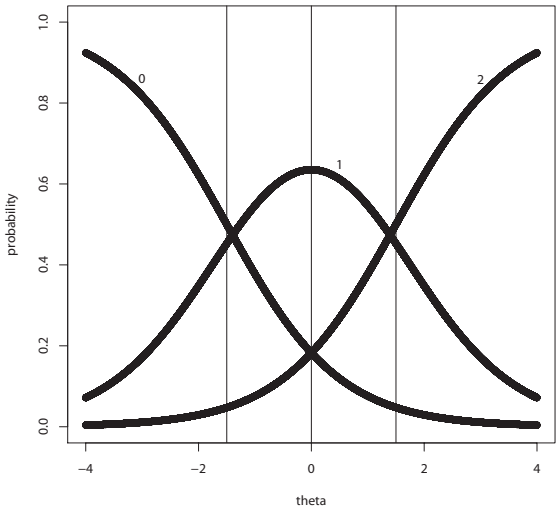

Figure 5.3. 3-category graded response model with parameter values $\alpha_i = 1.00$, $\beta_{i1} = -1.50$, $\beta_{i2} = 1.50$

of π_{i2} (see Figure 5.3). β_{i1} has to be smaller than β_{i2}, and curve 2 is the difference between π_{i1} and π_{i2}.

After collapsing curve 0 and curve 2 into observed response category $j = 0$, the probability of scoring in the collapsed observed categories on an item $P(X_i = j \mid \theta)$ is given by

$$P(X_i = 0 \mid \theta) = 1 - \pi_{i1}(m) + \pi_{i2}(n),$$

and

$$P(X_i = 1 \mid \theta) = \pi_{i1}(m) - \pi_{i2}(n). \tag{5.3}$$

5.2.4 The Quadratic Logistic Regression Model

The Quadratic Logistic Regression Model (QLOG) is different from the former three models, in that QLOG is not a collapsed version of a polytomous dominance IRT model, but a non-linear version of the common linear logistic regression models that are often used in IRT (e.g., 1PLM, 2PLM). The linear models assume a logit function that is equal to a linear regression equation (see Figure 5.4), which results in monotone increasing or decreasing item characteristic curves (ICCs). However, unfolding models consist of both monotone tracelines and single-peaked tracelines. To model ICCs with single-peaked curves a different formula for the logit function is needed. The logit function first has to increase and at a certain point

(the top) has to decrease again. The most well-known function to model a peaked curve is a parabolic regression function (Figure 5.4). Although the monotone tracelines do not seem to match this function, monotone tracelines actually are single-peaked tracelines with the top at minus infinity (decreasing traceline) or plus infinity (increasing traceline).

To describe unfolding models the parabolic regression function is used as the logit function, which is given by

$$logit = \log \frac{P(X_i = 1 \mid \theta)}{P(X_i = 0 \mid \theta)} = \alpha_i\theta + \beta_i + \gamma_i\theta^2,$$

with $\gamma_i < 0$. If

$$o = \alpha_i\theta + \beta_i + \gamma_i\theta^2,$$

the probability for scoring in each category is given by

$$P(X_i = 0 \mid \theta) = \frac{1}{1 + \exp(o)},$$

and

$$P(X_i = 1 \mid \theta) = \frac{\exp(o)}{1 + \exp(o)}. \tag{5.4}$$

5.3 Lagrange Multiplier test

The Lagrange Multiplier test (LM-test; Aitchison & Silvey, 1958), which is equal to the score test (Rao, 1947) or modification index (Sörbom, 1989), can be used to explicitly test against specific violations of the assumptions of an IRT model. The Lagrange Multiplier statistic (LM-statistic) is used to test differences in fit among two nested models; the most restricted model is the null model, which is the IRT model tested, and the alternative model is a more general model, that contains additional freely estimated parameters, which represent model violations. Under the null model the additional parameters of the alternative model are fixed to a constant, which is often equal to zero. So, two models are defined; H_0: the null model with a set of free item and person parameters, η_1, and one or more item and person parameters that are set to a constant value, $\eta_2 = c$, and H_a: an alternative model defined by the same item and person parameters, however, both η_1 and η_2 parameters are estimated freely. The LM-test is used to test the expected change in model fit between both models, however, the test is based on the loglikelihood of only the null model.

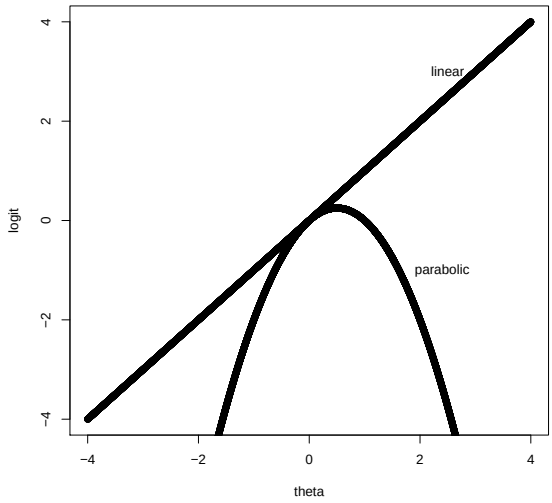

Figure 5.4. logit-function for linear and quadratic logistic regression model with parameter values $\alpha_i = 1.00$, $\beta_i = 0.00$, $\gamma_i = -1.00$

The LM-test statistic evaluates the slope of the likelihood function, L, of the full model, with respect to the values of all parameters of the restricted model. In maximum likelihood estimation the first derivatives $h(\eta_p) = \partial \log L / \partial \eta_p$ of the freely estimated parameters will be equal to zero. When estimating the alternative general model this results in first derivatives equal to zero, whereas the first derivatives of only the η_1 parameters are zero under the null model. The sign of the slope does not make any difference for the expected change in model fit, however, the rate at which it is changing, the curvature of the loglikelihood function, might differ. Therefore the slope is squared, but has to be weighted by the curvature of the loglikelihood function, the hessian, which can be expressed as $H(\eta_p, \eta_q) = \partial^2 \log L / \partial \eta_p \delta \eta_q$. In the LM-statistic, the weighting by the curvature is expressed by the observed values of the hessian, taking into account the influence of estimates of the η_1 parameters. In a formula the LM-statistic is expressed as

$$LM = h(\eta_2)^{'} \Sigma^{-1} h(\eta_2), \tag{5.5}$$

with

$$\Sigma = \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}, \tag{5.6}$$

and

$$\Sigma_{pq} = -\frac{\partial^2 \log L(\eta)}{\partial \eta_p \delta \eta_q'}.$$

The reciprocal of the negative expectation of the hessian by which the first derivatives in the LM-statistic are multiplied express the variance of the maximum likelihood estimators. The matrices $\sum_{pq}$ can be viewed as the asymptotic covariance matrices of the estimates (Glas, 1999).

The LM-statistic is asymptotically chi-square distributed with degrees of freedom equal to the number of additional freely estimated variables under the alternative model. In this paper, the LM-statistic is used to detect person fit under the four unfolding models. Before moving on to the person fit tests the likelihood for the models is explained.

5.3.1 Likelihood

For dichotomous data the likelihood of a response pattern x_n is given by

$$L(\theta, \eta_2 \mid x_n, \eta_1) = \prod_{i=1}^{K} P(\theta, \eta_2)^{x_{ni}} (1 - P(\theta, \eta_2))^{1-x_{ni}}$$

where $P(\theta, \eta_2)$ is the probability of endorsing an item i ($P(X_i = 1 \mid \theta, \eta)$), by the generalized graded unfolding model, collapsed generalized partial credit model, collapsed graded response model, or the quadratic logistic regression model. The loglikelihood is then equal to

$$\log L(\theta, \eta_2 \mid x_{ni}, \eta_1) = \sum_{i=1}^{K} [x_{ni} \log P(\theta, \eta_2) + (1 - x_{ni}) \log(1 - P(\theta, \eta_2))] \quad (5.7)$$

In this study item parameter values will be fixed. The test statistic becomes

$$LM(\eta_2) = \frac{(h)^2}{H} \quad (5.8)$$

in which

$$h = \sum_{i=1}^{K} \frac{\partial \log L}{\partial \eta_2}$$

$$H = \sum_{i=1}^{K} \left\{ -\frac{\partial^2 \log L}{\partial \eta_2^2} + \left(\frac{\partial^2 \log L}{\partial \eta_2 \partial \theta} \right)^2 \left(\frac{\partial^2 \log L}{\partial \theta^2} \right)^{-1} \right\}$$

For derivatives of the loglikelihood see the appendix.

5.3.2 Lagrange Multiplier tests for person fit

Two person fit tests based on the Lagrange Multiplier statistic were developed; a test for constancy of theta, and a test for tendency to agree.

LM-test for constancy of theta

During test administration persons may start unmotivated, become unmotivated (random response), or make mistakes in filling out the questionnaire (misalignment). This might result in an inconsistent pattern of item responses on the test. To check whether these forms of misfit take place the constancy of theta during test administration can be investigated. To test if the theta value for each person is equal in different parts of the test, a persons' response pattern can be divided into a number of subtests S. In this chapter the responses of persons are divided into two groups ($S = 2$).

The LM-test used to investigate constancy of theta compares two models. The general model for the CGRM is given by

$$P(X_i = 1) = \pi_{i1}(\alpha_i[\theta_1 + y_s\delta] - \beta_{i1}) - \pi_{i2}(\alpha_i[\theta_1 + y_s\delta] - \beta_{i2}), \qquad (5.9)$$

in which θ_1 is the estimated latent trait on the first part of the test. A dummy variable, y_s, is used to describe whether the shift, denoted by δ, takes place or not. If an item falls in the first subtest $S = 1$, the dummy variable is equal to $y_s = 0$, and the dummy variable is equal to $y_s = 1$ if the item falls in the second subset $S = 2$. If an item falls in the first subtest no shift is made, but if the items falls in the second subtest a shift on θ takes place. The null model states that during the test no shift in θ takes place, and $\delta = 0$ is assumed. The null model is the model given in equation 5.3. For the other three models the idea is the same, the null models are the models described in section 5.2, and the alternative model is the model in which the thetas of the null model are replaced by $\theta + y_s\delta$.

The LM-test tests the null hypothesis $H_0 : \delta = 0$ against the alternative hypothesis $H_a : \delta \neq 0$. The LM-test for constancy of theta is expressed by equation 5.8, in which $\eta_2 = \delta$. For the derivatives of the loglikelihood for the models see the appendix.

Lagrange Multiplier test for tendency to agree

Other reasons for person misfit are tendency to agree and extreme item responding (tendency to disagree). If persons have a tendency to agree,

they endorse more items than other persons from the population, who have the same trait level. On the other hand, extreme item responding on a dichotomous item might result in only endorsing items that have an item location very close to the person location, and endorsing less items that are located further away of the person location than other persons from the population, who have the same trait level. To model the tendency to agree and disagree the response curve of the agree response can be made higher and broader or lower and smaller respectively. The four unfolding models have different parameters that define the height and width of the curve, only the parameter definitions of the GGUM and CGPCM are similar.

Both GGUM and CGPCM model the height (and width) of the curve by the τ_i-parameter. To heighten (and broaden) the curve a shift δ in the τ_i-parameter has to be made. The alternative model for the CGPCM is expressed by

$$P(X_i = 1 \mid \theta) = \frac{\exp(\alpha_i[(\theta - \beta_i) - \tau_i - \delta)])}{1 + \exp(\alpha_i[(\theta - \beta_i) - \tau_i - \delta)]) + \exp(\alpha_i[2(\theta - \beta_i)])}. \tag{5.10}$$

Under the null model no shift is expected and $\delta = 0$. The null model for the CGPCM-test is the model as expressed in equation 5.2. The same shift δ in the τ_i-parameter is made under the alternative GGUM model in which $-\tau_i$ in equation 5.1 is replaced by $-\tau_i - \delta$. The null model in the GGUM-case with δ equal to 0 is equation 5.1.

For the CGRM the heighth (and width) of the curve is expressed in the distance between the item response curves of π_{i1} and π_{i2}. To increase the distance between both curves, and keep the location of the item equal, a shift δ has to be made in both β_{i1} and β_{i2}-parameters. This shift is expressed in the following formula for the alternative model

$$P(X_i = 1 \mid \theta) = \pi_{i1}(\alpha_i\theta - \beta_{i1} + \delta) - \pi_{i2}(\alpha_i\theta - \beta_{i2} - \delta). \tag{5.11}$$

The null model for the CGRM with $\delta = 0$ is given in equation 5.3.

The QLOG model expresses the logit as a parabolic function. The height (and width) of the response function under this model can be changed through a shift δ in the β_i-parameter. The alternative model for QLOG is given by

$$P(X_i = 1 \mid \theta) = \frac{\exp(\alpha_i\theta + \beta_i + \gamma_i\theta^2 + \delta)}{1 + \exp(\alpha_i\theta + \beta_i + \gamma_i\theta^2 + \delta)}. \tag{5.12}$$

and the null model under which $\delta = 0$ is given by equation 5.4.

The LM-test for tendency to agree then tests the alternative hypothesis $H_a : \delta \neq 0$ against the null hypothesis $H_0 : \delta = 0$. This test is given by

equation 5.8, in which $\eta_2 = \delta$. For the derivatives of the log-likelihood for the four models see the appendix.

5.4 Simulation study

Simulation studies of Type I error rate and power for the LM-tests for constancy of theta and tendency to agree were conducted. In the simulation studies fixed item parameters were used. The values of the item parameters were based on a real 20-item dataset measuring Censorship (Roberts, 1995). The real dataset was used to estimate person parameters with the original GGUM program (Roberts, Donoghue, & Laughlin, 2000). The item responses and the estimated person parameters of the original GGUM program were fixed at these values and item parameters were re-estimated for the GGUM model and estimated for each of the other three unfolding models. These item parameters were used as fixed parameters for the simulation study. Person parameters for $10,000$ persons were drawn from a normal distribution with mean zero and variance one. Data were generated under one model and based on the generated data and fixed item parameters, person parameters were estimated and LM-tests conducted under the four models. Type I error rate for both LM-tests, power of both LM-tests and agreement between models in flagging persons as fitting and misfitting were studied.

5.4.1 Type I error rate for LM-test of constancy of theta and LM-test of tendency to agree

In the simulation study on Type I error rate, the probability of rejecting the null model when this model is true was studied. A nominal significance level of 5% was used. The number of items, K, was varied from 10 to 80, in steps of 10. For $K = 10$ only the parameters of the first 10 items of the inventory were used, for $K = 20$, all items of the questionnaire were used, for $K = 40, 60$ and 80 the item parameters used were two, three, and four times the whole set respectively, and for $K = 30, 50$ and 70 one, two, and three times the whole set of items plus the first ten items were used respectively. Data were generated under the null models.

Table 5.1

Results of Type I error rate study of LM1 and LM2 test

		Estimation Model							
Generating		GGUM		CGPCM		CGRM		QLOG	
Model	K	LM1	LM2	LM1	LM2	LM1	LM2	LM1	LM2
GGUM	10	0.040	0.036	0.049	0.039	0.057	0.047	0.055	0.034
	20	0.049	0.046	0.053	0.044	0.051	0.051	0.058	0.048
	30	0.053	0.046	0.058	0.044	0.053	0.054	0.062	0.050
	40	0.051	0.045	0.054	0.046	0.048	0.062	0.060	0.052
	50	0.054	0.048	0.055	0.050	0.050	0.063	0.063	0.057
	60	0.046	0.047	0.050	0.051	0.047	0.063	0.055	0.058
	70	0.048	0.048	0.050	0.053	0.046	0.072	0.059	0.061
	80	0.049	0.052	0.051	0.052	0.048	0.073	0.060	0.066
CGPCM	10	0.037	0.029	0.041	0.032	0.048	0.036	0.040	0.029
	20	0.050	0.046	0.051	0.044	0.048	0.043	0.053	0.045
	30	0.050	0.054	0.050	0.043	0.044	0.044	0.055	0.055
	40	0.051	0.053	0.053	0.049	0.049	0.051	0.059	0.056
	50	0.051	0.063	0.052	0.048	0.046	0.050	0.060	0.065
	60	0.051	0.061	0.050	0.050	0.046	0.050	0.059	0.063
	70	0.050	0.061	0.053	0.047	0.046	0.051	0.062	0.064
	80	0.051	0.061	0.053	0.046	0.047	0.054	0.059	0.063
CGRM	10	0.050	0.034	0.057	0.041	0.059	0.040	0.049	0.040
	20	0.063	0.059	0.065	0.057	0.056	0.054	0.063	0.064
	30	0.060	0.064	0.062	0.052	0.051	0.047	0.063	0.063
	40	0.059	0.067	0.058	0.061	0.051	0.053	0.066	0.074
	50	0.060	0.083	0.060	0.062	0.051	0.050	0.066	0.085
	60	0.061	0.072	0.061	0.064	0.048	0.047	0.067	0.084
	70	0.059	0.084	0.059	0.066	0.051	0.049	0.063	0.092
	80	0.056	0.083	0.058	0.066	0.048	0.048	0.064	0.093
QLOG	10	0.040	0.032	0.042	0.034	0.050	0.037	0.039	0.031
	20	0.051	0.049	0.052	0.047	0.051	0.048	0.047	0.045
	30	0.047	0.052	0.047	0.044	0.045	0.049	0.049	0.042
	40	0.050	0.051	0.051	0.048	0.047	0.052	0.053	0.047
	50	0.049	0.061	0.048	0.054	0.045	0.055	0.051	0.052
	60	0.050	0.055	0.049	0.052	0.047	0.055	0.053	0.049
	70	0.043	0.058	0.044	0.053	0.039	0.055	0.047	0.046
	80	0.046	0.064	0.047	0.055	0.041	0.058	0.050	0.049

Table 5.1 gives the Type I error rate of the LM-test for constancy of theta (LM1) and of the LM-test for tendency to agree (LM2). The first column "Generating model" gives the model under which the data were generated on K items (column 2). Columns 3 to 10 give the test-results under the "estimation models". Proportions of rejections in $1,000$ replications at the 5%-level on the first and second LM-test are given in column 3 and 4 for GGUM, column 5 and 6 for CGPCM, column 7 and 8 for CGRM, and column 9 and 10 for QLOG. The results showed that the Type I error rates for both tests attained the 5% significance level. Estimation with the wrong model still gave an acceptable Type I error rate for most combinations of generating and estimation model. Only some minor deviations were found for the results on the LM2-test for GGUM and QLOG when the CGRM was the generating model. When the number of items increased the LM2-test results increased to .08 and .09 for the GGUM and QLOG model respectively. However, the deviations were still relatively small.

