PERSON FIT FOR TESTS WITH POLYTOMOUS RESPONSES

ANNA VILLA T. DAGOHOY
Twente University, Enschede
30 September 2005

SAMENSTELLING PROMOTIECOMMISIE

VOORZITTER/SECRETARIS	Prof. dr. H.W.A.M. Coonen
PROMOTOR	Prof. dr. C.A.W. Glas
ASSISTENT PROMOTOR	Dr. R.R. Meijer
LEDEN	Dr. P.A.T.M. Geurts
	Prof. dr. H. Hoijtink
	Prof. dr. W.J. van der Linden
	Prof. dr. J.K. Vermunt

ISBN: 90-365-2248-X
Druk: PrintPartners Ipskamp B.V., Enschede

PERSON FIT FOR TESTS WITH POLYTOMOUS RESPONSES

PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. W.H.M. Zijm,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op vrijdag 30 september 2005 om 16.45 uur.

door
Anna Villa T. Dagohoy
geboren op 25 february 1969
te Marawi City

Acknowledgement

Many people helped in making this research possible and bring this to fruition. Big or small, each piece of contribution was essential and greatly appreciated. I would like to convey my heartfelt gratitude to the following for their valuable contributions and encouragement.

Prof. dr. Cees Glas, for the invaluable discussions, critical comments, unselfish supervision and guidance, which have greatly improved my craft. Dr. Rob Meijer whose assistance and advice greatly contributed to this research.

Dr. Leonardo Sotaridona, who inspired me to carry out the task at hand and who despite his busy schedule, provided me with his valuable inputs and suggestions.

The Dutch people and the Dutch Government, who provided the funds for the entire course.

Dr. Eldigario G. Gonzales, President, Western Mindanao State University, for giving me the opportunity to pursue my doctorate degree. Dr. Elbia Aquino, Dr. Belinda Belisario, Dr. Florentina Lim, Mrs. Rosemary Legados and my CSM colleagues, whose support was always there and can be counted upon.

Prof. dr. Wim van der Linden, who was instrumental to my being able to avail of the Psychometric Society/ETS Travel Award competition and for the support extended.

Ir. Wim Tielen, whose expertise in programming enabled me to get away with most of the programming stuff.

My friends and the BLD-SE6 community in the Philippines, whose constant prayers and support helped me.

My "*sisters*" in Enschede, Irene and Adek; and Gingging for being with me especially during those arduous times.

To my friends in Germany, 4 JG's and Cecile, somehow you made Philippines not so far away.

To all my "*extended-family*" here in Netherlands, the Liem Family and the Purbojo Family, the experience that we shared will always be imprinted in me.

My parents, Pompey and Norma; sister; brothers and in-laws, whose love, understanding and prayers carried me thru.

My stay in The Netherlands was indeed a very pleasant one because of all of YOU!

But most all, TO THE ONE ABOVE, my CONSTANT COMPANION who enabled everything, PRAISE and GLORY to HIS NAME.

Anna Dagohoy

Kurt and Sam

Contents

1 Introduction **1**
1.1 Polytomous IRT Models 2
1.2 Person-fit Analysis 3
1.2.1 Standardized Log-Likelihood Statistic 4
1.2.2 The Corrected-Standardized Weighted Mean Square 5
1.3 Overview of the Thesis 6

2 Person Fit Tests for IRT Models for polytomous items with estimated person and item parameters **9**
2.1 The LM Test 11
2.2 The IRT Model 13
2.3 An LM Test for Constancy of Theta 14
2.4 An LM Test for Local Independence 17
2.5 Incorporating Item Parameter Estimates 19
2.6 Simulation Studies 21
2.6.1 Type I error rate 21
2.6.2 Power of Tests for Dichotomous Items 24
2.6.3 Power of Tests for Polytomous Items 29
2.7 Discussion 31

3 Lagrange Multiplier Person Fit Tests for Polytomous IRT models **37**
3.1 Models for polytomous items 38
3.1.1 The Graded Response Model 38
3.1.2 The sequential model 39
3.1.3 The generalized partial credit model 39
3.2 The LM Test 39
3.2.1 An alternative model for the unidimensional case . . 41
3.2.2 An alternative model of between-items multidimensionality 42
3.3 Simulation Study 1 43
3.3.1 Type I error rate 43
3.3.2 Power of the test 45
3.3.3 Agreement between models 47
3.4 Simulation Study 2 48
3.5 An Empirical Example 49
3.6 Discussion 55
3.7 Appendix A: Detailed characterization of the test statistics 56
3.7.1 First and second order derivatives for the graded response model 58
3.7.2 First and second order derivatives for the sequential model 58
3.7.3 First and second order derivatives for the generalized partial credit model 59
3.8 Appendix B: Details on estimation for simulation study 2 . 60

4 An Application of Person Fit Tests to Multidimensional Personality and Cognitive Tests **63**
4.1 Person Fit in Typical Performance Testing 64
4.2 The LM Test 66
4.2.1 LM Test For Multidimensional Data 67
4.3 Empirical Study 1 69
4.4 Empirical Study 2 73
4.5 Discussion 77

5 A Bayesian Approach to Evaluation of Person Fit to Polytomous IRT Models **81**
5.1 IRT Models and Person Fit 83

5.1.1 Models for polytomous items 83
5.1.2 Person Fit Tests 84
5.1.3 Evaluating the fit of an item score pattern. 86
5.2 Bayesian Estimation of the Models 87
5.3 Simulation Studies 88
5.3.1 Type I error rate 88
5.3.2 Power of the Tests 89
5.4 Discussion 94

6 Summary **97**

7 Samenvatting (Summary in Dutch) **101**

References **105**

1
Introduction

Researchers have given considerable attention to unexpected behavior individuals exhibit when answering psychological or educational tests. In the context of item response theory (IRT; e.g., van der Linden & Hambleton, 1997) person-fit statistics have been proposed (e.g., Drasgow & Levine, 1986; Meijer, 1994; Tatsuoka, 1984) that can be used to investigate the fit of an individual item score pattern to a test model and the effectiveness of these statistics for detecting misfitting item scores has been investigated (e.g., Drasgow, Levine & McLaughlin, 1987, 1991). In IRT, the probability of obtaining a correct or preferred answer to an item is a function of the latent trait value θ and the characteristics of the items such as the difficulty or popularity of the item. Item responses that do not fit the assumed IRT model can cause the latent trait value θ to be inaccurately estimated, or indicate that a person is unscalable on the trait measured (e.g., Reise & Waller, 1993). Possible interpretations of misfitting test behavior include test anxiety, guessing, cheating or response distortion as a result of faking on personality inventories (Zickar & Drasgow, 1996). Other reasons for misfitting response behavior are sleepers who get bored with a test and do poorly on later items, fumblers who do poorly in the beginning because the test format has confused them, plodders who never reach the end of

the test or persons who misalign their answer sheets or show exceptional creativity in interpreting questions.

Most person fit research has concentrated on dichotomously scored items. Only a few studies discussed person fit in the context of polytomous item scores (Drasgow et al., 1985; Wright & Masters, 1982; van Krimpen-Stoop & Meijer, 2002). The main topic of this thesis is the development and application of person fit statistics for polytomous IRT models. Before we introduce person-fit analysis in more detail, we first give a short introduction to polytomous IRT models.

1.1 Polytomous IRT Models

Three polytomous IRT models that are considered in this thesis are the graded response model (Samejima, 1969), the sequential model (Tutz, 1990) and the generalized partial credit model (Muraki, 1992). These models are briefly described below. For further details of these models, the readers are referred to van der Linden and Hambleton (1997).

Consider items $i = 1, ..., k$, with categories $j = 0, ..., m_i$. We will drop the index i of m_i for convenience. Let θ denote the ability of the person, and let α_i and β_i denote the item discrimination and item location parameters of item i, respectively. A response pattern is coded as $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_k)$, a response on an item i as $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{ij}, ..., x_m)$, and $x_{ij} = 1$ if a response was given in category j, and zero otherwise. We will use an abbreviation for the logistic function given by

$$\Psi(x) = \frac{\exp(x)}{1 + \exp(x)} .$$

The Graded Response Model

In the graded response model (GRM) the probability of a response in category j of item i, $P(X_{ij} = 1)$, is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Psi(\alpha_i(\theta - \beta_{i1})) & \text{if } j = 0 \\ \Psi\left(\alpha_i\left(\theta - \beta_{ij}\right)\right) - \Psi\left(\alpha_i\left(\theta - \beta_{i(j+1)}\right)\right) & \text{if } 0 < j < m \\ \Psi(\alpha_i(\theta - \beta_{im})) & \text{if } j = m \ . \end{cases} \tag{1.1}$$

To assure that the probabilities $P_{ij}(\theta)$ are positive, the restriction $\beta_{i(j+1)} > \beta_{ij}$, for $0 < j < m$ is imposed.

The sequential model

In the sequential model (SM) the probability of a response in category j of item i is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Psi\left(\alpha_i\left(\theta - \beta_{i1}\right)\right) & \text{if } j = 0 \\ \prod_{h=1}^{j} \Psi\left(\alpha_i\left(\theta - \beta_{ih}\right)\right)\left[1 - \left(\Psi(\alpha_i(\theta - \beta_{i(j+1)}))\right)\right] & \text{if } 0 < j < m \\ \prod_{h=1}^{m} \Psi\left(\alpha_i\left(\theta - \beta_{ih}\right)\right) & \text{if } j = m\ . \end{cases} \tag{1.2}$$

Verhelst, Glas and de Vries (1997) note that every item in the SM can be viewed as a sequence of virtual dichotomous items. These dichotomous items are considered to be presented as long as a correct response is given, and the presentation stops when an incorrect response is given. An important consequence of this conceptualization of the response process is that estimation and testing procedures for the 2PL model with incomplete data can be directly applied to the SM.

The generalized partial credit model

In the generalized partial credit model (GPCM) the probability of a response in category j of item i is given by

$$P_{ij}(\theta) = \frac{\exp[j\alpha_i\theta - \beta_{ij}]}{1 + \sum_{h=1}^{m_i} \exp[h\alpha_i\theta - \beta_{ih}]}, \tag{1.3}$$

where $\beta_{io} = 0$. The partial credit model (Masters, 1982) is the special case where $\alpha_i = 1$ for all items i and the item parameters are usually re-parameterized as $\beta_{ij} = \sum_{h=1}^{j} \eta_{ih}$.

1.2 Person-fit Analysis

Applications of IRT models to the analysis of test items, tests, and item score patterns are only valid if the IRT model holds. Fit of items can be investigated across persons and fit of persons can be investigated across

items. Item fit is important because in psychological and educational measurement, instruments are developed that are used in a population of persons; item fit then can help the test constructor to develop an instrument that fits an IRT model in that particular population (see, for instance, Andersen, 1973; Yen, 1981, 1984; Molenaar, 1983; Glas, 1988, 1999; Glas & Suarez-Falcon, 2003; and Orlando & Thissen, 2000).

As a next step, the fit of an individual's item score pattern can be investigated. Although a test may fit an IRT model, persons may produce patterns that are highly unlikely given the model. Using person fit statistics, the fit of a score pattern can be determined under the null-hypothesis that the IRT model holds. Meijer and Sijtsma (1995; 2001) give an overview of person fit statistics proposed for various IRT models.

Person-fit statistics have been proposed under various different names such as appropriateness measures (Levine & Rubin, 1979), caution indices (Tatsouka, 1982; 1984), and scalability indices (Reise & Waller, 1993). All these statistics are based, in some way, on the consistency of an individual's item response pattern to an IRT model. Some person-fit indices were developed within the context of a specific models such as the Rasch (1960) model or nonparametric models (Meijer, 1994). Some examples are given below.

1.2.1 Standardized Log-Likelihood Statistic

Drasgow et al. (1985) proposed a standardized log-likelihood statistic l_z for polytomous items. Let P_{ij} denote the response function both defined in Equations (1.1), (1.2) and (1.3). Assuming local independence between all items, the likelihood of a score pattern can be written as

$$L\left(\theta_j\right) = \prod_{i=1}^{I} \prod_{j=0}^{m} (P_{ij})^{x_{ij}}, \tag{1.4}$$

and the log-likelihood function l of $L(\theta_j)$ is

$$l = \log[L(\theta_j)] = \sum_{i=1}^{I} \sum_{j=0}^{m} x_{ij} \log(P_{ij}). \tag{1.5}$$

The standardized version of l is defined as

$$l_z = \frac{l - E[l]}{\sqrt{Var[l]}} \tag{1.6}$$

where

$$E[l] = \sum_{i=1}^{I} \sum_{j=0}^{m} P_{ij} \log(P_{ij}), \tag{1.7}$$

and

$$Var[l] = \sum_{i=1}^{I} \left[\sum_{h=0}^{m} \sum_{k=0}^{m} P_{ih} P_{ik} \log P_{ih} \log \frac{P_{ih}}{P_{ik}} \right]. \tag{1.8}$$

The performance of these statistics in the context of dichotomously scored tests are well-studied. For example, Reise and Due (1991), Drasgow and Levine (1986), and Nering (1996) found that the standardized log-likelihood statistic for dichotomous items proposed by Drasgow, Levine, and Williams (1985) is questionable when the examinees ability parameter is estimated. The fact that an estimated value of θ is plugged in will have an effect on the distribution of the person fit statistic. For short tests this may result in a decreased variance of the purported to be asymptotically standard normally distributed statistic. Snijders (2001) proposed a correction for the use of $\hat{\theta}$ for a family of statistics which are linear in the item scores. l_z is a member of this family. The correction applies to dichotomously scored items; the generalization to polytomously scored items is not yet available.

1.2.2 The Corrected-Standardized Weighted Mean Square

Another person-fit statistic that can be used for polytomous items was proposed by Wright & Masters (1982). They proposed to use the corrected-standardized weighted mean squared residual. Wright and Masters (1982) claimed that their statistic, *W*, follows a standard normal distribution when the model holds. Some research has been conducted using *W* for dichotomous data, where the PCM becomes the Rasch (1960) model and the statistic is equivalent to the statistic proposed by Wright & Stone (1979, Chapter 4). For example, Rogers and Hattie (1987) showed that this statistic was insensitive to the classification of fitting and misfitting item-score patterns.