Table 5.2

Results of power study of LM1-test (K=20)

		Estimation Model							
Generating		GGUM		CGPCM		CGRM		QLOG	
Model	δ	LM1	LM2	LM1	LM2	LM1	LM2	LM1	LM2
GGUM	0.5	0.129	0.052	0.137	0.052	0.115	0.063	0.141	0.057
	0.8	0.242	0.061	0.257	0.066	0.217	0.073	0.258	0.069
	1.0	0.326	0.064	0.349	0.071	0.299	0.075	0.351	0.079
	1.2	0.416	0.079	0.442	0.085	0.391	0.092	0.447	0.097
CGPCM	0.5	0.132	0.051	0.142	0.049	0.118	0.053	0.142	0.056
	0.8	0.232	0.057	0.251	0.060	0.213	0.066	0.254	0.065
	1.0	0.326	0.068	0.352	0.071	0.302	0.076	0.350	0.080
	1.2	0.408	0.085	0.444	0.089	0.390	0.092	0.446	0.101
CGRM	0.5	0.133	0.060	0.142	0.058	0.116	0.052	0.144	0.063
	0.8	0.232	0.070	0.255	0.072	0.219	0.062	0.253	0.077
	1.0	0.326	0.085	0.353	0.089	0.311	0.071	0.344	0.095
	1.2	0.410	0.100	0.440	0.108	0.401	0.077	0.435	0.114
QLOG	0.5	0.117	0.049	0.125	0.050	0.105	0.059	0.130	0.049
	0.8	0.219	0.058	0.238	0.062	0.200	0.073	0.248	0.061
	1.0	0.310	0.069	0.339	0.074	0.288	0.087	0.352	0.075
	1.2	0.409	0.076	0.444	0.082	0.384	0.096	0.461	0.086

5.4.2 Power of LM-test for constancy of theta

Power is the probability of detecting misfit or rejecting the null model when the alternative model is true for a person. In this simulation study power of the LM-test on constancy of theta (LM1) was investigated for a test of 20 and 40 items. For each person data on the first half of the test were generated under the null model, whereas data on the second half of the test were generated under the alternative model, with a shift in theta equal to $\delta = 0.5$, $\delta = 0.8$, $\delta = 1.0$, or $\delta = 1.2$. Data were generated under each model, and person parameters were estimated and LM-tests were computed under the four models. Item parameters were equal to the values in the previous study. Table 5.2 shows the results for a shift in theta for $K = 20$ and Table 5.3 shows the results for $K = 40$. The setup of the tables was similar to the setup of the Type I error rate table. The only difference is that in column 2 the δ-values for the shift in ability are given. Furthermore the results in the columns for the LM1-test give the power of the constancy of theta test, whereas the results in the columns of the LM2-test are the Type I error rates of the tendency to agree test.

In general, the LM1-test had reasonable power to detect model violations of constancy of theta. When the number of items and effect size increased the power increased. The different models showed similar power results in detecting misfitting persons, independent of the model used for generating the items, however power under CGPCM and QLOG was slightly higher than GGUM, which was slightly higher than CGRM. Type I error rate of the LM2-test for tendency to agree stayed relatively stable when the number of items increased and slightly increased when the effect size increased, however, Type I error rate was still maximum around .10. Type I error rate was slightly higher for GGUM, CGPCM and QLOG when CGRM was the generating model. However, results were still reasonable.

Table 5.3
Results of power study of LM1-test (K=40)

		Estimation Model							
Generating		GGUM		CGPCM		CGRM		QLOG	
Model	δ	LM1	LM2	LM1	LM2	LM1	LM2	LM1	LM2
GGUM	0.5	0.193	0.049	0.193	0.051	0.179	0.059	0.198	0.056
	0.8	0.398	0.051	0.396	0.059	0.371	0.063	0.394	0.066
	1.0	0.532	0.054	0.527	0.069	0.497	0.069	0.524	0.077
	1.2	0.656	0.061	0.638	0.090	0.619	0.085	0.641	0.095
CGPCM	0.5	0.191	0.051	0.202	0.046	0.185	0.051	0.205	0.059
	0.8	0.394	0.054	0.416	0.049	0.385	0.050	0.414	0.064
	1.0	0.529	0.064	0.553	0.060	0.518	0.061	0.551	0.078
	1.2	0.653	0.072	0.687	0.070	0.648	0.070	0.678	0.094
CGRM	0.5	0.186	0.067	0.192	0.066	0.179	0.047	0.192	0.079
	0.8	0.373	0.071	0.389	0.072	0.378	0.052	0.386	0.088
	1.0	0.512	0.076	0.529	0.082	0.527	0.055	0.520	0.100
	1.2	0.620	0.087	0.633	0.104	0.650	0.068	0.622	0.122
QLOG	0.5	0.177	0.052	0.184	0.051	0.168	0.054	0.196	0.049
	0.8	0.373	0.045	0.394	0.047	0.364	0.053	0.425	0.045
	1.0	0.513	0.054	0.535	0.057	0.504	0.069	0.577	0.056
	1.2	0.641	0.057	0.669	0.064	0.632	0.071	0.703	0.068

5.4.3 Power of LM-test of tendency to agree

In the simulation study on power of the LM-test of tendency to agree (LM2) all item responses for all persons were generated under the alternative model. Equal shifts in δ do not result in the same shift of the curve under the different models. Comparable shift values were searched for. They are shown in Table 5.4. Similar to the study on the power of the LM1 test, data were generated under each model and person parameters were estimated and LM-tests were computed under the four models for 20 and 40 items.

Results are given in Table 5.5 for $K = 20$ and in Table 5.6 for $K = 40$. Setup of the tables is equal to the setup of the tables of the power study on the LM1-test, except that the columns labeled LM2 show the power of the LM2-test, and the columns labeled LM1 show the Type I error rate of the LM1-test.

Table 5.4

Comparable values for shift δfor the four models

GGUM	CGPCM	CGRM	QLOG
-0.30	-0.30	0.25	0.30
-0.50	-0.50	0.45	0.50
-0.80	-0.80	0.70	0.80
-1.00	-1.00	0.90	1.00
-1.20	-1.20	1.10	1.20

Table 5.5

Results of power study of LM2-test (K=20)

		Estimation Model							
Generating		GGUM		CGPCM		CGRM		QLOG	
Model	δ	LM1	LM2	LM1	LM2	LM1	LM2	LM1	LM2
GGUM	-0.30	0.046	0.064	0.049	0.058	0.048	0.066	0.052	0.056
	-0.50	0.042	0.102	0.047	0.089	0.044	0.096	0.051	0.086
	-0.80	0.038	0.197	0.040	0.169	0.038	0.165	0.045	0.162
	-1.00	0.034	0.283	0.037	0.247	0.034	0.235	0.044	0.239
	-1.20	0.032	0.364	0.032	0.326	0.032	0.315	0.039	0.320
CGPCM	-0.30	0.045	0.071	0.046	0.062	0.043	0.060	0.049	0.061
	-0.50	0.041	0.106	0.043	0.093	0.044	0.093	0.048	0.089
	-0.80	0.039	0.197	0.041	0.172	0.038	0.168	0.045	0.170
	-1.00	0.035	0.274	0.037	0.244	0.038	0.237	0.041	0.238
	-1.20	0.032	0.360	0.032	0.328	0.035	0.309	0.038	0.320
CGRM	0.25	0.057	0.088	0.059	0.078	0.049	0.074	0.057	0.081
	0.45	0.054	0.133	0.058	0.116	0.051	0.114	0.060	0.116
	0.70	0.048	0.227	0.051	0.200	0.045	0.202	0.054	0.196
	0.90	0.048	0.321	0.049	0.285	0.044	0.288	0.053	0.278
	1.10	0.038	0.400	0.036	0.365	0.033	0.364	0.041	0.357
QLOG	0.30	0.043	0.073	0.043	0.065	0.042	0.064	0.041	0.062
	0.50	0.044	0.117	0.043	0.101	0.039	0.102	0.041	0.097
	0.80	0.037	0.206	0.036	0.184	0.034	0.178	0.038	0.180
	1.00	0.030	0.281	0.028	0.253	0.028	0.245	0.029	0.248
	1.20	0.028	0.385	0.027	0.352	0.027	0.331	0.029	0.347

Table 5.6
Results of power study of LM2-test (K=40)

		Estimation Model							
Generating		GGUM		CGPCM		CGRM		QLOG	
Model	δ	LM1	LM2	LM1	LM2	LM1	LM2	LM1	LM2
GGUM	-0.30	0.047	0.089	0.048	0.084	0.046	0.089	0.057	0.088
	-0.50	0.042	0.170	0.044	0.154	0.040	0.146	0.051	0.160
	-0.80	0.035	0.351	0.040	0.324	0.037	0.295	0.049	0.333
	-1.00	0.028	0.495	0.031	0.462	0.030	0.424	0.041	0.469
	-1.20	0.030	0.642	0.034	0.612	0.029	0.567	0.041	0.617
CGPCM	-0.30	0.044	0.094	0.045	0.084	0.041	0.080	0.051	0.096
	-0.50	0.041	0.173	0.044	0.158	0.041	0.141	0.053	0.168
	-0.80	0.033	0.358	0.035	0.339	0.033	0.300	0.045	0.348
	-1.00	0.030	0.492	0.034	0.467	0.030	0.421	0.044	0.479
	-1.20	0.028	0.618	0.033	0.599	0.029	0.550	0.040	0.606
CGRM	0.25	0.056	0.120	0.056	0.106	0.048	0.094	0.060	0.121
	0.45	0.051	0.218	0.052	0.198	0.044	0.182	0.059	0.213
	0.70	0.046	0.395	0.047	0.367	0.041	0.349	0.054	0.380
	0.90	0.037	0.542	0.040	0.510	0.033	0.490	0.047	0.527
	1.10	0.037	0.666	0.042	0.640	0.034	0.623	0.047	0.655
QLOG	0.30	0.045	0.094	0.047	0.086	0.043	0.081	0.051	0.089
	0.50	0.040	0.175	0.041	0.164	0.039	0.145	0.046	0.167
	0.80	0.032	0.355	0.033	0.337	0.031	0.303	0.038	0.351
	1.00	0.026	0.498	0.028	0.477	0.025	0.436	0.034	0.498
	1.20	0.022	0.627	0.026	0.613	0.024	0.573	0.030	0.634

It was shown that if number of items increased and if δ increased, the power of the LM2-test increased. Power of the GGUM estimation model was slightly higher than for the QLOG and CGPCM, which were slightly higher than the CGRM. Results were highest for the estimation models when CGRM was the generating model. Note that this was also the case for the Type I error rate of the LM2 test. Type I error rate values of the LM1-test were approximately the significance level of 5%, although values slightly decreased when δ increased. Values for the QLOG estimation model were slightly higher than for the other models.

5.4.4 Agreement between models

To check the robustness of the models, the degree of agreement between the models to detect fitting and misfitting response patterns is investigated. Agreement between models was high for all values of δ and K. Table 5.7 and 5.8 give the results of the degree of agreement between models for the LM-test for constancy of theta, and the LM-test for tendency to agree, respectively, for 20 and 40 item tests with δ equal to 0.5 and 1.0. Results for the LM1 power studies showed that agreement between all models was relatively stable when number of items increased, and decreased when δ increased. Results were similar under the generating models GGUM, CGPCM, and QLOG, and slightly lower when CGRM was the generating model. The degree of agreement between QLOG and the other models was highest in almost all analyses. Results on the power tests on LM2 showed that the agreement on the LM2 test slightly decreased when the number of items or delta increased. Results over generating models were similar, however for 40 items the CGRM results were slightly higher. Under all analyses highest degree of agreement was found between QLOG, GGUM and CGPCM.

5.5 Discussion

In this paper two person fit statistics based on the Lagrange Multiplier statistic were developed for four unfolding models for typical performance data. Simulation studies showed that Type I error rates on the LM-test for constancy of theta were reasonable for all models, and power increased when K and δ increased. QLOG had the highest degree of agreement with the other models, however, Type I error rate was highest as well. The Type I error rate for the LM-test for tendency to agree was around 5%, but increased slightly when misfit in constancy of theta was modeled. Power on the LM-test for tendency to agree increased with K and δ. Degree of agreement was slightly higher between GGUM, CGPCM and QLOG than between these three models and CGRM. Thus, in general the results of the simulation studies on both LM-tests were similar for all four models, and there is evidence that the models can be seen as comparable.

The present paper is a first contribution to detect person misfit under four unfolding IRT models. Only two specific types of person misfit are discussed, which are specified by the alternative model. This method can be used to develop tests to investigate other types of person misfit, like local

Table 5.7
Correlations between models for LM1 test

Generating Model	Estimation Models		$K = 20$ $\delta = 0.5$	$K = 20$ $\delta = 1.0$	$K = 40$ $\delta = 0.5$	$K = 40$ $\delta = 1.0$
GGUM	GGUM	- CGPCM	0.932	0.879	0.923	0.870
	GGUM	- CGRM	0.945	0.900	0.942	0.890
	GGUM	- QLOG	0.964	0.934	0.961	0.925
	CGPCM	- CGRM	0.932	0.882	0.932	0.885
	CGPCM	- QLOG	0.946	0.910	0.942	0.909
	CGRM	- QLOG	0.962	0.927	0.956	0.920
CGPCM	GGUM	- CGPCM	0.935	0.887	0.920	0.868
	GGUM	- CGRM	0.947	0.902	0.944	0.899
	GGUM	- QLOG	0.965	0.937	0.961	0.935
	CGPCM	- CGRM	0.937	0.883	0.928	0.876
	CGPCM	- QLOG	0.946	0.910	0.939	0.900
	CGRM	- QLOG	0.964	0.930	0.959	0.930
CGRM	GGUM	- CGPCM	0.920	0.867	0.916	0.863
	GGUM	- CGRM	0.941	0.883	0.935	0.901
	GGUM	- QLOG	0.962	0.931	0.955	0.932
	CGPCM	- CGRM	0.929	0.869	0.928	0.871
	CGPCM	- QLOG	0.937	0.890	0.935	0.901
	CGRM	- QLOG	0.955	0.914	0.951	0.913
QLOG	GGUM	- CGPCM	0.931	0.879	0.919	0.883
	GGUM	- CGRM	0.944	0.896	0.941	0.907
	GGUM	- QLOG	0.964	0.937	0.963	0.942
	CGPCM	- CGRM	0.932	0.884	0.928	0.891
	CGPCM	- QLOG	0.941	0.902	0.937	0.908
	CGRM	- QLOG	0.961	0.925	0.952	0.929

Table 5.8

Correlations between models for LM2 test

Generating Model	Estimation Models		$K = 20$		$K = 40$	
			$\delta = 0.5$	$\delta = 1.0$	$\delta = 0.5$	$\delta = 1.0$
GGUM	GGUM	- CGPCM	0.966	0.938	0.954	0.935
	GGUM	- CGRM	0.938	0.892	0.921	0.875
	GGUM	- QLOG	0.977	0.958	0.967	0.955
	CGPCM	- CGRM	0.948	0.914	0.918	0.878
	CGPCM	- QLOG	0.977	0.964	0.969	0.957
	CGRM	- QLOG	0.945	0.909	0.928	0.888
CGPCM	GGUM	- CGPCM	0.967	0.946	0.958	0.935
	GGUM	- CGRM	0.938	0.893	0.931	0.888
	GGUM	- QLOG	0.976	0.961	0.968	0.953
	CGPCM	- CGRM	0.951	0.916	0.935	0.887
	CGPCM	- QLOG	0.978	0.966	0.970	0.952
	CGRM	- QLOG	0.946	0.906	0.939	0.897
CGRM	GGUM	- CGPCM	0.955	0.934	0.950	0.939
	GGUM	- CGRM	0.931	0.897	0.922	0.901
	GGUM	- QLOG	0.974	0.958	0.964	0.958
	CGPCM	- CGRM	0.940	0.908	0.916	0.900
	CGPCM	- QLOG	0.970	0.960	0.963	0.954
	CGRM	- QLOG	0.939	0.903	0.928	0.907
QLOG	GGUM	- CGPCM	0.969	0.948	0.958	0.940
	GGUM	- CGRM	0.946	0.903	0.933	0.889
	GGUM	- QLOG	0.981	0.965	0.972	0.957
	CGPCM	- CGRM	0.951	0.923	0.933	0.893
	CGPCM	- QLOG	0.980	0.968	0.970	0.957
	CGRM	- QLOG	0.952	0.913	0.939	0.900

independence and multidimensionality. Furthermore it would be important to compare person fit statistics for unfolding models based on the LM-test with person fit statistics formerly developed for dominance IRT models as the likelihood statistic l_z (Drasgow, Levine, & Williams, 1985), the Pearson-type W-statistic (Wright & Stone, 1979), and Snijders' (2001) person fit test l_z.

Furthermore, results of the simulation studies show that the power to detect misfitting persons who show small shifts in constancy of theta or tendency to agree is low for 20-item and 40-item inventories. Typical performance measures often consist of only a small number of items (< 20). An option to detect person misfit better, and enhance power, might be to take into account information on external variables. This was already shown helpfull for person fit tests for dominance IRT models in Glas and Dagohoy (2007).

Although a lot is left to be done, at this time the two person fit statistics developed in this chapter show that detection of persons with inconsistent response patterns, and tendency to (dis)agree is possible under unfolding IRT models making use from LM-statistics.