Also, Hoijtink (1986) showed that the distribution of the dichotomous version of this statistic was far from the standard normal distribution in the case of the Rasch model. Li and Olejnik (1997) also found that in the case of the one-parameter logistic (Rasch) IRT model for dichotomous items, the obtained mean and variance for l_z and W were very close to each other and both deviated significantly from the standard normal distribution.

1.3 Overview of the Thesis

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.

In Chapter 2, two new person fit tests for polytomous IRT models are introduced. The first test is focused on shifts in ability, the second is focused on violation of local independence. Both are Lagrange multiplier tests. It is shown that the derivation of the distribution of Lagrange multiplier statistics can take the effects of the estimation of the item and person parameters into account. The Lagrange multiplier test has an asymptotic χ^2-distribution. Small sample Type I error rates and power are investigated using simulation studies. It is shown that naive test statistics that ignore the effects of estimation of the persons' ability parameters result in incorrect Type I error rates and a marked decrease of power. Incorporating a correction to account for the effects of estimation of the persons' ability parameters results in acceptable Type I error rates and power characteristics; incorporating a correction for the estimation of the item parameters had very little additional effect.

Chapter 3 presents a person fit test based on Lagrange multiplier tests for three IRT models for polytomous items : the GPCM, the SM and the GRM. It is shown that these tests can also be used in the framework of multidimensional ability parameters. A simulation study of the power and Type I error rate is presented. Further it is investigated to what extent the three models give comparable results. Finally, an example using data from NEO Personality Inventory-Revised is presented. In Chapter 4, we apply the multidimensional version of the person fit test proposed in Chapter 3 to personality and cognitive tests and show how we can interpret patterns that are flagged as aberrant. In Chapter 5, we generalize person fit tests for assessments with polytomous responses to the Bayesian framework. In

a Bayesian framework, a Markov chain Monte Carlo procedure can be used to generate samples of the posterior distribution of the parameters of interest. These draws can also be used to compute the posterior predictive distribution of the discrepancy variable. The procedure is worked out in detail for the graded response model and the sequential model. Type I error rate and the power of the test are evaluated using a number of simulation studies. The simulation study is conducted to investigate the differences between three person-fit statistics, model violations and test lengths. Finally, a summary of the main results is given and some suggestion for further research are made.

2

Person Fit Tests for IRT Models for polytomous items with estimated person and item parameters

Applications of item response theory (IRT) models to the analysis of test items, tests, and item score patterns are only valid if the IRT model used holds. Fit of items can be investigated across persons and fit of persons can be investigated across items. Item fit is important because in psychological and educational measurement, instruments are developed that are used in a population of persons; item fit then can help the test constructor to develop an instrument that fits an IRT model in that particular population (see, for instance, Andersen, 1973; Yen, 1981, 1984; Molenaar, 1983; Glas, 1988, 1999; Glas & Suárez-Falcón, 2003; and Orlando & Thissen, 2000).

As a next step, the fit of an individual's item score pattern can be investigated. Although a test may fit an IRT model, persons may produce patterns that are highly unlikely given the model. For instance, some persons may give random responses because they are unmotivated to take the test. Using person fit statistics, the fit of a score pattern can be determined under the null-hypothesis that the IRT model holds. Meijer and Sijtsma (1995; 2001) give an overview of person fit statistics proposed for various IRT models. Most person fit statistics were developed for IRT models for dichotomous items (Levine & Rubin, 1979; Wright & Stone, 1979; Tatsuoka, 1984; Smith, 1985, 1986; Drasgow, Levine, & McLaughlin, 1991).

This chapter has been submitted for publication as: Glas, C.A.W., & Dagohoy, A.V. Person Fit Tests for IRT Models for Polytomous items with estimated person and item parameters.

Person fit tests for polytomous items were developed by Drasgow, Levine, & Williams, 1985; Wright & Masters, 1982; van Krimpen-Stoop & Meijer, 2002). One of the problems of person fit statistics is that the derivation of the distribution of the statistics, both for dichotomously and polytomously scored items, has to account for the influence of the fact that item and person parameters are estimated. This usually decreases the asymptotic variance of most statistics proposed in literature. Therefore, their asymptotic distribution is usually unknown (see, for instance, Nering, 1995 and Reise, 1995). Recently, Snijders (2001) proposed a method for standardization of a broad class of person fit statistics for IRT models for dichotomous items, such that their asymptotic distribution can be properly derived. The method pertains to the influence of the estimates of the ability parameters. The influence of the estimates of the item parameters was not considered, maybe because a study of the effect of uncertainty of item parameter-estimation on ability estimates by Tsutakawa and Johnson (1990) showed only minor effects. This will probably also hold for person fit statistics, but this point has not been systematically investigated.

The purpose of this article is the following. Firstly, a general class of person fit statistics for parametric IRT models for polytomous items (with dichotomous items as a special case) based on the Lagrange multiplier (LM) test will be introduced. The proposed statistics take the effects of parameters estimation into account. The proposed method is an alternative to the approach by Snijders (2001). Secondly, the principle will be applied to a special model for polytomously scored items, the nominal response model (Bock, 1972). Thirdly, the Type I error rate will be assessed using simulation studies in the framework of dichotomous items. Three types of tests will be addressed: (1) naive tests that do not take estimation effects into account, (2) tests that take the effects of ability estimation into account and (3) tests that take both the effects of estimation of the item and person parameters into account. Fourthly, the Type I error rate and power of the tests will be studied in the framework of polytomous items, and finally, some conclusions will be drawn, and some directions for further research will be sketched.

2.1 The LM Test

Recently, LM tests for IRT models have been proposed by Glas (1998, 1999, 2001), Glas and Suárez-Falcón (2003) and Becher, Verhelst and Verstralen (2002). The LM test (Aitchison & Silvey,1958) is equivalent with the efficient score test (Rao, 1947) and the modification index that is commonly used in structural equation modelling (Sörbom, 1989). The purpose of the LM test is to compare two models, say the null-model and some more general model that is derived from the null model by adding parameters. Only the null-model needs to be estimated. Sörbom (1989) shows that the value of the LM statistic is proportional to the expected increase of the conditional likelihood should the additional parameters be estimated. In the score test formulation, the statistic is based on estimation of the null model and performing one Newton-Raphson step for the added parameters. So, the test is based on an estimate that improves the likelihood, but does not completely maximize it under the alternative model. The more common likelihood ratio (LR) test, on the other hand, is based on actual maximization of the likelihood under the alternative model. When applied to evaluate the fit of IRT models, the null model is the IRT model tested, and the alternative models represent model violations. The reason for considering the LM test, where a LR test is available, is that in complicated models with many parameters (such as IRT models) every item and person may be the source of various model violations. Instead of estimating all the alternatives for all person and items, and performing a vast number of LR tests, one can perform a number of LM tests using one estimate under the null model only. So the LM test must be seen as a diagnostic tool, and it derives its relevance from the fact that it serves another purpose than the LR test.

The LM test is grounded on the following rationale. Consider some general parameterized model, and a special case of the general model, the so-called restricted model. The restricted model is derived from the general model by imposing constraints on the parameter space. In many instances, this is accomplished by fixing one or more parameters of the general model to constants. The LM test is based on the evaluation of the first-order partial derivatives of the log-likelihood function of the general model, evaluated using the maximum likelihood estimates of the restricted model. The unrestricted elements of the vector of first-order derivatives are equal to zero, because their values originate from solving the likelihood equations.

The magnitudes of the elements of the vector of first-order partial derivatives corresponding to restricted parameters determine the value of the statistic: the closer they are to zero, the better the model fit.

More formally, the principle can be described as follows. Consider a general model with parameters $\boldsymbol{\eta}$. In the applications presented below, the special model is derived from the general model by fixing one or more parameters to zero. So if the vector of the parameters of the general model, say $\boldsymbol{\eta}$, is partitioned $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, the null hypothesis entails $\boldsymbol{\eta}_2 = 0$. Let $\mathbf{h}(\boldsymbol{\eta})$ be the first order partial derivatives of the log-likelihood of the general model, that is, $\mathbf{h}(\boldsymbol{\eta}) = \partial \log L(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$. This vector of partial derivatives gauges the change of the log-likelihood as a function of local changes in $\boldsymbol{\eta}$. Let the vector of partial derivatives $\mathbf{h}(\boldsymbol{\eta})$ be partitioned as $(\mathbf{h}(\boldsymbol{\eta}_1), \mathbf{h}(\boldsymbol{\eta}_2))$. Then the test is based on the statistic

$$\text{LM} = \mathbf{h}(\boldsymbol{\eta}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{h}(\boldsymbol{\eta}_2), \tag{2.1}$$

where

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \tag{2.2}$$

and

$$\boldsymbol{\Sigma}_{pq} = -\frac{\partial^2 \log L(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}_p \partial \boldsymbol{\eta}_q'},$$

for $p = 1, 2$ and $q = 1, 2$. An interpretation of the role of the matrices $\boldsymbol{\Sigma}_{pq}$ will be given below. The LM statistic is evaluated using the maximum likelihood estimates of the parameters of the special model under the null hypothesis. In the applications presented below, the model under the null hypothesis will be an IRT model. The LM statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters in $\boldsymbol{\eta}_2$ (Rao (1947, Aitchison & Silvey, 1958).

The variance of the parameter estimates plays the following role in the distribution of the LM statistics. Glas (1999) shows that the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{22}$ in (2.2) can be viewed as the asymptotic covariance matrices of $\mathbf{h}(\boldsymbol{\eta}_2)$ with $\boldsymbol{\eta}_1$ (in the present case, the ability parameter) estimated and known, respectively. Further, $\boldsymbol{\Sigma}_{11}^{-1}$ is the asymptotic covariance matrix of the estimate of $\boldsymbol{\eta}_1$, so the term $\boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$ accounts for the influence of the estimation of $\boldsymbol{\eta}_1$ on the covariance matrix of $\mathbf{h}(\boldsymbol{\eta}_2)$. So in the LM test, the variance of the estimates of the person parameters is explicitly taken into account.

Besides a test of significance, this approach also provide information with respect to the importance of the model violation. This is done by computing a new value of the fixed parameters, say ϕ_{02}^*, by performing one Newton-Raphson step, that is,

$$\boldsymbol{\eta}_2^* = \boldsymbol{\Sigma}^{-1}\ h(\boldsymbol{\eta}_2). \tag{2.3}$$

Testing whether $\boldsymbol{\eta}_2^*$ significantly differs can be done using Rao (1947) efficient score test. Rao shows that, assuming asymptotic normality of the estimates, $\boldsymbol{\eta}_2^*$ has a multivariate normal distribution with mean zero and dispersion matrix $\boldsymbol{\Sigma}$. Hence, $\boldsymbol{\eta}_2^*\boldsymbol{\Sigma}^{-1}\boldsymbol{\eta}_2^*$ has asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters fixed in the null-model. The test based on this statistic is asymptotically equivalent to the LM test (see, for instance, Buse, 1982).

2.2 The IRT Model

In the discussion section, it will be argued that the LM test proposed here can be used for a general class of parameterized IRT models for polytomous items. However, here we will consider the application to the nominal categories model by Bock (1972). Let the items in a test be labeled $i = 1, ..., K$. Every item has m_i response categories labeled $j = 0, ..., m_i$. Item responses will be coded by stochastic variables X_{ij} $(i = 1, ..., K;\ j = 0, ..., m_i)$ with realizations x_{ij}, and $x_{ij} = 1$ if a response was given in category j, and zero otherwise. The probability of a response in category j of item i is given by

$$P_{ij}(\theta) = P(X_{ij} = 1 \mid \theta) = \frac{\exp[\alpha_{ij}\theta - \beta_{ij}]}{1 + \sum_{h=1}^{m_i} \exp[\alpha_{ih}\theta - \beta_{ih}]}, \tag{2.4}$$

where it is assumed that $\alpha_{i0} = \beta_{i0} = 0$ for $i = 1, ..., K$. This general model has several interesting special cases. The first one is the generalized partial credit model (GPCM) by Muraki (1992), which follows upon introduction of the restriction $\alpha_{ij} = j\alpha_i$. Assuming that $\alpha_i > 0$, the model applies to ordered categories, such as used in tests of proficiency and ability. If the parameters α_i are considered as known constants, $\Sigma_i X_{ij} j\alpha_i$ is the sufficient statistic for θ, so an item score in a higher category reflects a higher ability level θ. The partial credit model by Masters (1982) is the special case where $\alpha_{ij} = j$ for all items $i = 1, ..., K$. In that model, the item parameters

are usually re-parameterized as $\beta_{ij} = \sum_{h=1}^{j} \eta_{ih}$. The rating scale model (Andersen, 1977, Andrich, 1978a, 1978b) is the special case where $\eta_{ij} = \xi_i + \tau_j$, that is in that case, the item parameters consist of an item location parameter ξ_i and category parameters τ_j. Finally, below the Rasch model (Rasch, 1960) will be used for some of the simulations; this is the special case where $m_i = 1$ and $\alpha_{i1} = 1$ for all items.

Assuming local independence of responses given θ, the likelihood of a response pattern $\mathbf{x}$ is given by

$$p_\theta(\mathbf{x}) = \prod_{i=1}^{k} \prod_{j=0}^{m_i} P_{ij}(\theta)^{x_{ij}}. \tag{2.5}$$

and the log-likelihood by

$$\log L(\theta) = \log p_\theta(\mathbf{x}) = \sum_{i=1}^{k} \sum_{j=0}^{m_i} x_{ij} \log P_{ij}(\theta), \tag{2.6}$$

with $P_{ij}(\theta)$ as defined in (2.4).

2.3 An LM Test for Constancy of Theta

For the Rasch model, Smith (1985, 1986) introduced a Pearson-type test statistic for evaluating the constancy of the ability parameter across subtests. For the UB test, the complete response pattern is split up into a number of parts, say the parts $g = 0, ..., G$, and it is evaluated whether the same ability parameter θ can account for all partial response patterns. In this section, this approach will be generalized to polytomously scored items, and the theory of the LM statistic will be used to derive the asymptotic distribution of the statistic.

Let A_g be the set of the indices of the items in part g. We pose the alterative model that the response pattern cannot be described by one ability parameter, that is, for $g > 0$, we pose the model

$$P_{ijg}(\theta) = P(X_{ij} = 1 \mid \theta, i \in A_g) = \frac{\exp[\alpha_{ij}(\theta_0 + \theta_g) - \beta_{ij}]}{1 + \sum_{h=1}^{m_i} \exp[\alpha_{ih}(\theta_0 + \theta_g) - \beta_{ih}]}. \tag{2.7}$$

For $g = 0$, the model given by (2.4) holds, so this partial response pattern is used as a reference. For the remainder of the response pattern, it is hypothesized that additional ability parameter θ_g $(g = 1, ..., G)$ are necessary to describe the response behavior.

In this section, an LM test accounting for the effects of estimation of the person parameter θ will be derived. An LM test that also accounts for the effects of estimation of the item parameters will be treated in a following section. To apply the framework of the LM statistic, we first derive the derivatives with respect to the ability parameters. It can be verified that

$$\frac{\partial P_{ij}(\theta)}{\partial \theta} = P_{ij}(\theta) \left[\alpha_{ij} - \sum_{h=1}^{m_i} \alpha_{ih} P_{ih}(\theta) \right], \tag{2.8}$$

so for the special model, the first order derivative of the log-likelihood is given by

$$\begin{aligned} \frac{\partial \log L(\theta)}{\partial \theta} &= \sum_{i=1}^{k} \sum_{j=0}^{m_i} \left[x_{ij} \left(\alpha_{ij} - \sum_{h=1}^{m_i} \alpha_{ih} P_{ih}(\theta) \right) \right] \\ &= \sum_{i=1}^{k} \left[y_i - E_\theta(Y_i) \right], \end{aligned}$$

where $y_i = \sum_{j=0}^{m_i} x_{ij} \alpha_{ij}$, that is, it is the weighted score on item i, and $E_\theta(Y_i)$ is its expectation. In the same manner, for the general model, we have

$$\frac{\partial \log L(\theta)}{\partial \theta_g} = \sum_{i \in A_g} \left[y_i - E_\theta(Y_i) \right].$$

For the second order derivatives of the log-likelihood we obtain

$$\begin{aligned}
\frac{\partial^2 \log L(\theta)}{\partial \theta^2} &= -\sum_{i=1}^{k}\sum_{j=0}^{m_i} \alpha_{ij} P_{ij}(\theta)\left[\alpha_{ij} - E_\theta(Y_i)\right], \\
\frac{\partial^2 \log L(\theta)}{\partial \theta_g^2} &= -\sum_{i\in A_g}^{k}\sum_{j=0}^{m_i} \alpha_{ij} P_{ij}(\theta)\left[\alpha_{ij} - E_\theta(Y_i)\right], \\
\frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta_g} &= -\sum_{i\in A_g}^{k}\sum_{j=0}^{m_i} \alpha_{ij} P_{ij}(\theta)\left[\alpha_{ij} - E_\theta(Y_i)\right], \\
\frac{\partial^2 \log L(\theta)}{\partial \theta_g \partial \theta_{g'}} &= 0,
\end{aligned}$$

where $g > 0$, and $g' > 0$. Inserting these formula's into (2.1) and (2.2) gives an LM statistics for testing the constancy of the ability parameter over partial response patterns.

We will now develop the test for the case where the test is split up into two subtests: the first part and the second part. So $G = 1$. Then the test statistic can be written as

$$LM = \left[\frac{\partial \log L(\theta)}{\partial \theta}, \frac{\partial \log L(\theta)}{\partial \theta_1}\right] \begin{bmatrix} \frac{\partial^2 \log L(\theta)}{\partial \theta^2} & \frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta_1} \\ \frac{\partial^2 \log L(\theta)}{\partial \theta \partial \theta_1} & \frac{\partial^2 \log L(\theta)}{\partial \theta^2} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial \log L(\theta)}{\partial \theta} \\ \frac{\partial \log L(\theta)}{\partial \theta_1} \end{bmatrix},$$

which we will write as

$$LM = [h, h_1] \begin{bmatrix} s & s_{10} \\ s_{10} & s_1 \end{bmatrix}^{-1} \begin{bmatrix} h \\ h_1 \end{bmatrix}.$$

But when the estimation equation is solved, it holds that $h = 0$, and inversion of the matrix of the quadratic form results in

$$LM = \frac{1}{s_1 s - s_{10}^2} [0, h_1] \begin{bmatrix} s_1 & -s_{10} \\ -s_{10} & s \end{bmatrix} \begin{bmatrix} 0 \\ h_1 \end{bmatrix},$$

that is,

$$LM = \frac{h_1^2}{s_1 - s_{10}^2/s}. \tag{2.9}$$

This LM statistic has an asymptotic χ^2-distribution with one degree of freedom.

Disregarding all covariances and disregarding the effects of estimation of θ results in a statistic

$$UB = \sum_{g=1}^{G} \frac{\left[\sum_{i \in A_g} [y_i - E_\theta(Y_i)]\right]^2}{\sum_{i \in A_g}^{k} \sum_{j=0}^{m_i} \alpha_{ij} P_{ij}(\theta) [\alpha_{ij} - E_\theta(Y_i)]}. \tag{2.10}$$

For the case of dichotomously scored items, the formulation of this statistic is almost similar to the formulation of the UB-statistic by Smith (1985, 1986). The main difference is that Smith also includes the base line response pattern, that is, the pattern for the items in A_0, and multiplies with a factor $1/G$. For polytomously scored items, this would result in

$$UB = \frac{1}{G} \sum_{g=0}^{G} \frac{\left[\sum_{i \in A_g} [y_i - E_\theta(Y_i)]\right]^2}{\sum_{i \in A_g}^{k} \sum_{j=0}^{m_i} \alpha_{ij} P_{ij}(\theta) [\alpha_{ij} - E_\theta(Y_i)]}. \tag{2.11}$$

The statistics given by (2.10), (2.11) and (2.13) will be compared in the simulation studies reported below.

2.4 An LM Test for Local Independence

Evaluation of local independence can be based on alternative models where the responses to particular items depend on responses to preceding items. In the framework of the Rasch model, such models were proposed by Kelderman (1984) and Jannarone (1986). In the application considered here, the dependence between the response on item i and the response on item k is modeled by the introduction of a parameter δ. Consider a model where the response on item i depends on a statistic Y_k, defined by $Y_k = \sum_{j=0}^{m_i} X_{kj} \alpha_{kj}$. The model is given by

$$\begin{aligned} P_{ij}(\theta) &= P(X_{ij} = 1 \mid Y_k = y_k) \\ &= \frac{\exp[\alpha_{ij}\theta - \beta_{ij}]}{1 + \sum_{h=1}^{m_i} \exp[\alpha_{ih}\theta - \beta_{ih}]} \\ &= \frac{\exp[\alpha_{ij}(\theta_0 + \delta y_k) - \beta_{ij}]}{1 + \sum_{h=1}^{m_i} \exp[\alpha_{ih}(\theta_0 + \delta y_k) - \beta_{ih}]}. \end{aligned} \tag{2.12}$$

Note that δy_k can be interpreted as a shift in ability that is proportional to the ability level that is reflected in the response on item k. That is, if α_{ij} $(j = 1, ..., m_k)$ were known, Y_k would be a sufficient statistic for ability, and, a score in category g of item k, would result in weighting δ with α_{kg}. For the Rasch model for dichotomous items, $\alpha_i = \alpha_k = 1$, and δ can also be interpreted as a shift in the ability parameter following a correct response on item k. For the 2PLM, this shift is weighted by the discrimination parameter α_k, which would be the contribution to the sufficient statistic for ability if the discrimination parameters were known scoring weights. In the general polytomous case, α_{kj} $(j = 1, ..., m_k)$ also act as scoring weights, which motivates modeling the dependence as the shift $\delta\alpha_{kg}$.

The derivatives needed for the LM statistics are derived as follows. First, it can be verified that

$$\begin{aligned}\frac{\partial P_{ij}(\theta)}{\partial \delta} &= \frac{\partial P_{ij}(\theta)}{\partial \theta}\frac{\partial(\theta_0 + \delta y_k)}{\partial \delta} \\ &= P_{ij}(\theta)\left[\alpha_{ij} - \sum_{h=1}^{m_i}\alpha_{ih}P_{ih}(\theta)\right] y_k.\end{aligned}$$

Suppose $k(i)$ is the item on which item i depends, for instance, $k(i) = i-1$. Then, for the special model, the first order derivative of the log-likelihood is given by

$$\begin{aligned}\frac{\partial \log L(\delta)}{\partial \delta} &= \sum_{i=1}^{K}\sum_{j=0}^{m_i} y_{k(i)}\left[x_{ij}\left(\alpha_{ij} - \sum_{h=1}^{m_i}\alpha_{ih}P_{ih}(\theta)\right)\right] \\ &= \sum_{i=1}^{K} y_{k(i)}\left[y_i - E_\theta(Y_i)\right].\end{aligned}$$

For the second order derivatives of the log-likelihood we obtain

$$\begin{aligned}\frac{\partial^2 \log L(\theta)}{\partial \delta^2} &= -\sum_{i=1}^{K}\sum_{j=0}^{m_i} y_{k(i)}^2\alpha_{ij}P_{ij}(\theta)\left[\alpha_{ij} - E_\theta(Y_i)\right], \\ \frac{\partial^2 \log L(\theta)}{\partial \theta_0 \partial \delta} &= -\sum_{i=1}^{k}\sum_{j=0}^{m_i} y_{k(i)}\alpha_{ij}P_{ij}(\theta)\left[\alpha_{ij} - E_\theta(Y_i)\right].\end{aligned}$$

The LM statistic follows upon inserting these expressions in (2.9). The statistic has an asymptotic χ^2-distribution with one degree of freedom.

Disregarding all covariances and disregarding the effects of estimation of θ results in a statistic

$$UD = \frac{\left[\sum_i y_{k(i)} \left[y_i - E_\theta(Y_i)\right]\right]^2}{\sum_i \sum_{j=0}^{m_i} y_{k(i)}^2 \alpha_{ij} P_{ij}(\theta) \left[\alpha_{ij} - E_\theta(Y_i)\right]}. \tag{2.13}$$

This statistic might be viewed as the generalization of the UB-statistic to a test for local independence.

2.5 Incorporating Item Parameter Estimates

Marginal maximum likelihood (MML) estimation is probably the most used technique for item calibration. For the 1-, 2- and 3PLM, the theory was developed by such authors as Bock and Aitkin (1981), Thissen (1982), Rigdon and Tsutakawa (1983), and Mislevy (1984, 1986), and computations can be made using the software package Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). In the MML approach, it is assumed that the ability parameters are independent and normally distributed. The approach derives its name from maximizing a log-likelihood that its marginalized with respect to θ, rather than maximizing the joint log-likelihood of all abilities parameters θ and all item parameters. The reason for maximizing the marginal rather than the joint likelihood is that maximizing the latter does not lead to consistent estimates. This is related to the fact that the number of ability parameters grows proportional with the number of observations and, in general, this leads to inconsistency (Neyman & Scott, 1948). Simulation studies by Wright and Panschapakesan (1969) and Fischer and Scheiblechner (1970) show that these inconsistencies can indeed occur in IRT models. Kiefer and Wolfowitz (1956) have shown that MML estimates are consistent under fairly reasonable regularity conditions.

The essential feature of MML estimation is that the number of parameters is constant in relation to the number of observations. Therefore, for the present application, we consider a model where we marginalize over all ability parameters except the one we are interested in. Further, we now consider all data available. Therefore, the log-likelihood is split up into two parts, one pertaining to the marginal likelihood of N respondents and one

pertaining to the respondent that is the focus of attention (say, observation $N+1$). So we have

$$\begin{aligned} \log L &= \log L_m + \log L_p \qquad (2.14) \\ &= \sum_{n=1}^{N} \log \int \prod_{i=1}^{k} \prod_{j=0}^{m_i} P_{ij}(\theta)^{x_{nij}} G(\theta) d\theta + \sum_{i=1}^{k} \sum_{j=0}^{m_i} x_{ij} \log P_{ij}(\theta), \end{aligned}$$

where x_{nij} are the responses of the N persons, and $G(\theta)$ is a, usually normal, ability distribution. The log-likelihood in (2.14) can be concurrently maximized with respect to the item-parameters, the population parameters (parameters of $G(\theta)$), and the ability parameter of the focus person.

In the section on the LM test, the parameter vector was partitioned $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, where, under the null model, $\boldsymbol{\eta}_1$ are the free parameters and $\boldsymbol{\eta}_2$ are the fixed parameters. In the present case, the item-parameters, the population parameters, and the ability parameter of the focus person are stacked in $\boldsymbol{\eta}_1$ and the parameters representing model violations (that is, θ_g , $g = 1, ..., G$, for violation of constancy or δ for violation of local independence) are stacked in $\boldsymbol{\eta}_2$. The parameters in $\boldsymbol{\eta}_1$ are partitioned into item and population parameters $\boldsymbol{\xi}$ and the focus-respondent's ability θ. To perform the test for a specific respondent, we proceed in two steps: first we estimate $\boldsymbol{\xi}$, and then we compute the LM statistic defined by (2.1). The first step boils down to solving the simultaneous system

$$\begin{aligned} \frac{\partial[\log L_m + \log L_p]}{\partial \xi} &= 0, \\ \frac{\partial \log L_p}{\partial \theta} &= 0. \end{aligned}$$

First and second order derivatives of $\log L_m$ with respect to item- and population parameters $\boldsymbol{\xi}$ can be found in Glas (1999), the derivatives of $\log L_p$ with respect to θ were given above, and the derivatives of $\log L_p$ with respect to the item parameters (population parameters don't figure in $\log L_p$) can be found in articles on joint maximum likelihood estimation for IRT, say, Wright and Linacre (1992). In practice, the estimates of the item and population parameters $\boldsymbol{\xi}$ will not change much when one respondent is singled out as a target; in practice, only a few iteration steps are needed.