Appendix

Detailed characterizations of the test statistics

The Lagrange Multiplier statistic (LM-statistic) for the differences in fit between two nested models is defined by

$$LM = h(\eta_2)^{'} \Sigma^{-1} h(\eta_2). \tag{5.13}$$

The test statistic is expressed by first and second derivatives of the loglikelihood function for a response pattern x_n on the alternative model. The loglikelihood of a response pattern for dichotomous data on the general model is given by

$$\log L = \sum_{i=1}^{K} \left[x_{ni} \log P(\theta, \eta_2) + (1 - x_{ni}) \log \left(1 - P(\theta, \eta_2)\right) \right], \tag{5.14}$$

where $P(\theta, \eta_2)$ is the probability of endorsing item i, $P(X_i = 1 \mid \theta)$, by the generalized graded unfolding model (GGUM), collapsed generalized partial credit model (CPCM), collapsed graded response model (CGRM), or the quadratic logistic regression model (QLOG). The first derivative of this

loglikelihood with respect to a parameter is given by

$$\frac{\partial \log L}{\partial \eta} = \sum_{i=1}^{K} \frac{P'(x_{ni} - P)}{P(1-P)}, \tag{5.15}$$

and the second derivative of the loglikelihood with respect to the parameters is given by

$$\frac{\partial^2 \log L}{\partial \eta^2} = -\sum_{i=1}^{K} \frac{\left(P'\right)^2}{P(1-P)}. \tag{5.16}$$

Derivatives of the loglikelihood for Lagrange Multiplier test for constancy of theta

Let the first and second derivatives of the log-likelihood with respect to θ be defined as follows

$$d_{ij} = \frac{\partial \log L}{\partial \theta}, \tag{5.17}$$

and

$$D_{ij} = \frac{\partial^2 \log L}{\partial \theta^2}. \tag{5.18}$$

The first and second derivatives of the log-likelihood with respect to parameter δ are then equal to

$$\frac{\partial \log L}{\partial \delta} = y_s d_{ij}, \tag{5.19}$$

$$\frac{\partial^2 \log L}{\partial \theta \partial \delta} = y_s D_{ij}, \tag{5.20}$$

and

$$\frac{\partial^2 \log L}{\partial \delta^2} = y_s^2 D_{ij} \tag{5.21}$$

For the four models the d_{ij} and D_{ij} differ. The first and second derivatives with respect to θ for all four models are given here. Derivatives for the loglikelihood with respect to δ can be computed by equations 5.19 to 5.21.

Derivatives for the Generalized Graded Unfolding Model

We introduce a concise notation

$$e1 = \exp(\alpha_i[(\theta + y_s\delta - \beta_i) - \tau_i]),$$

$$e2 = \exp(\alpha_i[2(\theta + y_s\delta - \beta_i) - \tau_i]),$$

and

$$e3 = \exp(\alpha_i[3(\theta + y_s\delta - \beta_i)]).$$

The first and second derivatives for the loglikelihood with respect to θ are

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{K} \frac{\alpha_i[e1 - 2e1e3 + 2e2 - e2e3][x_i - P]}{(e1 + e2)(1 + e3)}, \tag{5.22}$$

and

$$\frac{\partial^2 \log L}{\partial \theta^2} = \sum_{i=1}^{K} \frac{[\alpha_i(e1 - 2e1e3 + 2e2 - e2e3)]^2}{(e1 + e2)(1 + e3)(1 + e1 + e2 + e3)^2}. \tag{5.23}$$

Derivatives for the Collapsed Generalized Partial Credit Model

We introduce a concise notation

$$e1 = \exp(\alpha_i[(\theta + y_s\delta - \beta_i) - \tau_i]),$$

and

$$e2 = \exp(\alpha_i[2(\theta + y_s\delta - \beta_i)]).$$

Then the first and second derivatives for the loglikelihood with respect to θ are

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{K} \frac{\alpha_i e1(1 - e2)[x_i - P]}{e1(1 + e2)}, \tag{5.24}$$

and

$$\frac{\partial^2 \log L}{\partial \theta^2} = \sum_{i=1}^{K} \frac{[\alpha_i e1(1 - e2)]^2}{e1(1 + e2)(1 + e1 + e2)^2}. \tag{5.25}$$

Derivatives for the Collapsed Graded Response Model

When the abbreviation of the logistic function is given by

$$\pi(x) = \frac{\exp(x)}{1 + \exp(x)},$$

and $x_{i1} = \alpha_i(\theta + y_s\delta - \beta_{i1})$, and $x_{i2} = \alpha_i(\theta + y_s\delta - \beta_{i2})$, the first and second derivatives for the loglikelihood with respect to θ are

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{K} \frac{[\alpha_i\pi_{i1}(1 - \pi_{i1}) - \alpha_i\pi_{i2}(1 - \pi_{i2})][x_{ni} - (\pi_{i1} - \pi_{i2})]}{(\pi_{i1} - \pi_{i2})(1 - \pi_{i1} + \pi_{i2})}, \tag{5.26}$$

and

$$\frac{\partial^2 \log L}{\partial \theta^2} = \sum_{i=1}^{K} \frac{\left[\alpha_i \pi_{i1} \left(1 - \pi_{i1}\right) - \alpha_i \pi_{i2} \left(1 - \pi_{i2}\right)\right]^2}{\left(\pi_{i1} - \pi_{i2}\right)\left(1 - \pi_{i1} + \pi_{i2}\right)}. \tag{5.27}$$

Derivatives for the Quadratic Logistic Regression Model

The first and second derivatives for the loglikelihood with respect to θ are

$$\frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{K} (\alpha_i + 2\gamma_i \theta + 2\gamma_i y \delta)(x_{ni} - P), \tag{5.28}$$

and

$$\frac{\partial^2 \log L}{\partial \theta^2} = \sum_{i=1}^{K} (\alpha_i + 2\gamma_i \theta + 2\gamma_i y \delta)^2 PQ. \tag{5.29}$$

Derivatives of loglikelihood for Lagrange Multiplier test for tendency to agree

For the four models the shifts δ are linked to different parameters, only the parameter-shifts under the GGUM and CGPCM are similar. No general formula for the derivatives can be given, so the derivatives for each model are explained separately.

Derivatives for the Generalized Graded Unfolding Model

A concise notation is used

$$e1 = \exp(\alpha_i[(\theta - \beta_i) - \tau_i - \delta]),$$

$$e2 = \exp(\alpha_i[2(\theta - \beta_i) - \tau_i - \delta]),$$

and

$$e3 = \exp(\alpha_i[3(\theta - \beta_i)]).$$

The first derivatives of the loglikelihood with respect to δ and second derivatives for the loglikelihood with respect to θ and δ are equal to

$$\frac{\partial \log L}{\partial \delta} = -\sum_{i=1}^{K} (x_{ni} - P), \tag{5.30}$$

$$\frac{\partial^2 \log L}{\partial \delta^2} = \sum_{i=1}^{K} PQ, \tag{5.31}$$

$$\frac{\partial^2 \log L}{\partial\theta\partial\delta} = -\sum_{i=1}^{K} \frac{\alpha_i(e1 - 2e1e3 + 2e2 - e2e3)}{(1 + e1 + e2 + e3)^2}, \tag{5.32}$$

and

$$\frac{\partial^2 \log L}{\partial\theta^2} = \sum_{i=1}^{K} \frac{[\alpha_i(e1 - 2e1e3 + 2e2 - e2e3)]^2}{(e1 + e2)(1 + e3)(1 + e1 + e2 + e3)^2}. \tag{5.33}$$

Derivatives for the collapsed Generalized Partial Credit Model

A concise notation is used

$$e1 = \exp(\alpha_i[(\theta - \beta_i) - \tau_i - \delta]),$$

and

$$e2 = \exp(\alpha_i[2(\theta - \beta_i)]).$$

The first derivative for the loglikelihood with respect to δ and second derivatives for the loglikelihood with respect to δ and θ are

$$\frac{\partial \log L}{\partial\delta} = -\sum_{i=1}^{K}(x_{ni} - P), \tag{5.34}$$

$$\frac{\partial^2 \log L}{\partial\delta^2} = \sum_{i=1}^{K} PQ, \tag{5.35}$$

$$\frac{\partial^2 \log L}{\partial\delta\partial\theta} = -\sum_{i=1}^{K} \frac{\alpha_i e1(1 - e2)}{(1 + e1 + e2)^2}, \tag{5.36}$$

$$\frac{\partial^2 \log L}{\partial\theta^2} = \sum_{i=1}^{K} \frac{[\alpha_i e1(1 - e2)]^2}{e1(1 + e2)(1 + e1 + e2)^2}. \tag{5.37}$$

Derivatives for the collapsed Graded Response Model

When the abbreviation of the logistic function is given by

$$\pi(x) = \frac{\exp(x)}{1 + \exp(x)},$$

and $x_{i1} = \alpha_i(\theta - \beta_{i1} + \delta)$ and $x_{i2} = \alpha_i(\theta - \beta_{i2} - \delta)$ and , the first derivative with respect to δ and second derivatives for the loglikelihood with respect to δ and θ are

$$\frac{\partial \log L}{\partial\delta} = \sum_{i=1}^{K} \frac{[\pi_{i1}\,(1 - \pi_{i1}) + \pi_{i2}(1 - \pi_{i2})][x_{ni} - (\pi_{i1} - \pi_{i2})]}{(\pi_{i1} - \pi_{i2})\,(1 - \pi_{i1} + \pi_{i2})}, \tag{5.38}$$

$$\frac{\partial^2 \log L}{\partial \delta^2} = \sum_{i=1}^{K} \frac{\left[\pi_{i1}\left(1-\pi_{i1}\right)+\pi_{i2}(1-\pi_{i2})\right]^2}{\left(\pi_{i1}-\pi_{i2}\right)\left(1-\pi_{i1}+\pi_{i2}\right)}, \tag{5.39}$$

$$\frac{\partial^2 \log L}{\partial \delta \partial \theta} = \sum_{i=1}^{K} \frac{\alpha_i([\pi_{i1}\left(1-\pi_{i1}\right)]^2-[\pi_{i2}\left(1-\pi_{i2}\right)]^2)}{\left(\pi_{i1}-\pi_{i2}\right)\left(1-\pi_{i1}+\pi_{i2}\right)}, \tag{5.40}$$

and

$$\frac{\partial^2 \log L}{\partial \theta^2} = \sum_{i=1}^{K} \frac{\left(\alpha_i\left[\pi_{i1}\left(1-\pi_{i1}\right)-\pi_{i2}\left(1-\pi_{i2}\right)\right]\right)^2}{\left(\pi_{i1}-\pi_{i2}\right)\left(1-\pi_{i1}+\pi_{i2}\right)}. \tag{5.41}$$

Derivatives for the quadratic logistic regression model

The first derivatives for the loglikelihood with respect to δ and second derivatives for the loglikelihood with respect to δ and θ are

$$\frac{\partial \log L}{\partial \delta} = \sum_{i=1}^{K}(x_{ni}-P), \tag{5.42}$$

$$\frac{\partial^2 \log L}{\partial \delta^2} = \sum_{i=1}^{K} PQ, \tag{5.43}$$

$$\frac{\partial^2 \log L}{\partial \delta \partial \theta} = \sum_{i=1}^{K}(\alpha_i + 2\gamma_i\theta)PQ, \tag{5.44}$$

and

$$\frac{\partial^2 \log L}{\partial \theta^2} = \sum_{i=1}^{K}\left(\alpha_i + 2\gamma_i\theta\right)^2 P_iQ_i. \tag{5.45}$$

Chapter 6

Item fit for unfolding IRT models

6.1 Introduction

Analogous to dominance IRT models, unfolding IRT models follow two major assumptions: unidimensionality and local independence. Besides these two major assumptions the shape of the item response function or item characteristic curve (ICC) is assumed to be single-peaked. In this chapter, the focus is on items which require a dichotomous response, that is, the item can be endorsed or not endorsed. Persons only endorse an item if the location of the person parameter is close to the mode of the ICC. Such unfolding models are used to describe measures on typical performance inventories, like attitude and personality measures.

The first major assumption, unidimensionality, implies that the answers to items are based on only one underlying construct. Items share variance because of this one underlying construct, and after taking into account the score on this construct, covariance between item responses of a single respondent are assumed to be zero. This implies that item responses are independent given the value of the latent variable. The only dependency between items over persons is attributable to the dimension they are assumed to measure. The assumption on the shape of the ICC means that the probability of endorsing an item is expected to increase when trait level increases up till a certain point where the curve reaches its maximum, that is, the top of the curve or the ideal point. The curve decreases when moving along on the latent trait continuum from this ideal point onward. Items are therefore single-peaked, monotone increasing or monotone decreasing. For

extremely negatively formulated items curves are decreasing, because there are no persons located on the trait continuum lower than the item location, and for extremely positively formulated items the ICCs are increasing, because there are no persons located on the trait continuum higher than the item location. For items with locations in between these two extremes the single-peaked item response curve will represent the items.

In this study, four unfolding IRT models following these assumptions are investigated; the generalized graded response model (GGUM; Roberts, Donoghue, & Laughlin, 2000), the collapsed generalized partial credit model (CGPCM), the collapsed graded response model (CGRM), and a quadratic logistic regression model (QLOG). The first three models are collapsed versions of well-known polytomous IRT models. The GGUM is the best known and generally used model. The CGPCM and the CGRM are introduced here as simplified versions of the GGUM. The fourth model is introduced here as a straightforward application of logistic regression models.

Although, a lot of research is done on item fit for dichotomous and polytomous dominance IRT models (for an overview, see Glas & Suarez-Falcon, 2003), very little research is done for unfolding models. Lack of item fit emerges when the item responses do not follow the assumptions of the model. Items might not be unidimensional, might not be locally independent, or the shape of the ICCs is not as expected under the model.

Two types of item misfit for unfolding models are studied in this chapter; differential item functioning (DIF) and misfit of the ICC. DIF is a difference in item response behavior between equally proficient members of two or more groups. DIF occurs when external variables that should be irrelevant influence the responses to an item for persons with the same value on the underlying trait. Race, gender, and age may be examples. If DIF is present and the external variable is used as a background variable to form homogeneous subgroups, the observed values of the different groups do not follow the expected values under the model. That is, the shape of the ICC does not describe the item responses for these groups.

Item fit will be investigated using Lagrange Multiplier tests. These tests have formerly been used to detect DIF, violation of local independence, and violation of the shape of the ICCs for dominance models (Glas, 1998, 1999; Glas & Suarez-Falcon, 2003). The LM tests are defined in a marginal maximum likelihood (MML) framework.

In the following, first the four unfolding models will be explained. Second, a general framework for estimation and testing will be discussed.

Third the LM-tests to investigate item fit will be object of discussion, followed by a simulation study and a real data example. The chapter is finished with some conclusions and a discussion.

6.2 Unfolding IRT models

In this study, four unfolding IRT models for dichotomously scored items will be used; the generalized graded unfolding model (GGUM; Roberts, Donoghue, & Laughlin, 2000), and three alternatives: the collapsed generalized partial credit model (CGPCM), the collapsed graded response model (CGRM; Korobko, 2007), and the quadratic logistic regression model (QLOG). The probabilities of endorsement (response equals 1) of an item is given by $P(X_i = 1 \mid \theta)$, in which the stochastic variable X_i denotes the response on item i and θ is the latent trait variable for a person. The test consists of $i = 1, ..., K$ items.

6.2.1 Generalized Graded Unfolding Model

The generalized graded unfolding model (GGUM) is a collapsed version of a 4-category generalized partial credit model (GPCM; Muraki, 1992). The idea behind GGUM is that the two observed responses (agree and disagree) on an item are the result of collapsing four latent response categories following the GPCM. The four latent response categories are ordered as disagree from below, agree from below, agree from above, and disagree from above. Persons can both agree with a statement, and disagree with a statement for one out of two reasons. The two most extreme latent responses, disagree from below and disagree from above, form the observed disagree response, and the two "middle" latent responses, agree from below and agree from above, form the observed agree response. The probability of endorsement of a statement is equal to

$$P(X_i = 1 \mid \theta) = \frac{\exp(f) + \exp(g)}{1 + \exp(f) + \exp(g) + \exp(h)}, \tag{6.1}$$

in which

$$\begin{aligned} f &= \alpha_i((\theta - \beta_i) - \tau_i), \\ g &= \alpha_i(2(\theta - \beta_i) - \tau_i), \\ h &= \alpha_i(3(\theta - \beta_i)). \end{aligned}$$

The parameters α_i ($\alpha_i > 0$) and τ_i ($\tau_i < 0$) are parameters for the discrimination and the subjective response category threshold of the item,

and together describe the shape of the ICC. The location of the ICC is defined by the parameter β_i. This parameterization is according to the original GGUM model of Roberts, Donoghue, and Laughlin (2000).

6.2.2 Collapsed Generalized Partial Credit Model

The collapsed generalized partial credit model (CGPCM) is a simplification of GGUM. Also, it is a straightforward generalization of the collapsed partial credit model by Verhelst and Verstralen (1993). The idea of not endorsing a statement under the CGPCM is similar to the idea behind GGUM, however, the endorsement of a statement is modeled differently. The CGPCM describes the observed dichotomous response as a collapsed version of a 3-category GPCM with latent response options disagree from below, agree, and disagree from above. The idea behind this model is that persons can only agree for one reason; because their person location is close to the item location. The observed disagree response is the sum of the disagree from below response and the disagree from above response. The equation for the probability of endorsement of a statement is given by

$$P(X_i = 1 \mid \theta) = \frac{\exp(k)}{1 + \exp(k) + \exp(l)}, \tag{6.2}$$

in which

$$\begin{aligned} k &= \alpha_i((\theta - \beta_i) - \tau_i), \\ l &= \alpha_i(2(\theta - \beta_i)). \end{aligned}$$

The parameters have the same interpretation as under GGUM, so β_i defines the location of the ICC, and $\alpha_i(\alpha_i > 0)$ and $\tau_i(\tau_i < 0)$ define the shape of the ICC.

6.2.3 Collapsed Graded Response Model

The probability of endorsement of a statement for the collapsed graded response model (CGRM; Korobko, 2007) can be expressed by

$$P(X_i = 1 \mid \theta) = \pi_{i1}(m) - \pi_{i2}(n), \tag{6.3}$$

in which

$$\begin{aligned} m &= \alpha_i\theta - \beta_{i1}, \\ n &= \alpha_i\theta - \beta_{i2}, \end{aligned}$$

and

$$\pi(y) = \frac{\exp(y)}{1 + \exp(y)}.$$

The CGRM is a collapsed version of a latent 3-category GRM (Samejima, 1969), based on the same idea as the idea behind the CGPCM. Persons endorse an item because the item is located close to their ideal point, whereas they do not endorse an item because the item is located too far away from their ideal point. This may be for two reasons, they are located too far below the item location; disagree from below, or their location is too far above the item location, disagree from above. Parameter α_i ($\alpha_i > 0$) describes the discrimination of the item, and β_{i1} is equal to the item location β_i minus the items threshold τ_i, while β_{i2} is defined as the item location β_i plus the threshold τ_i. So, the two parameters β_{i1} and $\beta_{i2}$are ordered ($\beta_{i1} < \beta_{i2}$) and influence both item location and shape of the ICC.

6.2.4 Quadratic Logistic Regression Model

The quadratic logistic regression model (QLOG) is a different model than the models described above. QLOG is not a collapsed version of a polytomous dominance IRT model, but is based on a logistic regression equation. When it is expected that only persons whose person location is close to the item location will endorse an item, the curves are single-peaked. To get a curved function the logit of the regression model will have to follow that shape. This is the case when the logit is equal to a quadratic function. Then the probability of endorsement is equal to

$$P(X_i = 1 \mid \theta) = \frac{\exp(o)}{1 + \exp(o)}, \tag{6.4}$$

in which

$$o = \alpha_i\theta + \beta_i + \gamma_i\theta^2.$$

In this equation the β_i and $\gamma_i(\gamma_i < 0)$ parameters describe the shape of the response curve, whereas the α_i parameter defines the location.

6.3 A general framework for estimation and testing

Estimation of the parameters of the models will be done in the marginal maximum likelihood (MML) framework (Bock & Aitkin, 1981) and testing of the models will be done in the framework of the LM test.

6.3.1 Estimation of parameters

Let η represent all item parameters under the specified model. Maximizing the joint likelihood of both item and person parameters does not lead to consistent estimates. However, a solution is given by using the assumption that the person parameters are random variables from a probability distribution with density $g(\theta_n)$. In this chapter a standard normal distribution is assumed. Under this assumption MML estimation can be used to estimate the item parameters. The marginal maximum likelihood is given by

$$L(\eta) = \prod_{n=1}^{N} \int P(x_n \mid \theta_n, \eta) g(\theta_n) d\theta_n \tag{6.5}$$

in which $P(x_n \mid \theta_n, \eta)$ denotes the probability of response pattern x_n. The log of this marginal likelihood is maximized with respect to the item parameters. Details are given in the appendix.

The item parameters are estimated using an iterative procedure based on the Expectation-Maximization (EM) algorithm (Dempster, Laird & Rubin, 1977). Gauss-Hermite Quadrature with 40 quadrature points was used for the evaluation of the integrals.

6.3.2 Testing of models

The tests to detect item fit are based on the Lagrange Multiplier statistic (LM-statistic; Aitchison & Silvey, 1958). Two models are specified, and the difference in model fit between the two models is investigated. One model is a restricted model, or null model, which in this chapter is one of the IRT models described above. The other model is the alternative model. The alternative model is analogous to the restricted model, but has one or more additional parameters η_2. So the parameters can be divided into two groups, the parameters η_1 of the restricted model, and the additional parameters η_2. In the restricted model, the η_2 parameters can be seen as parameters which are fixed to zero.

The LM-statistic is evaluated with MML estimates of the parameters η_1 of the restricted model. The LM-statistic weights the first-order derivatives of the loglikelihood function $h(\eta_2)$ (the slope of the loglikelihood function evaluated at the fixed parameter values η_2 using the estimated parameter values η_1), with its covariance matrix Σ. The LM-statistic is given by

$$LM = h(\eta_2)' \Sigma^{-1} h(\eta_2), \tag{6.6}$$

with

$$\Sigma = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12},$$

and

$$\Sigma_{pq} = -\frac{\partial^2 \log L(\eta)}{\partial \eta_p \delta \eta_q'},$$

for $p, q = 1, 2$. The first-order derivatives of the parameters η_1 are usually available when the parameters are estimated by MML. They are given in the appendix for all four models. The second-order derivatives of the loglikelihood function, $-\partial^2 \log L(\eta)/\partial\eta_p \partial\eta_q$, can be computed as

$$\Sigma_{pq} \approx \sum_n E\left[\frac{\partial}{\partial \eta} \log p(x_n \mid \theta, \eta) | x_n, \eta\right] E\left[\frac{\partial}{\partial \eta} \log p(x_n \mid \theta, \eta) | x_n, \eta\right]^t. \quad (6.7)$$

$\sum$ can be viewed as an approximation of the observed information matrix (Mislevy, 1986). And so, matrix Σ produces the asymptotic variance-covariance matrix of the estimates of $h(\eta_2)$, based on only the first-order derivatives of the likelihood function. The LM-statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters in η_2.

An LM test for differential item functioning

When items show DIF this implies that, conditional on θ, the response probabilities differ across groups. Reasons for DIF can be that the item locations are ordered in a different way in the two groups, or that items are less discriminating in one group compared to the other groups. Two groups are defined by $s = 1$ (i.e. men) and $s = 2$ (i.e. women). The question is whether the responses on a certain item can be modeled in both groups by the same unfolding IRT model (null hypothesis) or that two different models are necessary (alternative hypothesis). The alternative model entails a shift in one or more parameters. All three item parameters might be influenced by DIF. The differences in item parameters between groups can be modelled by shifts δ_1, δ_2, and δ_3 for item location, item discrimination, and item threshold respectively. It is, of course, also possible to develop an LM-test targeted at the individual parameters δ_1, δ_2 or δ_3, but this is not considered here; we develop an omnibus test for the three parameters simultaneously. The alternative model uses a dummy variable y_n to describe group membership. If a person belongs to group $s = 1$ the dummy variable takes the value $y_n = 0$. If a person belongs to

group $s = 2$ the dummy variable has value $y_n = 1$. The alternative model for the GGUM can then be expressed by Formula 6.1, with

$$f = (\alpha_i + y_n\delta_2)\left[(\theta - (\beta_i + y_n\delta_1)) - (\tau_i + y_n\delta_3)\right], \tag{6.8}$$
$$g = (\alpha_i + y_n\delta_2)\left[2(\theta - (\beta_i + y_n\delta_1)) - (\tau_i + y_n\delta_3)\right], \tag{6.9}$$
$$h = (\alpha_i + y_n\delta_2)\left[3(\theta - (\beta_i + y_n\delta_1))\right]. \tag{6.10}$$

For the CGPCM the alternative model is expressed by Formula 6.2 with

$$k = (\alpha_i + y_n\delta_2)\left[(\theta - (\beta_i + y_n\delta_1)) - (\tau + y_n\delta_3)\right], \tag{6.11}$$
$$l = (\alpha_i + y_n\delta_2)\left[2(\theta - (\beta_i + y_n\delta_1))\right]. \tag{6.12}$$

Shifts in the CGRM are modeled by Formula 6.3 with

$$m = (\alpha_i + y_n\delta_2)\left[\theta - (\beta_{i1} + y_n\delta_1)\right], \tag{6.13}$$
$$n = (\alpha_i + y_n\delta_2)\left[\theta - (\beta_{i2} + y_n\delta_3)\right]. \tag{6.14}$$

Lastly for the QLOG model, the alternative model in which a shift takes place is modeled by Formula 6.4 with

$$o = (\alpha_i + y_n\delta_1)\theta + (\beta_i + y_n\delta_3) + (\gamma_i + y_n\delta_2)\theta^2. \tag{6.15}$$

The first-order derivatives of the loglikelihood for all four models, which have to be used in the calculations for the LM-test are given in the appendix. The LM-test checks whether there is DIF or not. The null model and the alternative models are compared. In case there is no DIF the null hypothesis $H_0 : \eta_2 = \delta_1 = \delta_2 = \delta_3 = 0$ holds and the expected item scores are the same in both subgroups. In this case the data in both groups can be described by the models as described in equations 6.1 – 6.4. If there are differences in expected item scores between both groups the alternative hypothesis may be $H_a : \eta_2 = \delta_1 \neq 0$ or $H_a : \eta_2 = \delta_2 \neq 0$ or $H_a : \eta_2 = \delta_3 \neq 0$ or a combination of two or three of the alternative models. The statistic to investigate DIF has an asymptotic χ^2distribution with 3 degrees of freedom.

An LM test for shape of item characteristic curve

In general, the development of the LM test for the shape of the ICC is along the same lines as the development for the LM test for DIF. However, there is a difference. The alternative DIF model in itself is a reasonable alternative for the restricted null model, that is, the original IRT model.

The alternative model is analogous to the null model with group-specific parameters for some of the items. The alternative model in the test for the ICCs is in itself not very realistic; it will become clear that it essentially plays the role of a diagnostic tool to assess the appropriateness of the assumed ICC. Similar as in the DIF test, the test for shape of the item characteristic curve partitions the population sample into a number of subgroups. The sample is partitioned based on persons' score levels on the test. The test measures whether the observed item responses for each subgroup are conform the predicted responses based on the model. It might be that an item is expected to be single-peaked, but persons on the higher end of the trait continuum do endorse this item because they are located higher on the trait than the location of the item, resulting in answers to statements according to a monotone increasing model.

The idea behind this test is that all items are ordered based on their location on the latent trait continuum. There is one item of interest, the target item labeled i, while the other items are labeled $j = 1, 2, ..., i - 1, i + 1, ..., K$, and their response pattern is labeled $x_n^{(i)}$. However, in the present case, we do not condition on the number-correct score as is done for dominance IRT models (Glas, 1999), but on the mean restscore on all items except the target item. This mean item restscore $m(x^{(i)})$ is calculated for each item, and based on this score the latent ability continuum is divided into a number of segments. If we order all items on the trait continuum and give them a rank number based on their position, the sum of all endorsed item indices can be computed and be used as the sum item restscore index. The mean item restscore is equal to the sum of the endorsed item indices divided by the total number of endorsed rest items,

$$m(x_n^{(i)}) = \frac{\sum_{j \neq i} j x_{nj}}{\sum_{j \neq i} x_{nj}}.$$

The range of mean item restscores is used to select persons for S_i disjoint subgroups, in which i indicates the partition for the specific target item. For each item, S_i subgroups are formed. In the present chapter, we use the same number of subgroups per item so the index i is dropped. Further, we use three subgroups: one with low mean item restscores, one with medium mean item restscores and one for high mean item restscores. The variable $w(s, x_n^{(i)})$ is used as an indicator function to express if the mean item restscore of the item response pattern $x_n^{(i)}$ is in score range s

$(s = 1, ..., S; S = 3)$ or not, that is,

$$w(s, x_n^{(i)}) = \begin{cases} 1 & \text{if } m_{s-1} \leq m(x_n^{(i)}) < m_s \\ 0 & \text{otherwise.} \end{cases}$$

To evaluate the form of the item characteristic curve the observed probability of a correct response is compared to the expected probability under the null model. Under the alternative model it is expected that the null model does not hold for one or more groups. Groups may change in difficulty parameter, discrimination parameter, or threshold parameter. Therefore shifts on all three parameters in all S subgroups are introduced in the alternative models: δ_{s1} for location, δ_{s2} for discrimination, and δ_{s3} for threshold $(s = 2, ..., S)$. Note that $s = 1$ is excluded to identify the alternative model. That is, subgroup $s = 1$ is used as a baseline. Proceeding analogous to the test for DIF, dummy variables are introduced such that $y_{ns} = 1$ if respondent n belongs to subgroup s and the alternative model for the GGUM can then be expressed by Formula 6.1, with

$$f = \alpha_i((\theta - \beta_i) - \tau_{i1}) + \sum_{s=2}^{S} y_{ns}(\delta_{s2}((\theta - \delta_1) - \delta_{s3})), \tag{6.16}$$

$$g = \alpha_i(2(\theta - \beta_i) - \tau_{i1}) + \sum_{s=2}^{S} y_{ns}(\delta_{s2}(2(\theta - \delta_{s1}) - \delta_{s3})), \tag{6.17}$$

$$h = \alpha_i(3(\theta - \beta_i - y_n\delta_i)) + \sum_{s=2}^{S} y_{ns}(\delta_{s2}(3(\theta - \delta_{s1}))). \tag{6.18}$$

Under the alternative model the probability of a response pattern is given by

$$P(x_n \mid \theta_n, \eta) = P(x_i \mid w(s, x_n^{(i)}), \theta_n, \eta_1, \eta_2) P(x_n^{(i)} \mid \theta_n, \eta_1). \tag{6.19}$$

The definitions for the other three models are analogous. For all four models the first-order derivatives of the loglikelihood, which have to be used in the calculations for the LM-test, are given in the appendix.

Under the null hypothesis it is expected that the observed score responses in each group are equal to the predicted score responses on the target item. The predicted responses under the null model are given in equations 6.1 to 6.4. The δ-values for the item are then assumed to be zero; $H_0 : \delta_{1s} = \delta_{2s} = \delta_{3s} = 0$ $(s = 2, ..., S)$. The alternative hypothesis states that one group, or all groups differ with respect to the location parameter $H_a : \delta_{1s} \neq 0$, the discrimination parameter $H_a : \delta_{2s} \neq 0$, the threshold

parameter $H_a : \delta_{3s} \neq 0$, or a combination of two or all three parameters for one or all groups. The LM-test checks whether the null model or alternative model holds. The statistic has an asymptotic χ^2-distribution with $3(S-1)$ degrees of freedom.

6.4 Simulation studies

Simulation studies were conducted to assess the Type I error rate and the power of the LM-tests for DIF and the shape of the ICC under the four models. In the studies, the test length, K, was varied from 10, to 20, and 30 items. To generate data with comparable response probabilities under the four models , the following procedure was used. Item parameters were generated for different test lengths for the GGUM model, similar to the procedure used by Roberts, Donoghue, and Laughlin (2002). The location parameters, β_i, were equally spread over the continuum and ranged from -2.0 to 2.0. Shape parameters, α_i and τ_i, were both drawn from a uniform distribution as follows. For the α_i-parameter the values were in between 0.5 and 2.0, and for the τ_i-parameter values were between -1.4 and -0.4. Data were generated under GGUM for all three test lengths for $2,000$ persons drawn from the standard normal distribution. Based on these data item parameter values for CGPCM, CGRM and QLOG were estimated. These item parameters and the chosen item parameters for GGUM were used to generate data. The person parameters were drawn from a standard normal distribution.

6.4.1 Type I error rate for LM-test for DIF and shape of ICC

The first simulation study investigated Type I error rate for the LM-statistics under the four models. For each model, data were simulated using item and person parameters as specified above. Then the item parameters were estimated back using MML estimation and the two LM-tests were computed for all items. This process of generating data and person parameters and subsequently estimating parameters and computing statistics was replicated 100 times in each condition. A nominal significance level of 5% was used.

Type I error rate was investigated over all three test lengths for 500 to $4,000$ persons with steps of 500 persons. In Table 6.1, the proportion of significant results aggregated over all K items in the test and over 100 replications are given. The first two columns contain the number of items

and the number of persons. In the next eight columns results on LM-test for differential item functioning (LM1) and for shape of the item characteristic curve (LM2) are given for GGUM (column 3 and 4), CGPCM (column 5 and 6), CGRM (column 7 and 8), and QLOG (column 9 and 10), respectively.

Table 6.1
Type I error rate of the LM-statistics

		GGUM		CGPCM		CGRM		QLOG	
K	N	LM1	LM2	LM1	LM2	LM1	LM2	LM1	LM2
10	500	0.218	0.448	0.172	0.413	0.156	0.334	0.127	0.262
	1000	0.183	0.396	0.095	0.178	0.082	0.170	0.070	0.101
	1500	0.128	0.314	0.075	0.139	0.059	0.078	0.060	0.070
	2000	0.146	0.313	0.067	0.089	0.057	0.099	0.046	0.093
	2500	0.113	0.233	0.054	0.072	0.063	0.063	0.064	0.076
	3000	0.081	0.229	0.039	0.044	0.050	0.069	0.044	0.064
	3500	0.093	0.223	0.046	0.053	0.054	0.068	0.056	0.061
	4000	0.128	0.236	0.067	0.087	0.042	0.052	0.053	0.063
20	500	0.180	0.605	0.225	0.584	0.227	0.656	0.192	0.711
	1000	0.085	0.281	0.096	0.279	0.071	0.201	0.079	0.213
	1500	0.073	0.231	0.069	0.126	0.074	0.133	0.072	0.104
	2000	0.069	0.178	0.049	0.127	0.053	0.084	0.076	0.107
	2500	0.066	0.225	0.049	0.085	0.058	0.087	0.056	0.095
	3000	0.067	0.167	0.059	0.083	0.056	0.087	0.060	0.076
	3500	0.070	0.148	0.051	0.091	0.058	0.077	0.047	0.092
	4000	0.061	0.137	0.045	0.073	0.055	0.072	0.047	0.079
30	500	0.382	0.711	0.435	0.755	0.389	0.839	0.515	0.903
	1000	0.146	0.552	0.191	0.614	0.146	0.644	0.144	0.644
	1500	0.089	0.360	0.105	0.390	0.087	0.417	0.085	0.425
	2000	0.067	0.270	0.066	0.213	0.071	0.181	0.069	0.190
	2500	0.066	0.148	0.059	0.119	0.060	0.087	0.059	0.107
	3000	0.057	0.127	0.056	0.099	0.053	0.074	0.061	0.089
	3500	0.063	0.102	0.058	0.077	0.056	0.076	0.062	0.075
	4000	0.061	0.099	0.053	0.065	0.057	0.061	0.059	0.069

Generally, the Type I error rates decreased with the test length and the sample size. A number of irregularities were present, but it must be taken into account that the standard errors of the percentages are between 0.02 (for a proportion of 0.05) and 0.05 (for a proportion of 0.50). The

results show that the Type I error rates for the LM1-statistic were lower than values for the LM2-statistic. Both statistics attained the nominal significance level, when the number of persons increased, except for the tests for GGUM when the numbers of items were $K = 10$, or 20. Note that the Type I error rates for the LM1 statistic were reasonable and approximately equal for GGUM, CGPCM, CGRM and QLOG when the number of persons is above $1,000$ and test length is above 10. Type I error rate on the LM2-test was approximately equal for CGPCM, CGRM, and QLOG and attained the nominal significance level when number of persons was above $2,000$. The Type I error rate for the LM2 test for the GGUM was still too high for $4,000$ persons and 30 items.