The LM statistic (2.1) can then be computed with $\mathbf{h}(\boldsymbol{\eta}_2) = \partial \log L_p / \partial \boldsymbol{\eta}_2$, where $\boldsymbol{\eta}_2$ is either δ or θ_g $(g = 1, ..., G)$ and a matrix of weights

$$\Sigma = \frac{\partial^2 \log L_p}{\partial \boldsymbol{\eta}_2^2} - \left[\begin{array}{cc} \frac{\partial^2 \log L_p}{\partial \boldsymbol{\eta}_2 \partial \xi^t} & \frac{\partial^2 \log L_p}{\partial \boldsymbol{\eta}_2 \partial \theta} \end{array} \right] W^{-1} \left[\begin{array}{c} \frac{\partial^2 \log L_p}{\partial \xi \partial \boldsymbol{\eta}_2} \\ \\ \frac{\partial^2 \log L_p}{\partial \theta \partial \boldsymbol{\eta}_2} \end{array} \right],$$

with $W = \left[\begin{array}{cc} \frac{\partial^2 [\log L_m + \log L_p]}{\partial \xi \partial \xi^t} & \frac{\partial^2 \log L_p}{\partial \xi \partial \theta} \\ \\ \frac{\partial^2 \log L_p}{\partial \theta \partial \xi^t} & \frac{\partial^2 \log L_p}{\partial \theta^2} \end{array} \right]$.Also for this expression, derivatives of $\log L_m$ with respect to $\boldsymbol{\xi}$ can be found in Glas (1999), derivatives of $\log L_p$ with respect to $\boldsymbol{\xi}$ can be found in articles on joint maximum likelihood estimation, and derivatives of $\log L_p$ with respect to $\boldsymbol{\eta}_2$ and θ can be found above.

2.6 Simulation Studies

Three sets of simulation studies will be reported. In the first set a comparison is made between the Type I error rate of naive tests, tests that take the ability estimate into account and tests that take both the estimates of the item and person parameters into account. These studies pertain to dichotomous items. In the second set of simulations, the power of the tests for dichotomous items will be studied. These studies also address the false alarm rate and the specificity of the tests, that is, the extent to which tests are sensitive to other model violations than the one they are targeted at. In the third set of simulations, this is repeated for polytomous items.

2.6.1 Type I error rate

The aim of this section is to assess whether the theoretical advantage of taking the effects of estimation into account pays off in practice. Obtaining item parameter estimates can be quite difficult. In the 2- and 3PLM, item parameter estimates are sometimes hard to obtain, because the parameters are poorly determined by the available data, in the sense that in the region of the ability scale where the respondents are located, the item characteristic curves can be appropriately described by a large number of sets of item parameter values. To obtain "reasonable" and finite estimates, Mislevy (1986) considers a number of Bayesian approaches, entailing the introduction of prior distributions on the parameters. In models for polytomous

items, item categories may not attract responses, and special adjustments have to be made to the model to obtain estimates (Wilson & Masters, 1993). These problems do not present any essential problem for the techniques presented here, but to keep the study simple and tractable, the Rasch model was used for this set of simulations. Sample sizes of $N = 100$, $N = 1000$, and $N = 4000$ were crossed with test lengths of $K = 20$, $K = 40$ and $K = 60$. For the test length $K = 20$, the item parameters were equal to $\beta_i = -2.00 + 0.20(i - 1)$, $i = 1, ..., 20$. For the test lengths $K = 40$ and $K = 60$ these values were repeated two and three times, respectively. The person ability parameters θ were drawn from a standard normal distribution. Item parameters were estimated using MML, person ability parameters using maximum likelihood. All tests were computed using a 5% significance level, and 100 replications were made for each branch in the simulation design.

The UB- and UD-statistics (Formulas (2.10) and (2.13), respectively) were computed in three conditions: (1) using the true item and person parameters, (2) using true item parameters and estimated ability parameters, and (3) using both estimated item and ability parameters. The UB-statistic given by (2.11) was also computed, but the results for this version of the test were so close to the results of the version given by (2.10) that they are not reported here. The LM statistics accounting for the ability parameter estimates will be labeled LM_{B1} and LM_{D1}. The first is targeted at the constancy of the ability parameter, the second to local independence. LM_{B1} was computed using a partition of the items into two subtests. These two statistics were computed in two conditions: (1) using true item parameters and estimated ability parameters, and (2) using both estimated item and ability parameters. Finally, LM statistics accounting for both the ability and item parameter estimates will be labeled LM_{B2} and LM_{D2}, they are defined analogous to LM_{B1} and LM_{D1}. The statistics were computed using estimates of the item and person parameters obtained as outlined in the previous section. The results for UB, LM_{B1} and LM_{B2} are shown in Table 2.1 and the results for UD, LM_{D1} and LM_{D2} are shown in Table 2.2, respectively.

In the third columns of the two tables, it can be seen that when the true values for all parameters were used, the Type I error rates of the UB- and UD-tests were very close to the nominal significance level. This is as expected, because computed this way, the test does not involve estimation.

Table 2.1
Type I error rate for tests for constancy of theta

			UB			LM_{B1}		LM_{B2}
	Theta	True	Estimated	Estimated	Estimated	Estimated	Estimated	Estimated
	Beta	True	True	Estimated	True	Estimated	Estimated	
K	N							
20	100	.042	.008	.007	.049	.041	.044	
	1000	.042	.007	.007	.043	.042	.042	
	4000	.042	.007	.007	.041	.041	.041	
40	100	.049	.008	.005	.068	.050	.055	
	1000	.047	.005	.005	.051	.045	.045	
	4000	.047	.005	.005	.046	.045	.045	
60	100	.047	.014	.004	.087	.050	.054	
	1000	.048	.006	.005	.055	.051	.051	
	4000	.048	.005	.005	.051	.051	.051	

Table 2.2
Type I error rate for tests for local independence

			UD			LM_{D1}		LM_{D2}
	Theta	True	Estimated	Estimated	Estimated	Estimated	Estimated	Estimated
	Beta	True	True	Estimated	True	Estimated	Estimated	
K	N							
20	100	.044	.005	.004	.051	.043	.048	
	1000	.044	.004	.004	.044	.043	.044	
	4000	.044	.004	.004	.043	.043	.043	
40	100	.049	.008	.005	.069	.047	.051	
	1000	.048	.006	.005	.050	.047	.048	
	4000	.048	.005	.005	.048	.048	.048	
60	100	.049	.014	.005	.101	.049	.054	
	1000	.050	.007	.006	.055	.049	.049	
	4000	.049	.006	.006	.051	.049	.049	

In the next two columns it can be seen that the Type I error rate decreased substantially when parameter estimates were used. The Type I error rates of the LM_{B1}- and LM_{D1}-tests were close to 5% in both in the cases. So the deterioration of the Type I error rate of the UB- and UD-tests was solved by using LM tests that explicitly take the effects of estimation of θ into account. Note that, for both LM_{B1} and LM_{D1}, the Type I error rate was slightly inflated for the case where $N = 100$ and the true item parameters were used. The effect vanished when estimates of the item parameters were used. Finally, the Type I error rates of the LM_{B2}- and LM_{D2}-tests were not substantially closer to 5% than the Type I error rates of the LM_{B1}- and LM_{D1}-tests. So explicitly taking the effects of the estimation of the item parameters into account did not result in a marked improvement.

2.6.2 Power of Tests for Dichotomous Items

The next set of simulation studies pertains to the power of the LM_{B1}-, LM_{B2}-, LM_{D1}-, and LM_{D1}-test to detect model violations in aberrant persons. (Given the results reported in the previous section, the UB- and UD-tests were not included here, but they will return in the next section). The MML estimation procedure was run using the data of all simulees, both the aberrant and non-aberrant ones. In all simulations, 10% of the simulees were aberrant. The presence of the aberrant simulees did, of course, produce some bias in the parameter estimates, but this setup was considered realistic because in many situations it is not a priori known which respondents are aberrant, and which are not. Item parameters were equal to the item parameters in the previous study, unless indicated otherwise, and the ability parameters θ were again drawn from a standard normal distribution. Three simulation studies will be reported, pertaining to detection of changes in ability, detection of violation of local independence and detection of guessing. In all simulations, test length was equal to $K = 40$ and $K = 60$. The samples sizes were $N = 400$ and $N = 1000$. The number of replications in each branch of the study was equal to 100.

In the simulations with respect to detection of changes in ability, besides test length and sample size, the effect size of the model violation and the number of items that were affected by this violation were varied. The model violation was imposed by assuming that the ability parameter shifted in the last part of the test and choosing the parameter θ_1 in (2.7) equal to 0.5 or 1.0. Further, this shift in ability occurred either in the last

Table 2.3
Detection of changes in ability

				LM_{B1}		LM_{B2}		LM_{D1}		LM_{D2}	
K	N	θ_1	Items Infected	False Alarms	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms	Hits
40	400	0.5	10	.05	.05	.05	.05	.05	.05	.05	.05
			20	.06	.08	.06	.08	.05	.05	.05	.05
		1.0	10	.06	.07	.06	.07	.05	.04	.05	.04
			20	.06	.15	.06	.15	.05	.05	.05	.05
	1000	0.5	10	.05	.05	.05	.05	.05	.04	.05	.05
			20	.05	.08	.05	.10	.05	.05	.05	.06
		1.0	10	.05	.07	.05	.07	.05	.05	.05	.04
			20	.05	.23	.05	.22	.05	.06	.05	.05
60	400	0.5	15	.05	.05	.05	.06	.05	.05	.05	.05
			30	.06	.09	.06	.08	.05	.05	.05	.05
		1.0	15	.06	.07	.06	.08	.05	.05	.05	.05
			30	.07	.20	.07	.21	.05	.05	.05	.05
	1000	0.5	15	.05	.06	.05	.06	.05	.05	.05	.05
			30	.05	.10	.05	.10	.05	.05	.05	.05
		1.0	15	.05	.09	.05	.09	.05	.05	.05	.06
			30	.06	.26	.06	.26	.05	.06	.05	.06

half or the last quarter of the test. The results are displayed in Table 2.3. The proportion of false alarms is the proportion of the 90% non-aberrant simulees where the test was significant at 5%. The proportion of hits is the proportion of significant tests for the 10% aberrant simulees. Note that the power of LM_{B1} and LM_{B2} is very limited. The are main effects on power of the test length and of the number of items infected, that is, of the number of items were the model violation was imposed. The explanation of the two effects is that the model violation can be better detected if the amount of aberrant data is larger. Further, a longer test length may provide more information for the estimation of the ability parameter. However, below it will become apparent that this does not always hold. The sample size did not have a marked effect, so the more precise estimation of the item parameters for the case where $N = 1000$ did not improve the power. The power of LM_{B1} has comparable to the power of LM_{B2} so taking the effects of the estimation of item parameters into account did not have a marked effect. In the last four columns, it can be seen that the power of LM_{D1} and LM_{D2} to detect changes in ability is negligible. Finally, it can be seen that, for all tests, the proportion of false alarms very close to the Type I error rate.

Before turning to the violation of local independence, first a second example of a change in ability will be studied. In this example, the simulees guess the response on part of the test. The probability of a correct response was equal to 0.2. This violation is more severe than the previous one. In the previous simulation the item parameters were not changed, so for aberrant simulees the probabilities of correct responses were uniformly shifted for the affected part of the test. In the present simulation, guessing implies that the original item parameters lose their meaning, that is, all item are equally difficult. Further, the setup of this study was analogous to the previous one, except that conditions were added where the respondents guessed three-quarters of the test of the while test. The results are shown in Table 2.4. Note that the power is now much larger than in the previous study. This is as expected, because the model violation is more serious. Also LM_{D1} and LM_{D2} have power to detect the violation, although their power is smaller that the power of LM_{B1} and LM_{B2}. Again, there are main effects on power of the test length and of the number of items infected, that is, of the number of items were the model violation was imposed. It is interesting to note that the power functions are single peaked in the proportion of affected items.

Table 2.4
Detection of guessing

K	N	Guessed Items	LM_{B1} False Alarms	LM_{B1} Hits	LM_{B2} False Alarms	LM_{B2} Hits	LM_{D1} False Alarms	LM_{D1} Hits	LM_{D2} False Alarms	LM_{D2} Hits
40	400	10	.05	.55	.05	.56	.04	.32	.04	.32
		20	.06	.85	.06	.84	.04	.35	.04	.35
		30	.05	.67	.05	.68	.04	.23	.04	.22
		40	.04	.39	.04	.38	.04	.05	.04	.05
	1000	10	.05	.55	.05	.55	.04	.32	.04	.32
		20	.06	.86	.06	.84	.04	.36	.04	.33
		30	.04	.55	.05	.55	.05	.31	.05	.29
		40	.04	.38	.04	.39	.04	.05	.04	.06
60	400	10	.05	.41	.05	.41	.05	.23	.05	.23
		20	.05	.72	.05	.72	.05	.31	.05	.31
		30	.06	.83	.06	.84	.04	.28	.05	.28
		45	.05	.74	.05	.75	.05	.13	.05	.13
		60	.05	.75	.05	.74	.04	.07	.05	.07
	1000	10	.05	.41	.05	.41	.05	.23	.05	.23
		20	.05	.72	.05	.72	.04	.31	.04	.32
		30	.06	.83	.06	.83	.05	.27	.05	.28
		45	.05	.75	.05	.75	.04	.13	.05	.13
		60	.05	.75	.05	.75	.04	.07	.05	.07

Table 2.5
Detection of violation of local independence

			LM_{B1}		LM_{B2}		LM_{D1}		LM_{D2}	
K	δ	Items Infected	False Alarms	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms	Hits
40	0.5	10	.04	.05	.04	.06	.05	.04	.05	.04
	0.5	20	.05	.07	.05	.07	.05	.05	.05	.05
	1.0	10	.05	.06	.05	.07	.05	.05	.05	.05
	1.0	20	.05	.13	.05	.13	.05	.07	.05	.07
	1.5	10	.05	.08	.05	.08	.05	.07	.05	.06
	1.5	20	.06	.23	.06	.24	.05	.12	.05	.12
60	0.5	15	.05	.06	.05	.06	.05	.04	.05	.04
	0.5	30	.05	.08	.05	.08	.05	.06	.05	.06
	1.0	15	.05	.07	.05	.07	.05	.06	.05	.06
	1.0	30	.05	.16	.05	.16	.05	.10	.05	.10
	1.5	15	.05	.11	.05	.11	.05	.08	.05	.08
	1.5	30	.05	.29	.06	.30	.05	.18	.05	.18

The reason is that the ability estimate deteriorates if the number of affected items becomes too large, so non-aberrant and aberrant responses can no longer be distinguished.