6.4.2 Power of the LM-test for differential item functioning

The second simulation study investigated the power of the LM-test for differential item functioning. The responses of persons were generated according to the alternative models given in equations 6.8 – 6.15. Two types of violations were used; either the location parameter was shifted or the shape parameters were shifted simultaneously. In each study two items were selected as misfitting the model. This implies that the percentage of misfitting items differed with test length. The same items were selected under the four models. Only the first half of the respondents were given a shift δ on the target items. For the second half of the respondents no shift was simulated. To model the same shift under all models, δ-values had to differ between the models. These equivalent δ-values were found in a process of curve fitting. The effect sizes for the shifts in the parameter values are given in Table 6.2. Number of persons was varied from $1,000$, via $2,000$ to $4,000$.

Table 6.2

Parameter-shifts for differential item functioning for the four models

label	GGUM			CGPCM			CGRM			QLOG		
shift δ	α_i	β_i	τ_i	α_i	β_i	τ_i	α_i	β_{i1}	β_{i2}	α_i	β_i	γ_i
L1	-	.25	-	-	.35	-	-	.35	.35	.28	-	-
L2	-	.50	-	-	.70	-	-	.70	.70	.56	-	-
S1	-.40	-	.40	-.55	-	.40	-.50	.35	-.35	-	-.40	.45
S2	-.75	-	.75	-.95	-	.75	-.95	.60	-.60	-	-.75	.70

L1 = shift in location number 1, L2 = shift in location number 2, S1 = shift in shape number 1, S2 = shift in shape number 2

In the first column of Table 6.2, labels are given for the different values

of shifts. Results of the simulation studies are shown in Table 6.3 for shift in location parameter and in Table 6.4 for shift in shape parameters respectively. The tables give the effect size (column 1), the number of items (column 2), the numbers of persons (column 3), and the power results for the misfitting items and the Type I error rate for the other items in the test. Results for GGUM are in column 4 and 5, for CGPCM in column 6 and 7, for CGRM in column 8 and 9, and for QLOG in column 10 and 11. The power was the proportion of times the two misfitting items were correctly detected averaged over the two items and the 100 replications, and the Type I error rate is the mean proportion of incorrect detections of fitting items over all fitting items and all replications.

Table 6.3 shows that for shift in β_i power increased when δ, K, and N increased. Power results were similar over the four models. Type I error rate increased as well when δ increased, but decreased as a result of increase in test length, except for $N = 1,000$. Type I error rate decreased when number of persons increased, for higher test length only. However, Type I error rates were only slightly above the nominal significance level for person numbers above $2,000$ and test length above 10 for the shift with label L2.

The results for shift in shape parameters (Table 6.4) showed that power of the LM-test for DIF increased when δ, K, and N increased. Type I error rate slightly increased when δ increased, however the values were attaining the nominal significance level only when number of items was above 10 and number of persons above $2,000$. When the number of items increased Type I error rate decreased. Type I error rate decreased when number of persons increased for all models when number of items was at least 20. Only for GGUM and QLOG the Type I error rate slightly increased for 10 items when number of persons increased, but also here a standard error between 0.02 and 0.05 has to be taken into account. Results for QLOG were best. This model had the highest power values and Type I error values were reasonable. Results for CGRM were slightly lower on power, and similar on Type I error rate, and results on GGUM and CGPCM were similar to each other for the higher test length, but CGPCM was better for lower test length.

Table 6.3
Power results for differential item functioning with a shift in location

			GGUM		CGPCM		CGRM		QLOG	
δ	K	N	Power	Type I	Power	Type I	Power	Type I	Power	Type I
L1	10	1000	0.130	0.179	0.165	0.102	0.205	0.094	0.115	0.063
		2000	0.205	0.137	0.200	0.065	0.285	0.062	0.250	0.054
		4000	0.350	0.155	0.470	0.067	0.515	0.070	0.425	0.055
	20	1000	0.255	0.106	0.215	0.096	0.260	0.086	0.235	0.079
		2000	0.350	0.075	0.315	0.060	0.370	0.066	0.400	0.064
		4000	0.640	0.076	0.600	0.060	0.680	0.058	0.620	0.064
	30	1000	0.260	0.148	0.315	0.172	0.310	0.167	0.260	0.160
		2000	0.390	0.071	0.355	0.078	0.415	0.066	0.385	0.074
		4000	0.715	0.070	0.700	0.065	0.680	0.061	0.660	0.062
L2	10	1000	0.350	0.160	0.385	0.120	0.475	0.123	0.380	0.069
		2000	0.660	0.183	0.765	0.090	0.800	0.082	0.755	0.071
		4000	0.915	0.176	0.975	0.077	0.960	0.096	0.955	0.096
	20	1000	0.615	0.113	0.580	0.107	0.615	0.073	0.685	0.095
		2000	0.925	0.089	0.905	0.066	0.910	0.066	0.930	0.071
		4000	1.000	0.096	1.000	0.060	1.000	0.070	1.000	0.082
	30	1000	0.685	0.144	0.690	0.196	0.730	0.165	0.680	0.166
		2000	0.935	0.079	0.920	0.063	0.925	0.073	0.905	0.081
		4000	1.000	0.077	1.000	0.058	0.995	0.066	0.995	0.068

The labels L1 and L2 refer to the parameter shifts in Table 6.2

Table 6.4
Power results for differential item functioning with a shift in shape

			GGUM		CGPCM		CGRM		QLOG	
δ	K	N	Power	Type I	Power	Type I	Power	Type I	Power	Type I
S1	10	1000	0.225	0.157	0.240	0.119	0.230	0.093	0.475	0.065
		2000	0.400	0.133	0.420	0.069	0.445	0.086	0.785	0.064
		4000	0.650	0.144	0.775	0.053	0.710	0.067	0.980	0.070
	20	1000	0.245	0.081	0.290	0.092	0.375	0.079	0.515	0.084
		2000	0.460	0.059	0.505	0.069	0.650	0.053	0.830	0.069
		4000	0.780	0.065	0.815	0.053	0.950	0.056	0.980	0.056
	30	1000	0.335	0.142	0.370	0.199	0.390	0.177	0.755	0.188
		2000	0.505	0.067	0.615	0.072	0.545	0.076	0.940	0.074
		4000	0.830	0.061	0.905	0.052	0.875	0.064	1.000	0.068
S2	10	1000	0.645	0.199	0.765	0.097	0.550	0.085	0.930	0.078
		2000	0.920	0.168	0.960	0.083	0.935	0.068	1.000	0.081
		4000	1.000	0.176	1.000	0.081	1.000	0.092	1.000	0.100
	20	1000	0.825	0.099	0.835	0.108	0.940	0.091	0.960	0.093
		2000	0.965	0.067	0.985	0.077	1.000	0.078	1.000	0.071
		4000	1.000	0.064	1.000	0.057	1.000	0.054	1.000	0.067
	30	1000	0.855	0.168	0.860	0.190	0.855	0.178	0.990	0.221
		2000	0.980	0.098	0.975	0.084	0.995	0.073	1.000	0.082
		4000	1.000	0.057	1.000	0.054	1.000	0.070	1.000	0.073

The labels S1 and S2 refer to the parameter shifts in Table 6.2

6.4.3 Power of LM-test for shape of item characteristic curve

In the third simulation study the sensitivity of the LM-test for shape of the item characteristic curve was investigated. The LM-test was computed with the model given by equations 6.16 – 6.18 as alternative for GGUM and the analogous alternative models for the other three models. The boundary values m_s were chosen in such a way that the sample of respondents was partitioned into three equal subsamples. However, for the generation of data under the alternative model, it was not considered realistic to base the

partitioning of the sample on observed scores. Therefore, the partitioning was based on the standard normal distribution of latent person parameters, again in such a way that the sample of respondents was partitioned into three equal subsamples. The shifts of the item parameters were either in the discrimination parameter or both shape parameters. The values are given in Table 6.5. Shifts were zero for the lowest scoring group, δ for the middle group and 2δ for the highest scoring group. In each study, two items were selected as misfitting under the models. Test length and number of persons were varied.

Column 1 of Table 6.5 gives labels for the different shifts in δ. Results of the power studies for shift in discrimination parameter are given in Table 6.6. The table has the same setup as the tables for the LM-test for differential item functioning. Both power and Type I error rate increased when δ increased. The table shows that for a test length of 10 items both power and Type I error rate were approximately equal, and relatively high. Similar values in power and Type I error rate were also found for $N = 1,000$. For the combinations of 20 and 30 items, and 2,000 and 4,000 persons, and the effect size labeled $D1$, the power increased when test length increased. Power increased when number of persons increased. Also, for the combinations of 20 and 30 items, and 2,000 and 4,000 persons, the Type I error rate decreased when number of persons increased. Type I error rates only attained the nominal significance level for 30 items and 4,000 persons. For the effect size labeled $D2$,the results were generally analogous.

Table 6.5

Parameter-shifts for shape of response function for the four models

Lable	GGUM			CGPCM			CGRM			QLOG		
shift δ	α_i	β_i	τ_i	α_i	β_i	τ_i	α_i	β_{i1}	β_{i2}	α_i	β_i	γ_i
D1	.50	-	-	.50	-	-	.50	-	-	-	-	-.50
D2	.75	-	-	.95	-	-	.95	-	-	-	-	-.70
S1	-.25	-	.25	-.35	-	.25	-.30	.20	-.20	-	-.25	.25
S2	-.40	-	.40	-.55	-	.40	-.50	.35	-.35	-	-.40	.45

D1 = shift in discrimination number 1, D2 = shift in discrimination number 2, S1 = shift in shape number 1, S2 = shift in shape number 2

Table 6.6
Power results for shape of item characteristic curve with shift in discrimination

			GGUM		CGPCM		CGRM		QLOG	
δ	K	N	Power	Type I	Power	Type I	Power	Type I	Power	Type I
D1	10	1000	0.433	0.374	0.153	0.187	0.135	0.112	0.060	0.099
		2000	0.551	0.428	0.140	0.073	0.120	0.065	0.115	0.054
		4000	0.486	0.484	0.067	0.078	0.047	0.074	0.065	0.088
	20	1000	0.429	0.283	0.298	0.222	0.305	0.203	0.255	0.209
		2000	0.480	0.188	0.225	0.109	0.240	0.095	0.170	0.103
		4000	0.760	0.134	0.390	0.071	0.415	0.075	0.325	0.088
	30	1000	0.633	0.551	0.596	0.565	0.670	0.634	0.650	0.604
		2000	0.430	0.236	0.330	0.220	0.255	0.142	0.290	0.179
		4000	0.665	0.091	0.390	0.061	0.435	0.077	0.435	0.061
D2	10	1000	0.580	0.490	0.384	0.392	0.310	0.280	0.105	0.095
		2000	0.648	0.563	0.392	0.340	0.345	0.176	0.120	0.070
		4000	0.663	0.646	0.288	0.305	0.128	0.170	0.035	0.120
	20	1000	0.551	0.251	0.442	0.246	0.495	0.244	0.330	0.196
		2000	0.695	0.194	0.665	0.304	0.515	0.118	0.270	0.097
		4000	0.965	0.172	0.825	0.288	0.885	0.092	0.630	0.085
	30	1000	0.753	0.618	0.702	0.596	0.780	0.638	0.730	0.598
		2000	0.655	0.222	0.555	0.178	0.705	0.152	0.445	0.163
		4000	0.910	0.092	0.820	0.074	0.915	0.066	0.595	0.070

The labels D1 and D2 refer to the parameter shifts in Table 6.5

Power and Type I error rate were higher for GGUM than for the other models for the effect size labeled $D1$, while results for the other models were similar. However results on GGUM, CGPCM and CGRM were similar for the effect size labeled $D2$, while results on QLOG were lower.

The results for shift in both shape parameters as shown in Table 6.7 were generally similar to the results in only shift in discrimination. For studies with a test length of 10 items or a number of persons of $1,000$, the Type I error rate and power were approximately equal. A small difference was seen with the former study in that for a combination of 10 items and $4,000$ persons the discrepancy between power and Type I error rate

increased for GGUM, CGPCM and CGRM, while for QLOG the difference was also valuable for 1,000 and 2,000 persons. For $K = 20$ and $K = 30$ the power increased when the test length increased, when the number of persons increased, and when δ increased under all models. The power of the LM2-test was reasonable and Type I error rate attained the nominal significance level for a test length of 20 items and 4,000 persons when a big shift was made, whereas this was the case at 30 items and 4,000 persons when a small shift was made. Results for GGUM, CGPCM and CGRM were similar, while results for QLOG were slightly better for a shorter test length, but worse for a longer test.

Table 6.7

Power results for shape of item characteristic curve for shift in shape

			GGUM		CGPCM		CGRM		QLOG	
δ	K	N	Power	Type I	Power	Type I	Power	Type I	Power	Type I
S1	10	1000	0.342	0.431	0.250	0.306	0.175	0.186	0.160	0.121
		2000	0.265	0.316	0.185	0.188	0.110	0.120	0.245	0.099
		4000	0.240	0.178	0.183	0.093	0.235	0.059	0.325	0.070
	20	1000	0.308	0.326	0.308	0.328	0.225	0.250	0.275	0.221
		2000	0.175	0.184	0.215	0.133	0.145	0.105	0.270	0.104
		4000	0.255	0.167	0.280	0.079	0.225	0.076	0.485	0.098
	30	1000	0.585	0.555	0.595	0.581	0.665	0.658	0.805	0.676
		2000	0.360	0.273	0.437	0.273	0.365	0.199	0.555	0.211
		4000	0.365	0.101	0.570	0.083	0.450	0.073	0.750	0.065
S2	10	1000	0.390	0.430	0.347	0.394	0.245	0.241	0.360	0.164
		2000	0.349	0.367	0.298	0.250	0.270	0.131	0.655	0.096
		4000	0.445	0.257	0.374	0.165	0.475	0.098	0.910	0.130
	20	1000	0.332	0.297	0.394	0.337	0.325	0.234	0.580	0.256
		2000	0.325	0.187	0.420	0.110	0.360	0.097	0.810	0.099
		4000	0.475	0.104	0.660	0.108	0.690	0.080	0.940	0.098
	30	1000	0.709	0.552	0.682	0.559	0.835	0.691	1.000	0.923
		2000	0.595	0.296	0.672	0.290	0.670	0.194	1.000	0.705
		4000	0.835	0.090	0.930	0.083	0.945	0.069	1.000	0.464

The labels S1 and S2 refer to the parameter shifts in Table 6.5

6.5 A real data example

The simulation results on the LM-test for shape of the item characteristic curve showed that this test only had a reasonable Type I error rate for more than 20 items and more than $2,000$ persons. Because the model violations generated above might be artificial, the present study investigated the characteristics of the tests on a real data set. Data of 223 persons on a 20-item inventory measuring Censorship (Roberts, 1995) were used for this study. Item and person parameters were estimated using the original GGUM program (Roberts, Donoghue, & Laughlin, 2000). The responses and estimated person parameters were used as input to estimate the item parameters back for GGUM and to estimate the item parameters for the CGPCM, CGRM and QLOG.

One of the problems with real data is that with large sample sizes, all tests become significant. On the other hand the asymptotic distribution is often not attained with low sample sizes. The present study investigates these opposite effects in real data. First, five mixtures of the real data and simulated data were generated for 223 persons for each model. The first dataset contained 100% simulated data based on the item parameters estimated on the real data and person parameters drawn from a standard normal distribution. The next four datasets were formed by replacing parts of the dataset by real data in steps of 25%. So the second dataset contained 75% simulated data and 25% real data, the third dataset 50% simulated data and 50% real data, the fourth dataset 25% simulated data and 75% real data, and the fifth dataset contained the real data only. Next, the datasets were multiplied four and eight times, to obtain data sets of 892 and $1,784$ persons, respectively. For each dataset item parameters were estimated under the four models and LM-statistics were computed.

In Table 6.8 the results of the study are given. The table shows the percentage of significant LM-tests for the 20-item tests. The quality of the test can be assessed by comparing the relationship between the percentage of significant item tests for completely simulated and completely real data. Note that the test attains the nominal significance level with 100% simulated data for $N = 1,784$, while the power is around 0.90. For $N = 892$ the results are only slightly less favorable. Note that for $N = 223$, for the CGPCM and CGRM, the Type I error rate increases to around 0.15 while the power markedly decreases. For the combination of GGUM and $N = 223$ the test completely fails: both the power and Type I error rate were around 0.60. The results for QLOG were somewhere in between.

Table 6.8

Percentages of significant item tests for various mixtures of a real data set with simulated data

Model	Percentage simulated	Percentage significant $N = 223$	$N = 892$	$N = 1784$
GGUM	100	61	11	5
	75	66	16	19
	50	62	47	65
	25	59	59	84
	0	62	75	89
CGPCM	100	15	9	8
	75	18	13	20
	50	53	44	60
	25	63	59	81
	0	30	88	92
CGRM	100	15	11	5
	75	40	30	25
	50	62	70	65
	25	75	85	90
	0	70	100	100
QLOG	100	33	5	9
	75	70	21	15
	50	80	70	55
	25	80	80	90
	0	55	85	95

For mixtures of real and simulated data, the effects of too low a sample size for an asymptotic distribution and a high power with a large sample size were found; for instance, for 50% simulated data in combination with GGUM, percentage significant decreased from 62 ($N = 223$) to 47 ($N = 892$) and then increased to 65 ($N = 1,784$). Analyses such as these can help to gain insight in the effect of sample size on the outcomes of the tests with a real data set and for the appraisal of the outcome of a test.

6.6 Discussion

Two Lagrange Multiplier statistics were developed to detect item misfit due to differential item functioning (LM1) and misfit of the shape of the item characteristic curve (LM2). When no misfitting items were present, the tests attained the nominal significance level when number of persons was at least $1,500$ and number of items was at least 20. The results of the simulation studies of the power showed that the LM-test for differential item functioning gave better results than the LM-test for shape of the item characteristic curve. Results on the LM-test for differential item functioning showed that this statistic gave reasonable results for all models when number of items was at least 20 and number of persons was at least $2,000$. Power increased with increasing shift in δ, increasing person numbers N, and increasing test lengths K. Type I error rate increased with increasing δ, and decreased with increasing K and N. Although no differences between models were found for shift in the location parameter, for shifts in shape the QLOG model seemed to work slightly better, followed by the CGRM, the CGPCM and the GGUM model.

The power study on the LM-test for the shape of the item characteristic curve showed reasonably high power values; in general, power increased with increasing values of δ, K, and N. However, the Type I error rate was as high as the power for a sample size of $1,000$. It attained values below 0.10 only for the combination of 30 items and $4,000$ persons for the CGPCM, CGRM, and QLOG. , However, when shape parameters were varied, the test gave poor results for the QLOG model, in the sense that the power and Type I error rate were of the same magnitude.

Because the model violations imposed might be somewhat artificial, a study with real data was conducted. The data set did not fit any of the four models. The research question was whether this was due to a large sample size resulting in high power or a low sample size resulting in a poor approximation of the asymptotic distribution. This study found a Type I error rate attaining the nominal significance level for 100% simulated data for a sample size of $N = 1,784$, which rose to 0.10 for $N = 892$, and to 0.15 for $N = 223$. For GGUM the test completely fails for $N = 223$.

Overall, the results show that large numbers of persons (more than $2,000$) and large numbers of items (more than 20) are needed to test the models. The test for differential item functioning performed better than the test for shape of the ICC.

To improve results, it might be possible to extend the LM-statistics to

forms that take into account additional information of other items which show similar wording or external variables as was done for person fit in Glas and Dagohoy (2007). A second extension of this research might be to use LM-tests to detect other forms of item misfit, such as local dependence or multidimensionality. Furthermore, a comparison study between the LM-statistics and other item fit statistics is needed. A fourth extension to this study might be to investigate power and Type I error rates of these item fit statistics for person parameters drawn from a different distribution than the standard normal distribution. In typical performance measurement person scores may not follow a standard normal distribution. The difference of person score distributions on item parameter estimation has been investigated by Roberts, Donoghue, and Laughlin (2002), who found only minor differences between different population distributions for GGUM.

The last remark pertains to the estimation of the models. In the process of searching for realistic item parameter values, we found that MML estimation of the parameters is not always unproblematic. The item parameters are highly correlated and therefore often poorly identified. An often used solution is found in a Bayesian framework where priors for the item parameters can be defined to limit their absolute values. Glas (1999) shows that LM tests for IRT models can be straightforwardly generalized to a Bayes modal framework (the term "modal" refers to the mode of the posterior distribution). Another possibility is using a fully Bayesian framework in combination with Markov chain Monte Carlo (MCMC) computational methods such as applied to IRT by Albert (1992) and Patz and Junker (1999a, 1999b). Also in this respect more research needs to be done.

Appendix

Define the loglikelihood of an item response as

$$L(x_{ni}, \theta_n) = x_{ni} \log P(x_{ni} = 1 \mid \theta_n) + (1 - x_{ni}) \log(P(x_{ni} = 0 \mid \theta_n)). \quad (6.20)$$

Then the log-likelihood for a response pattern given θ_n is

$$L(\eta, \theta_n \mid x_n) = \sum_{i=1}^{K} L(x_{ni}, \theta_n). \quad (6.21)$$

We apply Fisher's identity (Efron, 1977; Louis, 1982) to find the first-order derivatives easily. Fisher's identity is based on a complete data design, with

the complete data consisting of both observed variables x_n and unobserved variables θ_n. Let

$$\omega_n(\eta_{ij}) = \frac{\partial}{\partial \eta_{ij}} L(x_{ni}, \theta_n),$$

for some item parameter ij. Fisher's identity entails that the first-order derivative of the log of the likelihood given by (6.5) is given by

$$\frac{\partial}{\partial \eta_{ij}} \log L(\eta) = \sum_{n=1}^{N} E(\omega_n(\eta_{ij}) \mid x_n, \eta_i)$$

$$= \sum_{n=1}^{N} \int \omega_n(\eta_{ij}) P(\theta_n \mid x_n) d\theta_n,$$

in which $P(\theta_n \mid x_n)$ is the posterior density of θ given the response pattern. The posterior density is equal to $P(x \mid \theta)g(\theta)/p(x_n)$.

The first derivative of the loglikelihood with respect to the item parameters is used in the calculations and is given by

$$\omega_n(\eta_{ij}) = \frac{P_i^{'}(x_{ni} - P_i)}{P_i(1 - P_{i)}} \tag{6.22}$$

in which P_i =$P(x_{ni} = 1 \mid \theta_n)$ is one of the alternative models for GGUM, CGPCM, CGRM or QLOG.

Let the first derivatives of the log-likelihood with respect to the item parameters be defined as follows

$$\omega_n(\alpha_i) = \frac{\partial logL}{\partial \alpha_i} \tag{6.23}$$

$$\omega_n(\beta_i) = \frac{\partial L}{\partial \beta_i} \tag{6.24}$$

$$\omega_n(\tau_i) = \frac{\partial L}{\partial \tau_i} \tag{6.25}$$

$$\omega_n(\gamma_i) = \frac{\partial L}{\partial \gamma_i}. \tag{6.26}$$

For DIF, the first derivative of the log-likelihood with respect to parameters δ under the models for the differential item functioning test are then given in Table 6.9. The explanation of the table is as follows. Consider the GGUM model. The parameter δ_1 is a shift of the parameter β_i and from

Table 6.9
First derivatives of the log-likelihood for δ

	GGUM	CGPCM	CGRM	QLOG
$\omega_n(\delta_1)$	$y_n\omega_n(\beta_i)$	$y_n\omega_n(\beta_i)$	$y_n\omega_n(\beta_{i1})$	$y_n\omega_n(\alpha_i)$
$\omega_n(\delta_2)$	$y_n\omega_n(\alpha_i)$	$y_n\omega_n(\alpha_i)$	$y_n\omega_n(\alpha_i)$	$y_n\omega_n(\gamma_i)$
$\omega_n(\delta_3)$	$y_n\omega_n(\tau_i)$	$y_n\omega_n(\tau_i)$	$y_n\omega_n(\beta_{i2})$	$y_n\omega_n(\beta_i)$

Interpretation: $\omega_n(\delta_1) = y_n\omega_n(\beta_i)$, etc.

Equation (6.8) it can be seen that δ_1 has the same position as β_i, except that it is multiplied by y_n. Therefore, $\omega_n(\delta_1) = y_n\omega_n(\beta)$. The other entries in the table are explained analogously.

For the shape of the ICC, the first derivative of the log-likelihood with respect to parameters δ are analogous if we replace y_n by y_{ns}, for $s = 2, ..., S$. For the four models the ω_n differ. The first derivatives with respect to the item parameters for all four models are given below.

Derivatives for the Generalized Graded Unfolding Model

Using Equation (6.1) it can be easily verified that the first derivative of the loglikelihood with respect to the item parameters are

$$\begin{aligned}
\omega_n(\alpha_i) &= \frac{\theta[f - 2fh + 2g - gh][x_i - P_i]}{(f + g)(1 + h)}, \\
\omega_n(\beta_i) &= \frac{[-f + 2fh - 2g + gh][x - P_i]}{(f + g)(1 + h)}, \\
\omega_n(\tau_i) &= -[x_i - P_i].
\end{aligned}$$

Derivatives for the Collapsed Generalized Partial Credit Model

Using Equations (6.2) it can be easily verified that the first derivative of the loglikelihood with respect to the item parameters are

$$\begin{aligned}
\omega_n(\alpha_i) &= \frac{\theta(k - kl)[x_i - P_i]}{k(1 + l)}, \\
\omega_n(\beta_i) &= \frac{k(l - 1)[x_i - P_i]}{k(1 + l)}, \\
\omega_n(\tau_i) &= -[x_i - P_i].
\end{aligned}$$

Derivatives for the Collapsed Graded Response Model

Using Equations (6.3) it can be easily verified that the first derivative of the loglikelihood with respect to the item parameters are

$$\begin{aligned}
\omega_n(\alpha_i) &= \frac{\left[\theta\pi_{i1}\left(1-\pi_{i1}\right)-\theta\pi_{i2}\left(1-\pi_{i2}\right)\right]\left[x_i-\left(\pi_{i1}-\pi_{i2}\right)\right]}{\left(\pi_{i1}-\pi_{i2}\right)\left(1-\pi_{i1}+\pi_{i2}\right)} \\
\omega_n(\beta_{i1}) &= \frac{\left[-\pi_{i1}\left(1-\pi_{i1}\right)\right]\left[x_i-\left(\pi_{i1}-\pi_{i2}\right)\right]}{\left(\pi_{i1}-\pi_{i2}\right)\left(1-\pi_{i1}+\pi_{i2}\right)} \\
\omega_n(\beta_{i2}) &= \frac{\left[\pi_{i2}\left(1-\pi_{i2}\right)\right]\left[x_i-\left(\pi_{i1}-\pi_{i2}\right)\right]}{\left(\pi_{i1}-\pi_{i2}\right)\left(1-\pi_{i1}+\pi_{i2}\right)}.
\end{aligned}$$

Derivatives for the Quadratic Logistic Regression Model

Using Equations (6.4) it can be easily verified that the first derivative of the loglikelihood with respect to the item parameters are

$$\begin{aligned}
\omega_n(\alpha_i) &= \theta\left(x_i-P_i\right) \\
\omega_n(\beta_i) &= \left(x_i-P_i\right) \\
\omega_n(\gamma_i) &= \theta^2\left(x_i-P_i\right).
\end{aligned}$$

Chapter 7

Conclusions

In the educational, employment, and clinical context, attitude and personality inventories are used to measure typical performance traits and to predict outcomes. Statistical models are applied to obtain latent trait estimates. The models are often the same statistical models as the models used in maximum performance measurement. However, different models than the ones used in maximum performance measurement might be better applicable to describe typical performance measures. This thesis deals with the modeling of two systematic features in the typical performance domain: the factor structure of typical performance measures and response processes to typical performance measures.

The first part of this thesis discussed the multitude of related factors and facets. In Chapters 2 and 3 complex multidimensional models were used to describe the factor structure of both a personality inventory (Chapter 2) and an attitude inventory (Chapter 3). The dimensionality structure of measurement instruments is often explored using non-hierarchical multidimensional models or second-order models. However, the use of the more advanced bifactor model is increasing (Chen, West, & Sousa, 2006; Reise, Morizot, & Hays, 2007). In Chapters 2 and 3, the applicability of this model was investigated and compared to the applicability of the non-hierarchical multidimensional model and second-order model in Chapter 2, and to the non-hierarchical multidimensional model in Chapter 3. The non-hierarchical multidimensional model describes the item responses by domain-specific factors only, whereas second-order and bifactor models assume a general factor and both a general factor and domain-specific factors, respectively. The dichotomously scored personality inventory for adolescents in Chapter 2 could be described best by the bifactor model,

consisting of a general factor and three domain-specific factors. Results on the second-order model and the non-hierarchical multidimensional model supported these findings. Some items were found to be multidimensional, measuring both general and domain-specific factors, whereas other items did measure one of the factors only. Furthermore, it was found that sum scores and factor scores under different models resulted in a different ordering of persons on the traits, especially in the higher ranges of the trait continuum. The third chapter discussed the dimensionality structure of a polytomously scored attitude inventory, which was best described by the non-hierarchical multidimensional model. However, the bifactor analysis largely corroborated the non-hierarchical multidimensional model results. The general factor mainly consisted of items of one scale, and thus was representative of a strong domain-specific factor.

The second part of this thesis dealt with response processes to typical performance measures. Chapter 4 discussed the applicability of different response models. Recently it was suggested that typical performance measures might be better described by ideal point response processes and unfolding models with single-peaked item characteristic curves (ICCs), than by dominance response processes with monotone increasing and decreasing ICCs. In Chapter 4, it was investigated whether dominance or unfolding IRT models are best suited to describe the response processes to two personality inventories. These inventories were constructed using dominance response processes or ideal-point response processes. Both parametric dominance and unfolding IRT models and non-parametric dominance and unfolding IRT models were used to investigate the response processes. Results on all four models showed that inventories constructed using dominance response processes mainly consisted of items with monotone increasing and decreasing ICCs. However, some ICCs were single-peaked or showed a trend to single-peakedness at the higher or lower end of the trait continuum. The personality scale constructed based on ideal-point response processes consisted of items with monotone increasing, decreasing, and single-peaked ICCs. When the inventory was constructed making use of ideal-point response processes, estimated trait scores of the parametric dominance IRT model and parametric unfolding IRT model did not order persons on traits similarly, especially on the upper extreme of the trait continuum.

The results in Chapter 4 showed that there is evidence for the applicability of unfolding IRT models in the typical performance domain. In the literature only a small number of models for single-peaked

response processes are applied. In Chapters 5 and 6, an already existing unfolding IRT model, the generalized graded unfolding model (GGUM) and three newly developed alternatives, the collapsed generalized partial credit model (CGPCM), the collapsed graded response model (CGRM) and the quadratic logistic regression model (QLOG), were compared by investigating the statistical fit of the models. In Chapter 5, two person fit statistics are discussed and in Chapter 6 two item fit statistics are discussed. The fit statistics are developed based on the Lagrange Multiplier (LM) test.

The two person fit statistics measured constancy of theta during the test and tendency to agree. Type I error rate and power of both tests were investigated using simulation studies that were based on real data set parameters under one model, while person parameters were re-estimated and LM-statistics evaluated under all four models. The LM-statistics were good measures to investigate invalid person response processes based on inconsistent test responding and tendency to agree. The Type I error rate on both tests was reasonable, and power increased when test length and effect size increased. Results for all four models were similar, and there was evidence that the models are comparable.

The two item fit statistics in Chapter 6 investigated differential item functioning (DIF) and shape of the ICC. In simulation studies and in an empirical example, Type I error rate and power of the tests were evaluated for the four models separately. Type I error rate and power characteristics were better for the test for DIF than for the test for shape of the ICC. Type I error rate only attained nominal level when number of persons increased. Power on both tests was reasonable for large numbers of items, large numbers of persons and increasing effect size. Type I error rate on the test for DIF was only slightly above the nominal significance level and power was reasonable for at least 20 items and at least 2000 persons, whereas Type I error rate on the test for the shape of the ICC was reasonable for at least 30 items and 4000 persons. On the test for DIF, QLOG worked better than CGPCM and CGRM, which showed better results than GGUM. The CGPCM and CGRM models gave better results on the test for shape of the ICC. Furthermore, a real data example was used to investigate whether the results found in the simulation studies on the power of the test for shape of the ICC were artificial. Mixtures of both real data and simulated data were used to investigate power and the influences of sample size. Results showed that LM-tests do not reach the asymptotic distribution when sample size is too low, and become significant when sample size is too large. CGRM and CGPCM showed the best results.

In general, in this thesis it is shown that it is important to not simply apply models that are used in maximum performance measurement to typical performance measurement. It is important to investigate the validity of typical performance measures using models that take general and specific factors into account, and to investigate whether responses to typical performance measures follow an ideal-point response process. Regarding to the validity of typical performance measures, the bifactor model fitted the attitude measure slightly worse than the non-hierarchical model. However the bifactor model gave clear results on the dimensionality structure of instruments, dimensionality of items, interpretation of the factors, and scoring of individuals for both attitude and personality inventories. The bifactor model is more general, and gives a statistically based conclusion about the appropriateness of the non-hierarchical multidimensional model and the second-order model, even in case of two or three domain-specific factors. Furthermore, the bifactor model gave additional information compared to the other models, which makes it a valuable model to use for constructing and analyzing typical performance measures. Regarding the response processes, ideal-point response processes can be used to broaden the spectrum of measurement and the variety in items. Both monotone increasing (positively worded items), monotone decreasing (negatively worded items) and single-peaked items (neutrally worded items) can be written, and these items may add to the measurement precision in an area where dominance items are difficult to formulate. Four IRT models were introduced to model these types of responses, and two LM-tests to detect person fit and two tests to detect item fit were developed. These are first contributions to person, item and model fit for the unfolding models and thus further research is needed, but unfolding models have potential for measurement in the typical performance domain.

References

Aitchison, J., & Silvey, S. D. (1958). Maximum likelihood estimation of parameters subject to restraints. *Annals of Mathematical Statistics, 29*, 813-828.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*, 179-211.

Albert, J. H. (1992). Bayesian estimation of normal ogive item response functions using Gibbs sampling. *Journal of Educational Statistics, 17,* 251–269.

Andrich, D. (1996). Hyperbolic cosine latent trait models for unfolding direct-responses and pairwise preferences. *Applied Psychological Measurement, 20*, 269-290.