For the study to violation of local independence, three values of the parameter δ in defined in (2.12) were chosen: 0.5, 1.0, and 1.5. In the design of the study, these values were crossed with two test lengths, $N = 40$ and $N = 60$, and two values for the proportion of items infected: a quarter and a half . Because sample size had little effect in the previous studies, only a sample size of $N = 1000$ was chosen. The results are displayed in Table 2.5. It can be seen that the power of the tests is small. All three factors, test length, effect size and numbers of items infected had the expected, but small, main effects. Interestingly, the power of the tests specifically aimed at detection of violation of local independence was lower than the power of the tests aimed at shifts in the ability parameter. This means that the specificity of the tests is limited: they have power (above their significance level) for the violation they are targeted at, but they also have power (above their significance level) to detect other violations.

2.6.3 Power of Tests for Polytomous Items

The next set of simulations pertains to polytomously scored items with responses following the GPCM by Muraki (1992). The model is the special case of (2.4) with the restriction $\alpha_{ij} = j\alpha_i$. The setup of the studies was analogous to the previous power studies: two studies were conducted to the power in case of violation of constant ability parameters (one where the ability parameter was shifted, and one with random guessing behavior) and one study was conducted to violation of local independence. Ability parameters were drawn from a standard normal distribution, and the parameters α_i were drawn from a log-normal distribution with a mean equal to zero and a standard deviation of 0.25. Drawing the item parameters β_{ij} $(j = 1, ..., m_i)$ was not considered, because the dependence between these parameters may result in very unfavorable values with the consequence of item categories without responses (Wilson & Masters, 1993). The fixed values are difficult to interpret without transforming them to so-called category-bounds parameters defined in the framework of the partial credit model by Masters (1982). The parameters are defined by the transformation $\beta_{i1} = \eta_{i1}$ and $\beta_{ij} = \sum_{h=1}^{j} \eta_{ih}\,(j = 2, ..., m_i)$. The inverse transformation is $\eta_{i1} = \beta_{i1}$ and $\eta_{ij} = \beta_{ij} - \beta_{i(j-1)}\,(j = 2, ..., m_i)$. The

parameters η_{ij} are the points on the latent scale where the odds of scoring in category j rather than in category $j-1$ are one. The values of η_{ij} chosen for the first 5 items are given in Table 2.6. Note that the parameters of item 3 is located in such a way that the category-bounds are located symmetric with respect to the standard normal ability distribution. The first two items are shifted to the left on the latent scale, the last two items are shifted to the right.

Since taking into account the effects of the estimation of the item parameters in the previous study had little impact, the LM_{B2}- and LM_{D2}- tests were not considered in the present study. Thus the study focused on the UB-, LM_{B1}-, UD-, and LM_{D1}-tests. Again, 100 replications were made in every branch of the study. The sample size was always equal to 1000.

Table 2.6
Item parameter values
used for simulating the data
using the GPCM

Item	Category			
	1	2	3	4
1	-2.0	-1.5	-0.5	0.0
2	-1.5	-1.0	0.0	0.5
3	-1.0	-0.5	0.5	1.0
4	-0.5	0.0	1.0	1.5
5	0.0	0.5	1.5	2.0

For the simulations with respect to violation of constant ability parameters, test lengths were equal to 10, 20, 30 and 40 items. Therefore, the item parameter values of Table 2.6 were repeated 2, 4, 6, and 8 times. The model violation was imposed on the last quarter or halve of the test. The results for a shift of the ability parameter are shown in Table 2.7. Note that the shifts were either equal to -0.5 or -1.0. The columns labeled False Alarms pertain to the proportion of incorrectly flagged respondents in the sample of 900 non-aberrant simulees, the columns labeled Hits refer to the proportion of correctly flagged simulees in the sample of the 100 aberrant simulees. Note that the power of the UB-test is substantially lower than the power of the LM_{B1}-test. Main effects of test length and effect size are as expected. The false alarm rate of LM_{B1} was quite close to the nominal

significance level. The power of LM_{D1} was low, and the power of UD was extremely low.

Table 2.8 gives the results of a simulation study to random response behavior. In this simulation, the simulees gave a random response to the second half of the test, by choosing one of the five response categories at random. The power of the tests was now far less than the power in the study to guessing to dichotomous items reported in the previous section. The reason may be that for polytomous items the uniform distribution of item responses over the categories was too little removed from the non-aberrant response distribution, whereas in the dichotomous case, a drop in the probability of a correct response to 0.20 (and an increase in the probability of an incorrect response to 0.8) was a far more extreme deviation from a non-aberrant response, and therefore, the latter could be much easier detected. The last simulation studies relate to detection of violation of local independence. The design consisted of crossing test lengths of 10, 20, and 30, effect sizes $\delta = 0.2$ and $\delta = 0.4$, and 50% and 100% infected items. The results are displayed in Table 2.9. Note that the power is now much larger than in the previous studies. The reason is that in the polytomous case of the GPCM the number of possible outcomes of the response variables is much greater than in the dichotomous case of the Rasch model. As a result, in the polytomous case, the response variables are more reliable indicators of the latent trait and the model violation has more possibility to propagate itself through the consecutive responses. In the present case, the power of LM_{B1} and the power of LM_{D1} were approximately equal. Again the main effects of the test length, the number of items affected, and the effect size δ appeared as expected. The proportion of false alarms was close to the nominal significance probability.

2.7 Discussion

Two Lagrange multiplier test statistics for assessing person fit were introduced, where the effects of estimation of the item and person parameters are explicitly taken into account. Simulation studies showed that accounting for estimation of item parameters had little effect. However, the simulation studies also showed that taking the effects of estimating of the person parameter has a substantial effect on the precision of the Type I error rate and the power. The simulation studies showed that in many instances the

Table 2.7
Detection of changes in ability for polytomous items

K	θ_1	Items Infected	UB False Alarms	UB Hits	LM_{B1} False Alarms	LM_{B1} Hits	UD False Alarms	UD Hits	LM_{D1} False Alarms	LM_{D1} Hits
10	-0.5	5	.00	.08	.04	.28	.00	.02	.03	.13
	-1.0		.00	.08	.04	.31	.00	.03	.03	.13
20	-0.5	10	.00	.18	.04	.40	.00	.04	.04	.18
	-1.0		.00	.23	.03	.47	.00	.06	.04	.19
30	-0.5	15	.00	.04	.04	.20	.00	.00	.04	.04
	-1.0		.00	.22	.05	.54	.00	.01	.05	.09
40	-0.5	20	.01	.05	.05	.24	.00	.00	.05	.05
	-1.0		.01	.36	.05	.67	.00	.02	.05	.11

Table 2.8
Detection of random responses for polytomous items

		UB		LM_{B1}		UD		LM_{D1}	
K	Items Infected	False Alarms	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms	Hits
10	5	.00	.03	.04	.12	.00	.02	.03	.16
20	10	.00	.03	.04	.15	.00	.02	.03	.17
30	15	.00	.05	.04	.15	.00	.02	.03	.22
40	20	.00	.05	.03	.15	.00	.05	.03	.19

Table 2.9
Detection of violation of local independence for polytomous items

			UB		LM_{B1}		UD		LM_{D1}	
K	δ	Items Infected	False Alarms	Hits	False Alarms	Hits	False Alarms	Hits	False Alarms	Hits
10	.2	5	.00	.01	.05	.08	.00	.00	.04	.13
		10	.00	.20	.05	.55	.00	.08	.04	.48
	.4	5	.00	.02	.05	.19	.00	.00	.04	.41
		10	.00	.66	.05	.87	.00	.62	.04	.85
20	.2	10	.00	.23	.05	.64	.00	.05	.05	.71
		20	.00	.82	.05	.92	.00	.75	.05	.92
	.4	10	.00	.60	.05	.74	.00	.54	.05	.64
		20	.00	.70	.05	.75	.00	.65	.05	.65
30	.2	15	.00	.68	.05	.88	.00	.49	.05	.78
		30	.00	.94	.05	1.00	.00	1.00	.05	1.00
	.4	15	.00	.75	.05	1.00	.00	.55	.05	1.00
		30	.01	1.00	.06	1.00	.00	1.00	.04	1.00

power of person fit tests is quite low. This has already been noted by many others (see the references of research on person fit listed above). The highest power was achieved in a simulation with polytomous items were lack of local independence could propagate itself throughout the test.

The approach presented here was applied to a specific IRT model, that is, the generalized partial credit model, with the Rasch model as a special case. However, the approach can be easily generalized to a broader class of parametric IRT models. Parametric IRT models for polytomous items are usually categorized into three classes (Mellenbergh, 1995). Models in the first class are labeled adjacent-category models, and this class encompasses the models considered above (Bock, 1972; Masters, 1982; Muraki, 1992). Models in the second class are labeled continuation-ratio models (Tutz, 1990; Verhelst, Glas, & de Vries, 1997) and models in the third class are called cumulative probability models (Samejima, 1969, 1972, 1973). It is beyond the scope of the present article to outline the differences in the models; their common denominator is that they are all describe the probability of scoring in a certain category of an items by a function of a unidimensional person parameter, say θ. For all these models, person fit statistics for detection of shifts in ability or violation of local independence can be based on generalized models where either the ability parameter is shifted in certain groups by a quantity θ_g or by a quantity δ that is weighted by the response on previous items. The derivation of the test statistic is straightforward: derive first and second order derivatives of the log-likelihood function with respect to θ, θ_g and/or δ and substitute these expressions in (2.1) at the proper places. Further, above the models were presented in a logistic formulation, but the models in all three classes can also be formulated in a normal ogive framework (Lawley, 1943, 1944; Lord, 1952, 1953a and 1953b). Also here application of the presented framework for evaluation of model fit is straightforward.

3
Lagrange Multiplier Person Fit Tests for Polytomous IRT models

Item response theory (IRT) models for polytomous items are often used to analyze and score data from measurement instruments where the item scores are ordered categorical ratings. One may think of psychological rating scale instruments and educational tests where the ratings are made by judges. The fit of the individual's item score pattern to the IRT model are investigated using tests of person fit. Meijer and Sijtsma (1995, 2001) give an overview of person fit statistics proposed for various IRT models. Most person fit statistics were developed for IRT models for dichotomous items (Levine & Rubin, 1979; Wright & Stone, 1979; Tatsuoka, 1984; Smith, 1985, 1986; Drasgow, Levine, & McLaughlin, 1991). Person fit tests for IRT models for polytomous items were developed by Drasgow, Levine, & Williams, 1985; Wright & Masters, 1982; van Krimpen-Stoop & Meijer, 2002). One of the problems of person fit statistics is that the estimation of the item and person parameters complicates the derivation of the distribution of the statistics under the null hypothesis. In Chapter 2, two Lagrange multiplier test statistics for the generalized partial credit model (the GPCM, Muraki, 1992, also see Masters, 1982) were introduced where these effects are explicitly taken into account. Simulation studies showed that accounting for estimation of item parameters had little effect. However, the simulation studies also showed that taking the effects of estimating of the person pa-

rameter into account has a substantial effect on the precision of the Type I error rate and the magnitude of the power.

The purpose of this article is to generalize this approach to two alternatives for the GPCM: the sequential model (SM, Tutz, 1990) and the graded response model (GRM, Samejima, 1969). A general formulation for the person fit tests for the three models will also be given that includes between-item multidimensionality. An additional purpose of this article has to do with reports that, though the rationales underlying the models are very different, the models are hard to distinguish because their item-category response curves are very close (Verhelst, Glas, & de Vries, 1997). In this article, it will be investigated whether the exchangeability of the three models in practical situations also extends to person fit tests, or, put in another way, whether the three models can be distinguished using person fit tests.

Finally, the performance of the person fit tests will be evaluated using data set from NEO Personality Inventory-Revised test.

3.1 Models for polytomous items

Consider items $i = 1, ..., k$, with categories $j = 0, ..., m_i$. We will drop the index i of m_i for convenience. A response pattern is coded as $\mathbf{x} = (\mathbf{x}_1, ..., \mathbf{x}_i, ..., \mathbf{x}_k)$, a response on an item i as $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{ij}, ..., x_m)$, and $x_{ij} = 1$ if a response was given in category j, and zero otherwise. We will use an abbreviation for the logistic function given by

$$\Psi(x) = \frac{\exp(x)}{1 + \exp(x)} .$$

3.1.1 The Graded Response Model

In the graded response model (Samejima, 1969) the probability of a response in category j of item i, $P(X_{ij} = 1)$, is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Psi(\alpha_i(\theta - \beta_{i1})) & \text{if } j = 0 \\ \Psi\left(\alpha_i\left(\theta - \beta_{ij}\right)\right) - \Psi\left(\alpha_i\left(\theta - \beta_{i(j+1)}\right)\right) & \text{if } 0 < j < m \\ \Psi(\alpha_i(\theta - \beta_{im})) & \text{if } j = m . \end{cases} \tag{3.1}$$

To assure that the probabilities $P_{ij}(\theta)$ are positive, the restriction $\beta_{i(j+1)} > \beta_{ij}$, for $0 < j < m$ is imposed.