Arbuckle, J.L. (2007), *Amos 16.0 User's Guide*, Chicago, IL: SPSS Inc.

Atkinson, P. (2003). *Assessment 5-14: What do pupils and parents think?* (Spotlight No. 87). Edinburgh, UK: The SCRE Centre, University of Glasgow.

Barelds, D. P. H., & Luteijn, F. (2002). Measuring personality: a comparison of three personality questionnaires in the Netherlands. *Personality and Individual Differences, 33*, 499-510.

Blaikie, F., Schonau, D., & Steers, J. (2004). Preparing for portfolio assessment in art and design: A study of opinions and experiences of exiting secondary school students in Canada, England, and The Netherlands. *The International Journal of Art & Design Education, 23*(3), 302-315.

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of an EM-algorithm. *Psychometrika, 46,* 443-459.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Brookhart, S. M., & Bronowicz, D. L. (2003). 'I don't like writing. It makes my fingers hurt': Students talk about their classroom assessments. *Assessment in Education, 10*(2), 221-242.

Brouwer, D., Meijer, R. R., Weekers, A. M., & Baneke, J. J. (2008). On the dimensionality of the dispositional hope scale. *Psychological Assessment, 20*, 310-315.

Brown, G. T. L. (2004a). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Policy, Principles and Practice, 11*, 305-322.

Brown, G. T. L. (2004b). Measuring attitude with positively packed self-report ratings: Comparison of agreement and frequency scales. *Psychological Reports, 94*, 1015-1024.

Brown, G. T. L. (2006, September). *Secondary school students' conceptions of assessment. A survey of four schools.* Conceptions of Assessment and Feedback Project Report #5. Auckland, NZ: University of Auckland.

Brown, G. T. L. (2008). *Conceptions of Assessment: Understanding What Assessment Means to Teachers and Students.* New York: Nova Science Publishers.

Brown, G. T. L. & Hirschfeld, G. H. F. (2005, December). *Secondary school students' conceptions of assessment.* Conceptions of Assessment and Feedback Project Report #4. Auckland: University of Auckland.

Brown, G. T. L., & Hirschfeld, G. H. F. (2007). Students' conceptions of assessment and mathematics: Self-regulation raises achievement. *Australian Journal of Educational & Developmental Psychology, 7*, 63-74.

Brown, G. T. L., & Hirschfeld, G. H. F. (2008). Students' conceptions of assessment: Links to outcomes. *Assessment in Education: Principles, Policy and Practice, 15*, 3-17.

Brown, G. T. L., Irving, S. E., & Peterson, E. R. (2008). *Beliefs that make a difference: Students' conceptions of assessment and academic performance.* Paper presented at the Biannual Conference of the International Test Commission, Liverpool, UK.

Brown, G. T. L., Irving, S. E., Peterson, E. R., & Hirschfeld, G. H. F. (2009). Use of interactive-informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning and Instruction, 19*(2), 97-111.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *MMPI-2: Manual for administration and scoring.* Minneapolis: University of Minnesota Press.

Chen, F. F., West, S. G., & Sousa, K. H. (2006). A comparison of bifactor and second-order models of quality of life. *Multivariate Behavioral Research, 41*, 189-225.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research, 36*, 523-562.

Chernyshenko, O. S., Stark, S., Drasgow, F., & Roberts, B. W. (2007). Constructing personality scales under the assumptions of an ideal point response process: Toward increasing the flexibility of personality measures. *Psychological Assessment, 19*, 88-106.

Conn, S. & Rieke, M. L. (Eds.). (1994). *The 16PF fifth edition technical manual.* Champaign, IL: Institute for Personality and Ability Testing.

Coombs, C. H. (1964). *A theory of data.* New York: Wiley.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and the NEO Five-Factor Inventory (NEO-FFI): Professional manual.* Odessa, FL: Psychological Assessment Resources.

Cronbach, L. J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper and Row.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B. 39,* 1-38.

DeYoung, C. G. (2006). Higher-order factors of the big five in a multi-informant sample. *Journal of Personality and Social Psychology, 91*, 1138-1151.

Digman, J. M. (1997). Higher order factors of the Big Five. *Journal of Personality and Social Psychology, 73*, 1246-1256.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Efron, B. (1977). Discussion on maximum likelihood from incomplete data via the EM algorithm (by A. Dempster, N. Liard, and D. Rubin). *Journal of the Royal Statistical Society B. 39,* 29.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Erlbaum.

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment: Introduction to the special issue. *Higher Education, 22*, 201-204.

Fan, X. & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modelling, 12*, 343-367.

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model

misspecification and model types. *Multivariate Behavioral Research, 42*(3), 509-529.

Funder, D. C. (1997). *The personality puzzle.* (3rd ed.). New York: Norton & Company.

Glas, C. A. W. (1988). The derivation of some tests of the Rasch model from the multinomial distribution. *Psychometrika, 53*, 525-546.

Glas, C.A.W. (1998). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8, vol. 1,* 647-667.

Glas, C. A. W. (1999). Modification indices for the 2-pl and the nominal response model. *Psychometrika, 64*, 273-294.

Glas, C. A. W., & Dagohoy, A. V. T. (2007). A person fit test for IRT models for polytomous items. *Psychometrika, 72*, 159-180.

Glas, C.A.W., & Suarez-Falcon, J.C. (2003). A Comparison of Item-Fit Statistics for the Three-Parameter Logistic Model. *Applied Psychological Measurement, 27,* 87-106.

Gough, H. G., & Bradley, P. (1996). *CPI manual* (3rd ed.). Palo Alto, CA: Consulting Psychologists Press.

Gustafsson, J. -E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179-203.

Gustafsson, J. -E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*, 407-434.

Harlen, W. (2007). *Assessment of learning.* Los Angeles: Sage.

Hambleton, R. K., Swaminatan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hattie, J. A., Brown, G. T. L., Ward, L., Irving, S. E., & Keegan, P. J. (2006). Formative evaluation of an educational assessment technology innovation: Developers' insights into Assessment Tools for Teaching and Learning (asTTle). *Journal of Multi-Disciplinary Evaluation, 5*, Available online: http://survey.ate.wmich.edu/jmde/index.php/jmde_1/article/view/50/57.

Hendriks, A. A. J, Hofstee, W. K. B., & De Raad, B. (1999). *Handleiding bij de Five-Factor Personality Inventory (FFPI). [The Five-Factor Personality Inventory: Professional Manual].* Lisse: Swets Test Publishers.

Hirschfeld G. H. F., & Brown, G. T. L. (2009). Students' conceptions of assessment: Factorial and structural invariance of the SCoA across sex, age, and ethnicity. *European Journal of Psychological Assessment, 25*(1), 30-38.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Jöreskog, K. G. (2007). Factor analysis and its extensions. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 47-77). Mahwah, NJ: LEA.

Kline, R. B. (2005). *Principles and practice of structural equation modeling.*(2nd ed.). New York: The Guilford Press.

Klockars, A. J., & Yamagishi, M. (1988). The influence of labels and positions in rating scales. *Journal of Educational Measurement, 25*(2), 85-96.

Korobko, O. B. (2007). *Comparison of examination grades using item response theory: A case study.* Unpublished doctoral dissertation, University of Twente, Enschede, The Netherlands.

Lam, T. C. M., & Klockars, A. J. (1982). Anchor point effects on the equivalence of questionnaire items. *Journal of Educational Measurement, 19*(4), 317-322.

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching.* (8th ed.). Upper Saddle River, NJ: Merrill/Prentice Hall.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.

Louis, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226-233.

Luo, G. (1998). A general formulation for unidimensional unfolding and pairwise preference models: Making explicit the latitude of acceptance. *Journal of Mathematical Psychology, 42*, 400-417.

Luo, D., Petrill, S. A., & Thompson, L. A. (1994). An exploration of genetic g: Hierarchical factor analysis of cognitive data from the western reserve twin project. *Intelligence, 18*, 335-347.

Luteijn, F. (1974). *De konstruktie van een persoonlijkheidsvragenlijst (NPV) [*The construction of the Dutch Personality Questionnaire (NPV)]. Amsterdam: Swets & Zeitlinger.

Luteijn, F., van Dijk, H., & Barelds, D. P. H. (2005). *NPV-J: Junior Nederlandse Persoonlijkheidsvragenlijst. Herziene handleiding 2005 [*NPV-J: Dutch Personality Questionnaire-Junior: Professional manual (revised)]. Amsterdam: Harcourt Assessments B.V.

Luteijn, F., Starren, J., & van Dijk, H. (1985). *Herziene handleiding bij NPV.* Lisse: Swets & Zeitlinger.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin, 114*, 185-199.

Marsh, H. W., Hau, K.-T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling, 11*, 320-341.

Meijer, R. R. & Baneke, J. J. (2004). Analyzing psychopathology items: A case for nonparametric item response theory. *Psychological Methods, 9*, 354-368.

Meijer, R. R., & Sijtsma, K. (1995). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education, 8*, 261-272.

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., Eisman, E. J., Kubiszyn, T. W., & Reed, G. M. (2001). Psychological testing and psychological assessment: A review of evidence and issues. *American Psychologist, 56*, 128-165.

Mislevy, R.J. (1986). Bayes modal estimation in item response models. *Psychometrika, 51,* 177-195.

Molenaar, I. W., & Sijtsma, K. (2000). *MSP for Windows [Software manual].* Groningen: iec ProGAMMA.

Moni, K. B., van Kraayenoord, C. E., & Baker, C. D. (2002). Students' perceptions of literacy assessment. *Assessment in Education, 9*(3), 319-342.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Muthén, L. K., & Muthén, B. O. (1998-2006). *MPlus user's guide (4th ed.).* Los Angeles, CA: Muthén & Muthén.

Olsen, L., & Moore, M. (1984). *Voices from the classroom: Students and teachers speak out on the quality of teaching in our schools.* Oakland, CA: Students for Quality Teaching Project / Citizens Policy Center.

Ozer, D. J., & Benet-Martinez, V. (2006). Personality and the prediction of consequential outcomes. *Annual Review of Psychology, 57*, 401-421.

Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research, 62*, 307-332.

Pajares, M. F., & Graham, L. (1998). Formalist thinking and language

arts instruction: Teachers' and students' beliefs about truth and caring in the teaching conversation. *Teaching & Teacher Education, 14*(8), 855-870.

Patrick, J. C., Hicks, B. M., Nichol, P. E., & Krueger, R. F. (2007). A bifactor approach to modeling the structure of the psychopathy checklist-revised. *Journal of Personality Disorders, 21*, 118-141.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342–366.

Peterson, E. R., & Irving, S. E. (2008). Secondary school students' conceptions of assessment and feedback. *Learning and Instruction, 18*(3), 238-250.

Post, W. J. (1992). *Nonparametric unfolding models. A latent structure approach.* Leiden: DSWO Press.

Post, W. J., van Duijn, M. A. J., & van Baarsen B., (2001). Single peaked or monotone tracelines? On the choice of an IRT model for scaling data. In A. Boomsma, M. A. J. van Duijn & T. A. B. Snijders (Eds). *Essays on item response theory.* New York: Springer

Post, W. J., & Snijders, T. A. B. (1993). Nonparameteric unfolding models for dichotomous scaling data, *Methodica, 7*, 130-156.

Rao, C. R. (1947). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44*, 50-57.

Reay, D., & Wiliam, D. (1999). 'I'll be a nothing': Structure, agency and the construction of identity through assessment. *British Educational Research Journal, 25*(3), 343-354.

Reise, S. P., & Flannery, W. P. (1996). Assessing person-fit measurement of typical performance applications. *Applied Measurement in Education, 9*, 9-26.

Reise, S. P., & Henson, S. P. (2003). A discussion of modern and traditional psychometrics as applied to personality assessment scales. *Journal of Personality Assessment, 81*, 93-103.

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research, 16*,19-31.

Reise, S. P., & Waller, N. G. (2003). How many IRT parameters does it take to model psychopathology items? *Psychological Methods, 8*, 164-184.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*, 51-67.

Roberts, J. S. (1995). *Censorship* [Data file]. Available from School of Psychology, Georgia Institute of Technology website, http://www.psychology.gatech.edu/Unfolding/Intro.html

Roberts, J. S. (2001). GGUM2000: Estimation of parameters in the generalized graded unfolding model. *Applied Psychological Measurement, 25*, 38.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2000). A general item response theory model for unfolding unidimensional polytomous responses. *Applied Psychological Measurement, 24*, 3-32.

Roberts, J. S., Donoghue, J. R., & Laughlin, J. E. (2002). Characteristics of MML/EAP parameter estimates in the generalized graded unfolding model. *Applied Psychological Measurement, 26*, 192-207.

Roberts, J. S., Fang, H., Cui, W., & Wang, Y. (2004). *GGUM2004: A Windows-based program to estimate parameters of the generalized graded unfolding model*. Manuscript in preparation.

Roberts, J. S., Laughlin, J. E., & Wedell, D. H. (1999). Validity issues in the likert and thurstone approaches to attitude measurement. *Educational and Psychological Measurement, 59*, 211-233.

Rotter, J. B. (1982). Social learning theory. In N. T. Feather (Ed.), *Expectations and actions: Expectancy-value models in psychology* (pp. 241-260). Hillsdale, NJ: Erlbaum.

Ryan, R. M., Connell, J. P., & Deci, E. L. (1985). A motivational analysis of self-determination and self-regulation in education. In C. Ames & R. Ames (Eds.). *Research on motivation in education, 2*. New York: Academic.

Samejima, F. (1969). Estimation of latent ability using a pattern of graded scores. *Psychometrika, Monograph Supplement, No. 17.*

Sanderman, R., Arrindell, W. A., Ranchor, A. V., Eysenck, H. J., & Eysenck, S. B. G. (1995). *Het meten van persoonlijkheidskenmerken met de Eysenck Personality Questionnaire (EPQ)*. [Measuring personality aspects with the Eysenck Personality Questionnaire (EPQ)]. Groningen: Noordelijk Centrum voor Gezondheidsvraagstukken.

Schunk, D. H., & Zimmerman, B. J. (2006). Competence and control beliefs: Distinguishing the means and ends. In P. A. Alexander & P. H. Winne (Eds.), *Handbook of Educational Psychology* (2nd ed., pp. 349-367). Mahwah, NJ: LEA.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

Smith, T. W., Mohler, P. P., Harkness, J., & Onodera, N. (2005). Methods for assessing and calibrating response scales across countries and languages. *Comparative Sociology, 4*(3-4), 365-415.

Snijders, T. (2001). Asymptotic distribution of person-fit statistics with estimated person parameter. *Psychometrika, 66*, 331-342.

Sörbom, D. (1989). Model modification. *Psychometrika, 54*, 371-384.

Stark, S. (2001). *MODFIT: A computer program for model-data fit.* Unpublished manuscript, University of Illinois at Urban-Champaign.

Stark, S., Chernyshenko, O. S., Drasgow, F., & Williams, B. A. (2006). Examining assumptions about item responding in personality assessment: should ideal point methods be considered for scale development and scoring? *Journal of Applied Psychology, 91*, 25-39.

Steiger, J. H. (2000). Point estimation, hypothesis testing, and interval estimation using the RMSEA: Some comments and a reply to Hayduk and Glaser. *Structural Equation Modeling, 7*(2), 149-162.

Stevens, J. P. (1992). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Erlbaum.

Stralberg, S. (2006). Reflections, journey, and possessions: Metaphors of assessment used by high school students. *Teachers College Record.* Retrieved from http://www.tcrecord.org. doi:12570

Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education, 30*(4), 325-341.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393-408.

Thurstone, L. L. (1928). The measurement of social attitudes. *Journal of Abnormal and Social Psychology, 26*, 249-269.

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology, 43*, 39-55.

Van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory.* New York: Springer.

Van Schuur, W. H., & Post, W. J. (1998). *MUDFOLD. A program for multiple unidimensional unfolding.* Version 4.0 [Software manual]. Groningen: ProGAMMA.

Verhelst, N. D., & Glas, C. A. W. (1995). The generalized one parameter

model: OPLM. In: G. H. Fischer & I. W. Molenaar (Eds.). *Rasch models: their foundations, recent developments and applications.* (pp. 215-238). New York: Springer.

Verhelst, N. D., Glas, C. A. W., & Verstralen, H. H. F. M. (1995). *OPLM: computer program and manual.* Arnhem: Cito, the National Institute for Educational Measurement, the Netherlands.

Verhelst, H. D., & Verstralen, H. H. F. (1993). A stochastic unfolding model derived from partial credit model. *Kwantitatieve Methoden, 42*, 93-108.

Walpole, M., McDonough, P. M., Bauer, C. J., Gibson, C., Kanyi, K., & Toliver, R. (2005). This test is unfair: Urban African American and Latino high school students' perceptions of standardized college admission tests. *Urban Education, 40*(3), 321-349.

Weeden, P., Winter, J., & Broadfoot, P. (2002). *Assessment: What's in it for schools?* London: RoutledgeFalmer.

Weekers, A. M., & Meijer, R. R. (2008). Scaling response processes on personality items using unfolding and dominance models: An illustration with a dutch dominance and unfolding personality inventory. *European Journal of Psychological Assessment, 24*, 65-77.

Weiner, B. (1985). An attributional theory of achievement motivation and emotion. *Psychological review, 92*, 548-573.

Wright, B. D., & Stone, M. H. (1979). *Best Test Design.* Chicago, IL: MESA Press University of Chicago.

Yung, Y. -F., Thissen, D., & McLeod, L. D. (1999). On the relationship between the higher-order factor model and the hierarchical factor model. *Psychometrika, 64*, 113-128.

Zeidner, M. (1992). Key facets of classroom grading: A comparison of teacher and student perspectives. *Contemporary Educational Psychology, 17*, 224-243.