3.1.2 The sequential model

In the sequential model (Tutz, 1990) the probability of a response in category j of item i is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Psi\left(\alpha_i\left(\theta - \beta_{i1}\right)\right) & \text{if } j = 0 \\ \prod_{h=1}^{j} \Psi\left(\alpha_i\left(\theta - \beta_{ih}\right)\right)\left[1 - \left(\Psi(\alpha_i(\theta - \beta_{i(j+1)}))\right)\right] & \text{if } 0 < j < m \\ \prod_{h=1}^{m} \Psi\left(\alpha_i\left(\theta - \beta_{ih}\right)\right) & \text{if } j = m\ . \end{cases} \tag{3.2}$$

Verhelst, Glas and de Vries (1997) note that every item in the SM can be viewed as a sequence of virtual dichotomous items. These dichotomous items are considered to be presented as long as a correct response is given, and the presentation stops when an incorrect response is given. An important consequence of this conceptualization of the response process is that estimation and testing procedures for the 2PL model with incomplete data can be directly applied to the SM.

3.1.3 The generalized partial credit model

In the generalized partial credit model (Muraki, 1992) the probability of a response in category j of item i is given by

$$P_{ij}(\theta) = \frac{\exp[j\alpha_i\theta - \beta_{ij}]}{1 + \sum_{h=1}^{m_i} \exp[h\alpha_i\theta - \beta_{ih}]}. \tag{3.3}$$

The partial credit model (Masters, 1982) is the special case where $\alpha_i = 1$ for all items i and the item parameters are usually re-parameterized as $\beta_{ij} = \sum_{h=1}^{j} \eta_{ih}$.

3.2 The LM Test

The LM test is grounded on the following rationale. Consider some general parameterized model, and a special case of the general model, the so-called restricted model. The restricted model is derived from the general model by imposing constraints on the parameter space. In many instances, this

is accomplished by fixing one or more parameters of the general model to constants. The LM test is based on the evaluation of the first-order partial derivatives of the log-likelihood function of the general model, evaluated using the maximum likelihood estimates of the restricted model. The unrestricted elements of the vector of first-order derivatives are equal to zero, because their values originate from solving the likelihood equations. The magnitudes of the elements of the vector of first-order partial derivatives corresponding to restricted parameters determine the value of the statistic: the closer they are to zero, the better the model fit.

More formally, the principle can be described as follows. Consider a general model with parameters $\boldsymbol{\eta}$. In the applications to be presented below, the special model is derived from the general model by fixing one or more parameters to zero. So if the vector of the parameters of the general model, say $\boldsymbol{\eta}$, is partitioned $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, the null hypothesis entails $\boldsymbol{\eta}_2 = 0$. Let $\mathbf{h}(\boldsymbol{\eta})$ be the first order partial derivatives of the log-likelihood of the general model. This vector of partial derivatives gauges the change of the log-likelihood as a function of local changes in $\boldsymbol{\eta}$. Let the vector of partial derivatives $\mathbf{h}(\boldsymbol{\eta})$ be partitioned as $(\mathbf{h}(\boldsymbol{\eta}_1), \mathbf{h}(\boldsymbol{\eta}_2))$. If the likelihood equations are solved under the null model, $\mathbf{h}(\boldsymbol{\eta}_1) = 0$. Then the test is based on the statistic

$$\text{LM} = \mathbf{h}(\boldsymbol{\eta}_2)' \boldsymbol{\Sigma}^{-1} \mathbf{h}(\boldsymbol{\eta}_2), \tag{3.4}$$

where

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} \tag{3.5}$$

and $\boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{22}$, and $\boldsymbol{\Sigma}_{12}$ are the matrices of second order derivative with respect to $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$, and $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$, respectively. The LM statistic is evaluated using the maximum likelihood estimates of the parameters of the special model of the null hypothesis. The LM statistic has an asymptotic χ^2-distribution with degrees of freedom equal to the number of parameters in $\boldsymbol{\eta}_2$ (Rao (1947, Aitchison & Silvey, 1958).

The variance of the estimates of θ plays the following role in the distribution of the LM statistics. Glas (1999) shows that the matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_{22}$ in (3.5) can be viewed as the asymptotic covariance matrices of $\mathbf{h}(\boldsymbol{\eta}_2)$ with $\boldsymbol{\eta}_1$ (in the present case, the ability parameter) estimated and known, respectively. Further, $\boldsymbol{\Sigma}_{11}^{-1}$ is the asymptotic covariance matrix of the estimate of $\boldsymbol{\eta}_1$, so the term $\boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$ accounts for the influence of the estimation of $\boldsymbol{\eta}_1$ on the covariance matrix of $\mathbf{h}(\boldsymbol{\eta}_2)$. So in the LM test, the

variance of the estimates of the person parameters is explicitly taken into account.

3.2.1 An alternative model for the unidimensional case

Two model violations will be targeted: differences of the person parameter between test items and violation of local independence. To assess the first violation let the complete response pattern be split up into a number of parts, say the parts $g = 0, ..., G$, and let A_g be the set of the indices of the items in part g. It can be evaluated whether the same ability parameter θ can account for the responses in all partial response patterns by comparing the null model against the alterative model that this is not the case, that is, for $g > 0$, it holds that $P(X_{ij} = 1) = P_{ij}(\theta + \delta_g)$. In the sequel, it will be assumed that $G = 1$ because, in general, tests are too short to consider more than two subtests. Very short subtests would result in too low power for the fit statistics.

A general alternative model that violates local independence is given by $P(X_{ij} = 1|Y_i = y_i) = P_{ij}(\theta + y_i\delta)$. The variable Y_i can be the score on some other item or set of items in the test (without item i).

In the situation where Y_i is defined as a dichotomous variable that partitions the item into two subtests, the formal definition of the alternative model for violation of constancy of θ and local independence become similar. For constancy of θ, Y_i is defined as an indicator function assuming a value one if the item is in the second subtest, and zero otherwise. For local independence, Y_i is some dichotomous indicator depending on the score of one or more other items.

If the null hypothesis $\delta = 0$ is tested, the matrices $\mathbf{\Sigma}_{11}$, $\mathbf{\Sigma}_{22}$, and $\mathbf{\Sigma}_{12}$ in formulas (3.4) and (3.5) become scalars and the LM statistic specializes to

$$LM = \frac{h_1^2}{\boldsymbol{\sigma}_{22} - \boldsymbol{\sigma}_{12}^2/\boldsymbol{\sigma}_{11}}, \tag{3.6}$$

where h_1 is the first order derivative of the log-likelihood of the general alternative model. In Appendix A it is shown that for the GPCM and SM, h_1 turns out to be a difference between observed and expected values. So in these two cases h_1 can be viewed as a residual. Further $\boldsymbol{\sigma}_{11}$, $\boldsymbol{\sigma}_{22}$, and $\boldsymbol{\sigma}_{12}$ are the second order derivatives with respect to θ, δ, and θ and δ, respectively. The statistic has one degree of freedom.

Note that $\boldsymbol{\sigma}_{12}^2/\boldsymbol{\sigma}_{11}$ takes into account the loss of variation due to the estimation of θ. In the simulation studies reported below, we shall also consider the "naive" version of the test statistic where the term $\boldsymbol{\sigma}_{12}^2/\boldsymbol{\sigma}_{11}$ is deleted. Exact expression for h_1, $\boldsymbol{\sigma}_{11}$, $\boldsymbol{\sigma}_{22}$, and $\boldsymbol{\sigma}_{12}$ are given in appendix A.

3.2.2 An alternative model of between-items multidimensionality

In many situations, the assumption that an individual's response behavior can be explained by a unidimensional person parameter θ does not hold. In that case the assumption of unidimensionality can be replaced by the assumption of a multidimensional person parameter $\theta_1, ..., \theta_q, ..., \theta_Q$. The multidimensional versions of the models given by (3.1), (3.2) and (3.3) are defined by replacing the term $\alpha_i\theta$ by $\sum_{q=1}^{Q} \alpha_{iq}\theta_q$. Further, it is assumed that $\theta_1, ..., \theta_q, ..., \theta_Q$ have a Q-variate normal distribution (McDonald, 1997; Reckase, 1997).

In the present article we will only consider the special case of between-items multidimensionality. This entails the assumption that the test can be split up in a number of subtests and every subtest relates to a specific person parameter θ_t. So this is the special case where all α_{iq} $(q = 1, ..., Q; Q = T)$ are zero except for the proficiency parameter θ_t to which the item figure relates.

Assuming local independence of responses given θ, the likelihood of a response pattern $\mathbf{x}$ for multidimensional data is given by

$$p_{\boldsymbol{\theta}}(\mathbf{x}) = \prod_{t=1}^{T} \prod_{i|t} \prod_{j=0}^{m_i} P_{ij}(\theta_t)^{x_{ij}} g\left(\theta_1, \theta_2, ..., \theta_T | \boldsymbol{\Sigma}_\theta\right). \tag{3.7}$$

where $g\left(\boldsymbol{\theta}|\boldsymbol{\Sigma}_\theta\right)$is a multivariate normal density with covariance matrix $\boldsymbol{\Sigma}_\theta$ and mean equal to zero. The mean is set equal to zero to identify the latent scale.

A general alternative model is given by $P(X_{ij} = 1|Y_i = y_i) = P_{ij}(\theta_t + y_i\delta)$. For testing the null hypothesis that $\delta = 0$ the statistic given by (3.4) specializes to a case where $\boldsymbol{\Sigma}_{12}$ is now a T-dimensional vector of the second order derivatives with respect to δ and θ_t $(t = 1, ..., T)$ and $\boldsymbol{\Sigma}_{11}$ is a TxT matrix of the second order derivative with respect to θ_t $(t = 1, ..., T)$. In Appendix A it is shown that also for the multidimensional versions of the

GPCM and SM, h_1 is a difference between observed and expected values. So also here h_1 can be viewed as a residual.

3.3 Simulation Study 1

In the first simulation study, only the unidimensional test statistics will be considered.

3.3.1 Type I error rate

First, the Type I error rate of LM tests for the GPCM, GRM and SM was studied using a simulation study. To assess the impact of ignoring the effects of estimation of θ, the "naive" test statistic was also included in the study.

The set up of the simulation study was as follows. For all three models, ability parameters were drawn from a standard normal distribution and the α-parameters were fixed to one. For the GPCM, drawing the item parameters β_{ij} $(j = 1, ..., m_i)$ was not considered, because the dependence between these parameters may result in very unfavorable values with the consequence of item categories without responses (Wilson & Masters, 1993).Therefore, the β-parameter were fixed. In the GPCM, the fixed values are difficult to interpret without transforming them to so-called category-bounds parameters defined in the framework of the partial credit model by Masters (1982). The parameters are defined by the transformation $\beta_{i1} = \eta_{i1}$ and $\beta_{ij} = \sum_{h=1}^{j} \eta_{ih}$ $(j = 2, ..., m_i)$. The inverse transformation is $\eta_{i1} = \beta_{i1}$ and $\eta_{ij} = \beta_{ij} - \beta_{i(j-1)}$ $(j = 2, ..., m_i)$. The parameters η_{ij} are the points on the latent scale where the odds of scoring in category j rather than in category $j - 1$ are one. The values of η_{ij} chosen for the first 5 items are given in Table 3.1.

Table 3.1
Item parameter values
used for simulating the data
using the GPCM

Item	Category			
	1	2	3	4
1	-2.0	-1.5	-0.5	0.0
2	-1.5	-1.0	0.0	0.5
3	-1.0	-0.5	0.5	1.0
4	-0.5	0.0	1.0	1.5
5	0.0	0.5	1.5	2.0

Note that the parameters of item 3 is located in such a way that the category-bounds are located symmetric with respect to the standard normal ability distribution. The first two items are shifted to the left on the latent scale, the last two items are shifted to the right.

The item parameter for the SM and GRM were chosen in such a way that the item-category response curves were close. To achieve this, data were generated under the GPCM, and using this data, the item parameters of the SM and GRM were estimated using maximum marginal likelihood (Bock & Aitkin, 1981). These estimated values were then used as generating values for the data following SM and GRM.

100 replications were made in every branch of the study and the sample size always equal to 1000.

Table 3.2 gives the Type I error rate of "naive" and LM tests. The first column labeled 'Generating Model' gives the model used for generating the data. The estimation model is given in the second column labelled 'Estimation Model'. The third column labeled 'K' gives the number of items and UB, UD, LM_B and LM_D give the proportions of rejections at a 5% significance level for "naive" tests for constancy of theta and violation of local independence and LM tests for constancy of theta and violation of local independence, respectively.

The results show that the empirical type I error rate of naive tests were much smaller than the nominal ones but was greatly improved for the LM tests. Note that the "wrong model" still gives good Type I error rate. In general, we have an acceptable Type I error rate.

Table 3.2
Type 1 error rates

Generating Model	Estimation Model	K	UB	UD	LM_B	LM_D
GPCM	GPCM	20	.005	.002	.051	.046
		40	.006	.005	.054	.051
	SM	20	.007	.005	.054	.050
		40	.009	.007	.055	.050
	GRM	20	.006	.003	.048	.048
		40	.007	.007	.050	.052
SM	GPCM	20	.007	.002	.056	.048
		40	.008	.005	.057	.051
	SM	20	.004	.003	.032	.031
		40	.008	.006	.037	.038
	GRM	20	.006	.003	.045	.042
		40	.009	.004	.050	.045
GRM	GPCM	20	.004	.0005	.042	.032
		40	.008	.001	.048	.036
	SM	20	.003	.001	.030	.026
		40	.005	.003	.039	.030
	GRM	20	.004	.001	.032	.031
		40	.009	.002	.040	.034

3.3.2 Power of the test

Second, the power of the test was studied and this was done by creating a shift on the ability parameter to the left. The MML estimation procedure was run using the data of all simulees, both the aberrant and non-aberrant ones. In all simulations, 10% of the simulees were aberrant. The presence of the aberrant simulees did, of course, produce some bias in the parameter estimates, but this setup was considered realistic because in many situations it is not a priori known which respondents are aberrant, and which are not. Item parameters were equal to the item parameters in the previous study, unless indicated otherwise, and the ability parameters θ were again drawn from a standard normal distribution. In all simulations, test length was equal to $K = 20$ and $K = 40$ and the samples size is always equal to 1000. Therefore, the item parameter values in Table 3.1 were repeated 4 and 8 times. The number of replications in each branch of the study was equal to 100. The results for a shift of the ability parameter are shown in Tables 3.3 and 3.4. Note that the shifts were either equal to -0.5 or -1.0. The columns labeled 'False Alarms' pertain to the proportion of incorrectly flagged respondents in the sample of 900 non-aberrant simulees, the columns labeled 'Hits' refer to the proportion of correctly flagged simulees in the sample of the 100 aberrant simulees. LM_B and LM_D stands for LM tests for constancy of theta and violation of local independence, respec-

Table 3.3
Detection of changes in ability
20 items

Generating Model	Estimation Model	δ	LM_B		LM_D	
			False Alarms	Hits	False Alarms	Hits
GPCM	GPCM	-0.5	.051	.131	.045	.049
		-1.0	.054	.358	.045	.082
	SM	-0.5	.054	.136	.049	.053
		-1.0	.056	.373	.049	.086
	GRM	-0.5	.049	.123	.046	.050
		-1.0	.051	.339	.048	.078
SM	GPCM	-0.5	.045	.145	.034	.040
		-1.0	.047	.384	.033	.084
	SM	-0.5	.029	.148	.026	.034
		-1.0	.031	.420	.025	.083
	GRM	-0.5	.036	.140	.032	.039
		-1.0	.037	.376	.031	.083
GRM	GPCM	-0.5	.042	.151	.031	.040
		-1.0	.045	.393	.030	.090
	SM	-0.5	.030	.151	.025	.037
		-1.0	.031	.426	.025	.089
	GRM	-0.5	.033	.146	.030	.039
		-1.0	.035	.386	.029	.085

Table 3.4
Detection of changes in ability
40 items

Generating Model	Estimation Model	δ	LM_B		LM_D	
			False Alarms	Hits	False Alarms	Hits
GPCM	GPCM	-0.5	.052	.235	.049	.057
		-1.0	.057	.641	.049	.126
	SM	-0.5	.056	.240	.054	.062
		-1.0	.060	.645	.053	.131
	GRM	-0.5	.051	.225	.051	.055
		-1.0	.055	.616	.050	.122
SM	GPCM	-0.5	.055	.289	.040	.055
		-1.0	.056	.647	.040	.188
	SM	-0.5	.034	.312	.030	.051
		-1.0	.036	.677	.030	.202
	GRM	-0.5	.043	.283	.036	.056
		-1.0	.045	.639	.038	.192
GRM	GPCM	-0.5	.050	.294	.037	.059
		-1.0	.057	.662	.036	.183
	SM	-0.5	.035	.315	.031	.056
		-1.0	.039	.678	.029	.207
	GRM	-0.5	.040	.294	.036	.060
		-1.0	.043	.638	.032	.194

Table 3.5
Agreement between models
20 items

Generating Model	Estimation Model	δ	GPCM			
			$LM_B(\%)$		$LM_D(\%)$	
			Normal	Aberrant	Normal	Aberrant
GPCM	SM	-0.5	98.5	81.4	98.5	79.3
		-1.0	98.3	84.7	98.5	79.8
	GRM	-0.5	99.1	80.8	98.9	81.0
		-1.0	99.0	83.2	98.9	82.4
SM	SM	-0.5	99.4	62.5	99.5	62.9
		-1.0	99.1	74.6	99.4	66.4
	GRM	-0.5	99.3	71.4	99.3	75.3
		-1.0	99.3	79.1	99.3	77.0
GRM	SM	-0.5	99.3	66.7	99.4	65.6
		-1.0	99.1	76.2	99.4	69.1
	GRM	-0.5	99.4	72.6	99.3	75.9
		-1.0	99.3	78.5	99.3	77.4

tively. The false alarm rate of our LM tests were relatively close to the nominal significance level though the power of the LM test for local independence was much lower than the power to detect changes in ability. Also, it would seem like the power of the test for GRM is lower than the power of the test for GPCM with the power of the test for SM performing better than the other two models. In general, the LM test have reasonable power to detect model violations of constancy of theta and local independence.

3.3.3 Agreement between models

To investigate what extent the three models give comparable results the degree of agreement to detect normal and aberrant responses between the three models was determined. Tables 3.5 and 3.6 gives the results of the degree of agreement. The proportion of the degree of agreement between models for LM tests for constancy of theta and violation of local independence are given in columns 4, 5, 6, and 7.

The degree of agreement for normal simulees was higher than aberrant simulees and was not greatly improved as we increase the number of items from 20 to 40. Note that the highest degree of agreement to detect aberrant simulees occurs with GPCM as the generating model and lowest if the generating model is the SM. In general, a comparable result to detect normal simulees is more evident than to detect aberrant simulees between the three models.

Table 3.6
Agreement between models
40 items

Generating Model	Estimation Model	δ	GPCM			
			$LM_B(\%)$		$LM_D(\%)$	
			Normal	Aberrant	Normal	Aberrant
GPCM	SM	-0.5	98.3	82.2	98.4	79.5
		-1.0	98.2	88.1	98.4	80.8
	GRM	-0.5	98.9	81.9	98.7	79.8
		-1.0	98.9	87.0	98.9	83.2
SM	SM	-0.5	99.0	65.1	99.4	62.0
		-1.0	99.1	78.2	99.0	68.9
	GRM	-0.5	99.1	72.8	99.2	74.1
		-1.0	99.1	82.2	98.9	77.4
GRM	SM	-0.5	98.9	69.3	99.3	67.6
		-1.0	98.9	78.3	99.1	74.3
	GRM	-0.5	99.0	74.1	99.1	76.4
		-1.0	99.2	80.1	99.2	78.0

In conclusion, the LM person fit tests developed under this framework have an acceptable type 1 error rate, reasonable power with respect to the model violations targeted at though SM seems to be uniformly larger or better than the other two models and seems to give the result that the three models can be distinguished using person fit tests.

3.4 Simulation Study 2

The multidimensional version of our test statistics will be considered in this simulation study which will also be used in the real data example.

If a scale consists of more than one subscale, a person fit statistic pertaining to one of the subscales can be computed in two ways: using the estimate of θ_t obtained on the focussed subscale alone, or using an estimate of all ability parameters pertaining to all subscales. The purpose of this simulation study is to asses the effect of using such auxiliary information from other subscales on the power of the fit statistic as a function of the correlation between the subscales.

We consider three subscales, $t = 1, ..., 3$, associated with three ability parameters $\theta_1, \theta_2, \theta_3$. Every subscale has 16 items, so there are 48 items in total. We will test the null-hypothesis that the last 8 items of the first subscale relate to the same ability parameter θ_1 as the first 8 items. The item parameters are the same as in the previous study. In particular, for all items, the number of categories equals five, that is, $m = 4$.

The variances of the ability parameters are all equal to one.
The design of the simulation studies is crossed with three facets:

1. Three values are chosen for the correlations between the ability dimensions: $\rho_{\theta\theta} = 0.4$, $\rho_{\theta\theta} = 0.6$, and $\rho_{\theta\theta} = 0.8$.

2. The effect size of the model violation: $\delta = -0.5$ and $\delta = -1.0$.

3. Using an estimate of θ_1 alone or using an estimate of all person parameters $\theta_1, \theta_2, \theta_3$.

Every data set consists of $N = 1,000$ simulees, and 100 replications are made in every condition.

The results are given in Table 3.7.

First, the false alarm rate between the two estimation procedures used were relatively close to the nominal significance level for the correct estimation model but not for the wrong estimation model. GPCM always has a higher hit rate even if it is not the starting or generating model. As expected, there is no main effect of correlation when we use the estimate of θ_1 alone but there is a slight though not substantial enough correlation effect when we use the estimate of all person parameters $\theta_1, \theta_2, \theta_3$. Also as expected, there is a main effect of the effect size on the power of the test.

3.5 An Empirical Example

Data from the NEO Personality Inventory data were used to get an impression of the degree of agreement between the tree IRT models. This personality test is designed to provide a general description of normal personality that is relevant to clinical, counselling and educational situations. It is based on the Five-Factor model of personality (Costa & McCrae, 1992). NEO consists of five broad domains and for each of these domains, six facet scores or subfactors have been developed in to provide the specific level of information necessary for clinicians. Each of the six facets is measured by eight items. All items are rated on a 5-point scale. Three validity items are also included.

The empirical example presented here pertians to the the neuroticism domain. To obtain subscales of reasonable length, pairs of two factes within the domain were grouped together on the bases of their correlation. That

Table 3.7a
Power of LM_B as a function of the correlation between the subscales

Generating Model	Estimation Model	δ	$\rho_{\theta\theta}$	Estimation of θ_1		Estimation of $\theta_1, \theta_2, \theta_3$	
				False Alarms	Hits	False Alarms	Hits
	GPCM	-0.5	.40	.047	.138	.047	.142
			.60	.049	.144	.049	.146
			.80	.050	.138	.049	.150
		-1.0	.40	.047	.415	.047	.425
			.60	.049	.418	.049	.433
			.80	.050	.414	.049	.450
GPCM	SM	-0.5	.40	.068	.070	.071	.071
			.60	.064	.072	.065	.072
			.80	.071	.068	.052	.078
		-1.0	.40	.068	.209	.071	.213
			.60	.064	.212	.065	.229
			.80	.071	.212	.052	.259
	GRM	-0.5	.40	.040	.047	.046	.037
			.60	.040	.045	.047	.036
			.80	.043	.043	.047	.035
		-1.0	.40	.040	.112	.046	.087
			.60	.039	.113	.047	.086
			.80	.043	.112	.047	.105
	GPCM	-0.5	.40	.130	.228	.132	.232
			.60	.134	.222	.130	.222
			.80	.138	.224	.132	.224
		-1.0	.40	.130	.397	.132	.402
			.60	.134	.396	.130	.405
			.80	.138	.409	.129	.421
SM	SM	-0.5	.40	.051	.098	.052	.100
			.60	.053	.098	.050	.101
			.80	.055	.097	.052	.103
		-1.0	.40	.051	.230	.052	..235
			.60	.053	.226	.049	.238
			.80	.055	.230	.052	.260
	GRM	-0.5	.40	.025	.060	.024	.049
			.60	.027	.059	.026	.046
			.80	.025	.058	.026	.045
		-1.0	.40	.025	.151	.024	.123
			.60	.027	.145	.026	.115
			.80	.025	.148	.026	.127

Table 3.7b
Power of LM_B as a function of the correlation between the subscales

Generating Model	Estimation Model	δ	$\rho_{\theta\theta}$	Estimation of θ_1		Estimation of $\theta_1, \theta_2, \theta_3$	
				False Alarms	Hits	False Alarms	Hits
	GPCM	-0.5	.40	.105	.164	.116	.189
			.60	.105	.171	.112	.192
			.80	.108	.165	.104	.186
		-1.0	.40	.105	.291	.116	.326
			.60	.105	.277	.112	.316
			.80	.108	.277	.104	.325
GRM	SM	-0.5	.40	.072	.075	.066	.083
			.60	.077	.083	.068	.098
			.80	.075	.077	.064	.101
		-1.0	.40	.072	.140	.066	.169
			.60	.077	.136	.068	.179
			.80	.075	.139	.064	.203
	GRM	-0.5	.40	.043	.053	.043	.054
			.60	.044	.059	.044	.062
			.80	.046	.054	.043	.057
		-1.0	.40	.043	.112	.043	.118
			.60	.044	.107	.044	.122
			.80	.046	.114	.043	.135

is, facets with the highest mutual correlation were grouped together. So 3 subscales of 16 items each were analysed. Note that this setup is analogous to the setup used in the second simulation. Further, also in the present analysis, the null-hypothesis tested was that the last 8 items of the first subscale relate to the same ability parameter as the first 8 items. The test was performed in two versions: one version using the parameter estimates of the first subscale only, and one version using the item parameter estimates of all three subscales and the corrlations between the latent veriables of the three subscales as collateral information. MML estimates of the item-parameter for the unidimensional model were computed using Multilog (Thissen, Chen, & Bock, 2003) and the MML estimates of the item- and latent covariance parameters of the multidimensional model were computed using dedicated software by the authors. Table 3.8 gives the manifest correlations between the three subscales and the MML estimated latent correlations. Note that, as expected, the manifest correlations are attenuated. For all three models, Figure 1 gives plots of the β-parameters of the first subscale obtained using the unidimensional and multidimensional version of the models. Note that the parameter estimates are quite close.