Samenvatting

Attitude- en persoonlijkheidsvragenlijsten worden in de onderwijs- beroeps- en klinische context vaak gebruikt om kenmerkend gedrag ("typical performance") te meten en externe variabelen (uitkomsten) te voorspellen. Door gebruik te maken van statistische modellen is het mogelijk om schattingen te maken van de trek die wordt gemeten en om de structuur van de gemeten gedragingen in kaart te brengen. De gebruikte modellen zijn vooral veel toegepast bij onderwijskundige toetsen en andere prestatietoetsen ("maximum performance"). Omdat persoonlijkheids- en attitudevragenlijsten kwalitatief verschillen van onderwijskundige toetsen is het niet vanzelfsprekend psychometrische modellen die worden toegepast bij onderwijskundige metingen toe te passen bij persoonlijkheids- en attitudemetingen. In dit proefschrift wordt aan het modelleren van twee systematische kenmerken van vragenlijsten voor kenmerkend gedrag aandacht besteed. Ten eerste is de complexiteit van de factorstructuur van de vragenlijsten vaak complexer dan bij prestatietoetsen. Ten tweede zijn er mogelijke verschillen tussen de responsprocessen bij prestatietoetsen en vragenlijsten.

Na een algemene inleiding (hoofdstuk 1) wordt in het eerste deel van dit proefschrift aandacht besteed aan het modelleren van de structuur van persoonlijkheids- en attitudevragenlijsten. Het tweede deel van dit proefschrift besteed aandacht aan het modelleren van responsprocessen op vragenlijsten.

In de twee hoofdstukken (hoofdstuk 2 en 3) in het eerste deel van het proefschrift worden complexe multidimensionele modellen gebruikt om de factorstructuur van een dichotoom gescoorde persoonlijkheidsvragenlijst (hoofdstuk 2) en een polytoom gescoorde attitudevragenlijst (hoofdstuk 3) te beschrijven. Hoewel de dimensionaliteitsstructuur van meetinstrumenten vaak wordt onderzocht met behulp van niet-hierarchische multidimensionele modellen of tweede-orde modellen, neemt het gebruik van bifactor modellen toe. Het niet-hierarchische model beschrijft de items als

metingen van domein-specifieke factoren, terwijl het tweede-orde model en bifactor model de items beschrijven als metingen van respectievelijk een algemene factor en zowel algemene als domein-specifieke factoren. De structuur van de persoonlijkheidsvragenlijst wordt onderzocht door gebruik te maken van alle drie de modellen, terwijl de structuur van de attitudevragenlijst alleen wordt onderzocht voor het niet-hierarchisch multidimensionele model en het bifactor model. De resultaten tonen dat de persoonlijkheidsvragenlijst het best wordt beschreven door het bifactor model, terwijl de attitudevragenlijst het best wordt beschreven door het niet-hierarchisch multidimensioneel model. De resultaten voor de onderzochte modellen laten elkaar ondersteunende en bevestigende resultaten zien. Het bifactor model geeft in beide onderzoeken aanvullende suggesties over de algemeenheid van de constructen en de multidimensionaliteit van de items binnen de gemeten constructen. Daarnaast wordt geconstateerd dat het scoren van personen onder de verschillende modellen leidt tot verschillende ordeningen van personen op trekken, vooral in de hogere regionen van het latente trek continuum.

Het tweede deel, startend met hoofdstuk 4, onderzoekt de toepasbaarheid van ontvouwingsmodellen binnen het kenmerkend gedrag domein (hoofdstuk 4), alsmede statistische procedures om de passing van deze ontvouwingsmodellen te onderzoeken (hoofdstuk 5 en 6). Recentelijk is gesuggereerd dat binnen het persoonlijkheidsdomein responsprocessen op vragenlijsten wellicht beter worden beschreven door ontvouwingsmodellen met belvormige itemkarakteristieke curves dan door de gangbare dominantiemodellen. Ondanks dat ontvouwingsmodellen wel worden gebruikt bij attitudemetingen, worden deze heden ten dage nauwelijks gebruikt om persoonlijkheidsvragenlijsten te construeren en te analyseren. Om meer inzicht te krijgen in de structuur van persoonlijkheidsdata, wordt in hoofdstuk 4 onderzocht of ontvouwings-IRT-modellen (ontvouwings-itemresponsetheorie-modellen) een alternatief kunnen vormen voor dominantie-IRT-modellen. Zowel vragenlijsten die zijn geconstrueerd op basis van dominantieresponsprocessen als ontvouwingsresponsprocessen worden geanalyseerd. Zowel parametrische als niet-parametrische IRT-modellen worden gebruikt voor de analyses. De resultaten laten zien dat vragenlijsten die geconstrueerd zijn door gebruik te maken van dominantieresponsprocessen hoofdzakelijk bestaan uit items met monotoon stijgende itemkarakteristieke curven. Sommige itemkarakteristieke curves zijn echter belvormig of tonen een trend naar belvormige curves aan het hogere of lagere extreem van het

trek continuum. De persoonlijkheidsvragenlijst die geconstrueerd is met behulp van ontvouwingsresponsprocessen bestaat inderdaad uit items met zowel monotoon stijgende, monotoon dalende als belvormige itemkarakteristieke curves. Voor de vragenlijst geconstrueerd op basis van ontvouwingsresponsprocessen zijn de geschatte trekscores van het parametrische dominantie-IRT-model en het parametrische ontvouwings-IRT-model niet in dezelfde volgorde geordend.

De resultaten van hoofdstuk 4 tonen enige evidentie voor de toepasbaarheid van ontvouwings-IRT-modellen binnen het kenmerkend gedrag domein. Het gebruik van ontvouwings-IRT-modellen in het persoonlijkheids- en attitude domein is relatief schaars, en in tegenstelling tot dominantie-IRT-modellen worden slechts enkele modellen voor ontvouwingsresponsprocessen gebruikt. In de hoofdstukken 5 en 6 worden een bestaand ontvouwings-IRT-model, het "generalized graded unfolding model" (GGUM) en drie nieuw ontworpen alternatieven, het "collapsed generalized partial credit model" (CGPCM), het "collapsed graded response model" (CGRM) en het "quadratic logistic regression model"(QLOG) vergeleken door de statistische passing van de modellen te onderzoeken. In hoofdstuk 5 worden passingsmaten voor personen onderzocht en in hoofdstuk 6 passingsmaten voor items. In beide hoofdstukken worden passingstoetsen ontworpen die gebaseerd zijn op de Lagrange Multiplier (LM) toets.

In hoofdstuk 5 worden twee "personfit" toetsen ontwikkeld: een toets om te onderzoeken of de waarde van de latente trek gedurende de test niet veranderd, en een toets om de tendens tot instemming te onderzoeken. Simulatiestudies gebaseerd op de karakteristieken van echte data worden uitgevoerd om de Type I fout en het onderscheidend vermogen van de toetsen te onderzoeken. Data worden gegenereerd onder één model en vervolgens worden de persoonsparameters teruggeschat en de LM-toetsen geëvalueerd voor de vier modellen. Type I fouten van beide LM-toetsen zijn acceptabel voor alle modellen. Het onderscheidend vermogen van de LM-toetsen neemt toe wanneer de testlengte en effectgrootte toeneemt. Tussen de vier modellen worden slechts kleine verschillen gevonden. Beide LM-toetsen blijken geschikt om invalide respons processen van personen door respectievelijk inconsistentie van de latente trek en tendens tot instemming te onderzoeken.

In hoofdstuk 6 worden de assumpties van eendimensionaliteit en vorm van de itemkarakteristieke curve van ontvouwings-IRT-modellen onderzocht. Twee LM-toetsen worden ontwikkeld, een passingstoets om

"differential item functioning" (DIF) te onderzoeken, en een passingstoets om de vorm van de itemkarakteristieke curve te onderzoeken. Beide toetsen worden apart geëvalueerd voor de vier modellen. In simulatiestudies worden de Type I fout en het onderscheidend vermogen van de toetsen onderzocht. De test voor DIF presteert beter dan de test voor vorm van de itemkarakteristieke curve. De Type I fout van beide toetsen is boven nominaal niveau, maar benadert het nominale niveau als het aantal personen toeneemt. Resultaten voor de studie naar onderscheidend vermogen van beide toetsen laten zien dat een groot aantal items en een groot aantal personen nodig zijn om acceptabele resultaten te verkrijgen. Ondanks dat het onderscheidend vermogen van de LM-toets voor DIF toeneemt als de test lengte, het aantal personen en de effectgrootte toenemen, en de Type I fout net boven nominaal niveau blijft steken, worden pas acceptabele resultaten gevonden voor ten minste 20 items en ten minste 2000 personen. Voor de LM-toets naar vorm van de itemkarakteristieke curve worden acceptabele waarden enkel gevonden voor ten minste 30 items en ten minste 4000 personen. Voor de DIF-toets presteert QLOG iets beter dan de andere drie modellen, terwijl GGUM slechter presteert dan de overige modellen. De modellen CGPCM en CGRM geven betere resultaten dan de andere twee modellen voor de toets naar de vorm van de itemkarakteristieke curve. Empirische data worden geanalyseerd om te onderzoeken of de resultaten, die worden gevonden in de simulatiestudie naar het onderscheidend vermogen van de toets voor de vorm van de itemkarakteristieke curve, gemaakt zijn. Gemixte designs van echte en gesimuleerde data tonen dat de toets significante resultaten geeft als de populatiegrootte te groot wordt en dat de LM-toets de asymptotische verdeling niet bereikt als de populatiegrootte te klein is. Resultaten zijn het meest hoopvol voor CGPCM en CGRM.

In het algemeen worden in dit proefschrift twee onderwerpen besproken; de veelheid aan gerelateerde factoren en het modelleren van responsprocessen. De twee hoofdstukken over de dimensionaliteitsstructuur van metingen laten zien dat het belangrijk is de validiteit van vragenlijsten te onderzoeken door gebruik te maken van modellen die zowel algemene als specifieke factoren in ogenschouw nemen. Het tweede deel van het proefschrift laat zien dat het belangrijk is ontvouwingsresponsprocessen in ogenschouw te nemen, omdat deze het meetspectrum kunnen verbreden, alsmede de variëteit in persoonlijkheids- en attitudeitems en de meetprecisie van vragenlijsten. Vier modellen met bijbehorende passingsmaten worden besproken.

Dankwoord

Dit is het dan, mijn proefschrift, het resultaat van vier jaar werk. Ik mag het dan wel *mijn* proefschrift noemen, maar ik had dit boekje niet kunnen schrijven zonder de hulp, medewerking en steun van velen. Promoveren gaat niet altijd over rozen, maar desondanks is het een mooi proces waarin je leert, groeit en je eigen persoonlijkheid en attitude goed leert kennen. Nee, het schrijven van een proefschrift is geen simpel monotoon stijgend proces met alleen maar mooie ontwikkelingen. Het is meer te omschrijven als een multi-peaked monotoon stijgend proces. Naast de groei, die welliswaar door blijft gaan zijn er ook kleine tegenslagen en gaat het niet altijd helemaal zoals je graag zou willen. Maar juist deze tegenslagen heb je nodig om verder te kunnen, immers alleen door fouten te maken doe je ervaring op en leer je. Op deze momenten heb je de hulp, medewerking en steun van anderen het hardst nodig. Natuurlijk kan ik niet iedereen hier persoonlijk bedanken, maar enkelen wil ik toch graag noemen.

In het bijzonder bedank ik mijn promotoren en assistent-promotor. Rob, de eerste schreden van mijn promotietraject zette ik onder jouw begeleiding. Je kennis over de psychologie, het meten van psychologische constructen, en de mogelijkheden voor het toepassen van de psychometrie binnen de psychologie hebben mij zeer geholpen. Altijd was je enthousiast, altijd stond de deur open en was je bereid bevindingen te bespreken, iets uit te leggen of een discussie te voeren. Ik heb hier ontzettend veel van geleerd. Zelfs toen je verhuisde naar Groningen bleef je je met mijn project bezig houden en stond je klaar indien nodig. Mijn dank hiervoor, en natuurlijk ook voor alle leuke gesprekken over sport, verhuizingen en andere dagelijkse bezigheden.

Bernard, toen mijn project een aantal jaar liep, kwam ik binnen jouw interessegebied te werken. Ik ben je dankbaar voor alle gesprekken die we hebben gehad over mijn project. Elke week wilde je dat ik even langsliep om je op de hoogte te houden van wat ik had gedaan, hoe ik verder wilde, en ook al had ik niet echt iets nieuws te vertellen, dan kregen we dat uurtje

ook wel vol door over de dagelijkse gang van zaken te praten. Bedankt!

Cees, hoewel je mijn promotor bent, hebben we in de eerste jaren van mijn promotietraject weinig samengewerkt of overleg gevoerd. Toen Rob vertrok naar Groningen, stond je erop dat wij samen toch echt die paar meer technische projecten zouden gaan doen, die al zo lang op de plank lagen. In het begin was ik daar nogal huiverig voor. Ik vond de mathematische statistiek-bijeenkomsten die we regelmatig hadden leuk, leerzaam en interessant, maar om nog een ommezwaai te maken in het laatste jaar van mijn promotie vond ik toch wel moeilijk. Maar jij zette door en overtuigde mij ervan dat dit echt ten goede kwam voor mijn proefschrift, mijn ontwikkeling en mijn toekomst. Ik kan nu zeggen, het is een zwaar jaar geweest dat laatste, maar ik heb er zoveel van geleerd. Niet alleen leerde je me op een andere manier naar psychometrie en wetenschap te kijken, maar dat stukje wat ik ergens steeds miste doordat ik niet wist wat er achter al die modellen zat, leerde jij mij inzien. Ik leerde in een half jaar programmeren, en we schreven twee artikelen. Heel erg bedankt dat je doorzette om mij ook dit stukje bij te brengen en voor de andere gesprekken die we hadden over levenswijzen, idealen en wat dan ook meer.

Naast mijn directe begeleiders wil ik mijn collega's van OMD bedanken voor de gezellige sfeer, die maakte dat ik met plezier naar mijn werk ging. Bij naam wil ik hier noemen, Rinke en Hanneke, mijn kamergenoten, die er voor zorgden dat ik niet alleen en eenzaam op een kamertje zat. Jullie allebei bedankt voor de gezelligheid op onze werkplek en de leuke professionele en dagelijkse gesprekken.

Marie, Jose en Servan zorgden voor de afwisseling tijdens het werk. Het was fijn af en toe te gaan sporten, wandelen of gewoon gezellig bij te kletsen bij een kopje thee.

Verder bedank ik Iris voor de gezelligheid en daarnaast de hulp bij het verzamelen van de data in het eerste jaar. Voor de dataverzameling ben ik ook dank verschuldigd aan Ilse van den Bosch, Egon Wevers, en René Delnooz, en natuurlijk alle leerlingen en studenten die de tijd namen de vragenlijsten in te vullen.

Gavin bedank ik voor het beschikbaar stellen van zijn gegevens en de samenwerking in het schrijven van het derde hoofdstuk van dit proefschrift.

Cees Aarts, Theo Eggen, Karin Sanders en Klaas Sijtsma, bedank ik voor het zitting nemen in mijn promotiecommissie.

Naast alle wetenschappelijke collega's ben ik natuurlijk ook dank verschuldigd aan mijn vrienden en familie. Ook van hen noem ik slechts enkelen bij naam. Petra wil ik bedanken voor onze vriendschap, voor

de sportieve pauze's en dat ze me als paranimf wil begeleiden bij de verdediging. Suzanne bedank ik voor de begeleiding tijdens mijn promotie als paranimf, voor het ontwerpen van het omslag van mijn proefschrift en haar en Koen bedank ik voor alle leuke avonden, uitstapjes en vakanties, die we met zijn vieren ondernemen.

Dan ben ik aangekomen bij mijn ouders Jan en Els, en mijn broertje Theo. Het is heel fijn te weten dat jullie altijd voor me klaar staan, interesse tonen in wat ik doe, en net zo trots zijn op dit resultaat als ik dat ben.

En als laatste, mijn lieve vriend Björn, voor de hulp bij de lay-out van mijn proefschrift, voor al het geduld dat je hebt moeten opbrengen in met name het laatste half jaar, maar bovenal voor het feit dat je altijd voor me klaar staat, me altijd steunt, en voor alle leuke dingen die we samen doen.

Anke Weekers
november 2009

Stellingen
behorend bij het proefschrift
Modeling typical performance measures
door Anke M. Weekers

1. De stelling "er zijn geen perfecte indicatoren voor kenmerkend gedrag: er zijn alleen maar aanwijzingen, en aanwijzingen zijn altijd dubbelzinnig (Funder, 1997)" maakt de moeilijkheden bij de ontwikkeling en interpretatie van vragenlijsten duidelijk.

2. Wat standaard is in onderzoek naar prestatie- en onderwijskundige toetsen, hoeft dat voor vragenlijsten niet te zijn. Ieder type test of toets heeft zijn eigen doelen en kan dus een andere aanpak vereisen (dit proefschrift).

3. Om de validiteit van vragenlijsten te onderzoeken is het belangrijk gebruik te maken van modellen die zowel algemene als specifieke factoren in ogenschouw nemen (dit proefschrift).

4. Ontvouwingsresponsprocessen kunnen de variëteit in kenmerkend gedrag items vergroten, alsmede het meetspectrum verbreden (dit proefschrift).

5. Een goed model hangt niet alleen af van de passing van het gehele model, de items en de personen. De combinatie met interpretatie maakt het waardevol (dit proefschrift).

6. De kloof tussen de universitaire psychometrie en de daadwerkelijke toepassingen in de praktijk is te groot en lijkt alleen maar toe te nemen.

7. Een onderzoeksvraag levert niet altijd datgene op wat je ervan had verwacht: dat houdt de spanning erin.

8. Ook een niet significante uitkomst of een uitkomst waarin geen verschil gevonden wordt is een waardevolle uitkomst.

9. Het beste moment voor reflectie is in de auto op de snelweg tussen Enschede en Vianen.

10. Geen proefschrift zonder koffie (en IRThee).

Funder, D. C. (1997). The personality puzzle. (3rd ed.). New York: Norton & Company.