Table 3.8 Observed and latent correlation matrices of the subscales

Observed			latent-GPCM		
1.000	0.694	0.596	1.000	0.848	0.758
0.694	1.000	0.585	0.848	1.000	0.778
0.596	0.585	1.000	0.758	0.778	1.000
latent-GRM			latent-SM		
1.000	0.864	0.767	1.000	0.861	0.772
0.864	1.000	0.797	0.861	1.000	0.802
0.767	0.797	1.000	0.772	0.802	1.000

Table 3.9 Cross tabulation tables for the Unidimensional item-parameter estimates

Panel A

		SM +	SM -
GPCM	+	.128	.061
	-	.083	.728
Kappa = 0.55			

Panel B

		GRM +	GRM -
GPCM	+	.074	.038
	-	.096	.792
Kappa = 0.45			

Panel C

		SM +	SM -
GRM	+	.108	.132
	-	.081	.679
Kappa = 0.37			

Table 3.10
Cross tabulation tables for the Multidimensional item-parameter estimates

Panel A

		SM +	SM -
GPCM	+	.109	.061
	-	.089	.741
Kappa = 0.50			

Panel B

		GRM +	GRM -
GPCM	+	.061	.049
	-	.077	.813
Kappa = 0.42			

Panel C

		SM +	SM -
GRM	+	.098	.103
	-	.088	.711
Kappa = 0.39			

Table 3.11
Cross tabulation tables for the degree of agreement between Unidimensional and Multidimensional item-parameter estimates

GPCM

		UNI +	UNI -
MULTI	+	.112	.019
	-	.048	.821
Kappa = 0.73			

GRM

		UNI +	UNI -
MULTI	+	.093	.065
	-	.051	.791
Kappa = 0.55			

SM

		UNI +	UNI -
MULTI	+	.074	.099
	-	.055	.772
Kappa = 0.40			

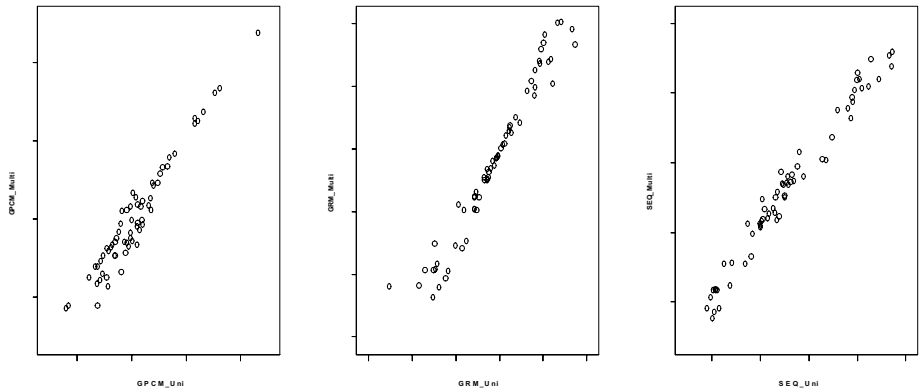

Figure 1 Plots of the β-parameter unidimensional vs multidimensional

Using the MML estimates, the θ-parameters were estimatd by maximum likelihood, and the fitt-statistics were computed. Tables 3.9 and 3.10 give a cross-tabulation of the persons identified as aberrant and non-aberrant under the three models for the unidimensional and multidimensional case, respectively. The “plus ” and “minus ” signs in all the tables

refer to aberrant and normal persons, respectively. Coefficient Kappa was used as a measure of the degree of agreement, the values are given at the bottom of the table. The values of Kappa indicate that the degree of agreement between the models is moderate. The largest agreement to detect normal and aberrant respondents is between GPCM and GRM and GPCM and SM, respectively, for both parameter estimation procedures. However, estimating the item-parameters unidimensionally and multidimensionally has the effect of slighly increasing the degree of agreement.

The degree of agreement between the unidimensional and multidimensional versions of the models is presented in Table 3.11. It can be seen that the GPCM is performing better than the other two models.

3.6 Discussion

Our main goal in this study was comparing the performance of person fit tests based on the Lagrange multiplier tests for three IRT models for polytomous items: the GPCM, the SM and the GRM. The test was developed both for the unidimensional and multidimensional case, and two simulation studies were conducted. We have also conducted a small empirical study.

The first simulation study pertained to the unidimensional version where the power and Type I error rate of the test was investigated. Here we see that in general we have an acceptable Type I error rate even if a "wrong model " (a model not used to generate the data) was used in the analysis. The power of the test to detect model violations of contancy of theta and local independence was reasonable. Further, in our investigation to what extent the three models give comparable results we found that contrary to earlier findings (Verhelst, Glas, & de Vries, 1999) the models are not completely exchangeable.

The second simulation study pertained to the multidimensional version of the test statistics. For simplicity, we only considered the special case of between-items multidimensionality. Here we assessed the effect of using auxiliary information from other subscales on the power of the fit tests as a function of the correlation between the subscales. Results showed that, in general, there was no main effect of correlation. Further, model violation on two levels of effect sizes was introduced. Results showed that there was a main effect of the effect size on the power of the test.

In our empirical example, we used data from the NEO PI-R to get the impression of the degree agreement between the three IRT models in a real situation. Results based on the Kappa coeffecient showed that, the degree of agreement betweeen the three models was mostly moderate and always slightly higher for GPCM and SM. Also, the degree of agreement between the unidimensional and multidimensional versions of the models was highest for GPCM.

In conclusion, the performance of the test is reasonable and there is evidence that the results are not comparable between models.

3.7 Appendix A: Detailed characterization of the test statistics

We will now give a detailed characterization of the LM tests. We will consider the tests for the multidimensional versions of the GPCM, SM and GRM, the unidimensional versions follow directly as a special case. Only the case where the test is split up into two subtests is considered, so $G = 1$. The generalization to $G > 1$ is straightforward. A general formulation for all statistics considered is given by

$$LM = \frac{h_1^2}{\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^{'}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}}. \tag{3.8}$$

The log-likelihood of a response pattern $\mathbf{x}$ for the general alternative model for multidimensional data is given by

$$\log L(\theta, \delta) = \sum_{t=1}^{T} \sum_{i|t} \sum_{j=0}^{m_i} x_{ij} \log P_{ij}(\theta_t + y_i\delta) + \log g\left(\theta_1, ..., \theta_T | \boldsymbol{\Sigma}_\theta\right). \tag{3.9}$$

where the first summation is over subtests t, the second summation is over the items i in subtest t, and the last summation is over the response categories. Further, $g\left(\boldsymbol{\theta}|\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)$ is the multivariate normal density with mean zero and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$, and $P_{ij}(\theta_t + y_i\delta)$ is the probability of a response on item i in category j, given by either the GPCM, the SM, or the GRM. We define

$$d_{tij} = \frac{\partial \log P_{ij}(\theta_t + y_i\delta)}{\partial \theta_t}$$

and

$$D_{tij} = \frac{\partial^2 \log P_{ij}(\theta_t + y_i\delta)}{\partial \theta_t^2}.$$

It directly follows that

$$\frac{\partial \log P_{ij}(\theta_t + y_i\delta)}{\partial \delta} = y_i d_{tij},$$

$$\frac{\partial^2 \log P_{ij}(\theta_t + y_i\delta)}{\partial \theta_t \partial \delta} = y_i D_{tij}$$

and

$$\frac{\partial^2 \log P_{ij}(\theta_t + y_i\delta)}{\partial \delta^2} = y_i^2 D_{tij}.$$

Further,

$$\frac{\partial \log g\left(\boldsymbol{\theta}|\boldsymbol{\Sigma}_\theta\right)}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}.$$

Therefore, the first order derivatives in (3.8) are given by

$$h_1 = \frac{\partial \log L(\boldsymbol{\theta}, \delta)}{\partial \delta} = \sum_{t=1}^{T} \sum_{i|t} \sum_{j=0}^{m_i} y_i x_{ij} d_{tij},$$

and the second order derivatives by

$$\begin{aligned}
\boldsymbol{\Sigma}_{11} &= -\frac{\partial^2 \log L(\boldsymbol{\theta}, \delta)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = -\sum_{t=1}^{T} \sum_{i|t} \sum_{j=0}^{m_i} \frac{x_{ij} \partial^2 \log P_{ij}(\theta_t + y_i\delta)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \frac{\partial^2 \log g\left(\boldsymbol{\theta}|\boldsymbol{\Sigma}_\theta\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \\
&= -Diag\left[\sum_{i|t} \sum_{j=0}^{m_i} x_{ij} D_{tij}\right] + \boldsymbol{\Sigma}_\theta^{-1} \\
\boldsymbol{\Sigma}_{22} &= -\frac{\partial^2 \log L(\boldsymbol{\theta}, \delta)}{\partial \delta^2} = -\sum_{t=1}^{T} \sum_{i|t} \sum_{j=0}^{m_i} x_{ij} y_i D_{tij} \\
\boldsymbol{\Sigma}_{12} &= -\frac{\partial^2 \log L(\boldsymbol{\theta}, \delta)}{\partial \boldsymbol{\theta} \partial \delta} = -Vec\left[\sum_{i|t} \sum_{j=0}^{m_i} x_{ij} y_i D_{tij}\right],
\end{aligned}$$

where $Vec\left[\sum_{i|t} \sum_j x_{ij} y_i D_{tij}\right]$ and $Diag\left[\sum_{i|t} \sum_j x_{ij} D_{tij}\right]$ are a vector and a diagonal matrix of the elements as indexed by t, $t = 1, ..., T$.

Next, we derive expressions for d_{tij} and D_{tij} for the GRM, SM and GPCM, respectively.

3.7.1 First and second order derivatives for the graded response model

We introduce a concise notation

$$P_{ij} = \Psi_{ij} - \Psi_{i(j+1)}$$

with $P_{ij} = P_{ij}(\theta)$, $\Psi_{ij} = \Psi\left(\alpha_i\left(\theta - \beta_{ij}\right)\right)$, $\Psi_{i0} = 1$, and $P_{ij}(\theta) = \Psi_{i(m+1)} = 0$. Further Ψ'_{ij} and P'_{ij} are first order derivative with respect to θ. Then

$$\begin{aligned} d_{tij} &= \frac{\partial \log P_{ij}(\theta + y_i\delta)}{\partial\theta} = \frac{P'_{ij}}{P_{ij}} \\ &= \alpha_i\left[\frac{\Psi_{ij}(1-\Psi_{ij}) - \Psi_{i(j+1)}(1-\Psi_{i(j+1)})}{\Psi_{ij} - \Psi_{i(j+1)}}\right] \\ &= \alpha_i\left[1 - \Psi_{ij} - \Psi_{i(j+1)}\right], \end{aligned}$$

for $j = 0, .., m_i$. Note that for $j = 0$, we have $d_{tij} = -\alpha_i\Psi_{i(j+1)}$ and for $j = m_i$ we have $d_{tij} = -\alpha_i(1 - \Psi_{ij})$.

For the second order derivatives of the log-likelihood we obtain

$$D_{tij} = \frac{\partial^2 \log P_{ij}(\theta + y_i\delta)}{\partial\theta^2} = -\alpha_i^2\left[\Psi_{ij}(1-\Psi_{ij}) + \Psi_{i(j+1)}(1-\Psi_{i(j+1)})\right],$$

for $j = 0, .., m_i$. Note that for $j = 0$, we have $d_{tij} = -\alpha_i^2\Psi_{i(j+1)}(1-\Psi_{i(j+1)})$ and for $j = m_i$ we have $d_{tij} = -\alpha_i^2\Psi_{ij}(1-\Psi_{ij})$.

3.7.2 First and second order derivatives for the sequential model

We introduce a concise notation

$$P_{ij} = \left[\prod_{h=1}^{j}\Psi_{ih}\right]\left[1 - \Psi_{i(j+1)}\right]$$

were the product from $j = 1$ to 0 is assumed to result in unity, and $1 - \Psi_{i(m+1)} = 1$. Then

$$\begin{aligned} d_{tij} &= \frac{\partial \log P_{ij}(\theta + y_i\delta)}{\partial\theta} = \frac{P'_{ij}}{P_{ij}} \\ &= \alpha_i \left[\sum_{h=1}^{j} \frac{P_{ij}(1 - \Psi_{ih}) + P_{ij}(-\Psi_{i(h+1)})}{P_{ij}} \right] \\ &= \alpha_i \left[\sum_{h=1}^{j} (1 - \Psi_{ih}) - \Psi_{i(h+1)} \right] \end{aligned}$$

Note that it follows that the elements of h_1 are a difference between observed and expected values: If z_i is defined as the score obtained on item i, and $z = \sum_i y_i z_i \alpha_i$, then

$$h_1 = z - \sum_i y_i \alpha_i \sum_{h=1}^{\min(z_i+1, m_i)} \Psi_{ih}.$$

The second order derivative is given by

$$D_{tij} = \frac{\partial^2 P_{ij}(\theta + y_i\delta)}{\partial\theta^2} = -\alpha_i^2 \sum_{h=1}^{z_i+1} \Psi_{ih}(1 - \Psi_{ih}).$$

3.7.3 First and second order derivatives for the generalized partial credit model

We introduce a concise notation $P_{ij} = P_{ij}(\theta + y_i\delta)$. Then

$$d_{tij} = \frac{\partial \log P_{ij}(\theta + y_i\delta)}{\partial\theta} = \left(j\alpha_i - \sum_{h=1}^{m_i} h\alpha_i P_{ih} \right).$$

and

$$D_{tij} = \frac{\partial^2 \log P_{ij}(\theta + y_i\delta)}{\partial\theta^2} = -\sum_{j=0}^{m_i} j\alpha_i P_{ij} \left[j\alpha_i - \sum_{h=1}^{m_i} h\alpha_i P_{ih} \right].$$

Note that also in this case the elements of h_1 are a difference between observed and expected values: If z_i is defined as the score obtained on item i, and $z = \sum_i y_i z_i \alpha_i$, then

$$h_1 = z - \sum_{i=1}^{k} y_i \alpha_i \sum_{j=1}^{m_i} j P_{ij}.$$

3.8 Appendix B: Details on estimation for simulation study 2

We use two estimation procedures.

1. Estimation taking into account one dimension (one subscale, say subscale $t = 1$) only. The first order derivative is set equal to zero, so we solve

$$0 = \frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_1} = \sum_{i|t=1} \sum_{j=0}^{m_i} x_{ij} \frac{\partial \log P_{ij}(\theta_1)}{\partial \theta_1}$$

that is

$$\sum_{i|t=1} \sum_{j=0}^{m_i} x_{ij} d_{tij} = 0.$$

We use the Newton-Raphson algorithm, we iterate

$$\theta_1^* = \theta_1 - \left[\sum_{i|t=1} \sum_{j=0}^{m_i} x_{ij} d_{tij} \right] / \left[\sum_{i|t=1} \sum_{j=0}^{m_i} x_{ij} D_{tij} \right]$$

until convergence.

2. Estimation taking into account all dimensions (all subscales,$t = 1, .., T$). The first order derivative is set equal to zero, so we solve

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_1} \\ \frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_2} \\ \frac{\partial \log L(\boldsymbol{\theta})}{\partial \theta_3} \end{bmatrix} = \begin{bmatrix} \sum_{i|t=1} \sum_{j=0}^{m_i} x_{ij} \frac{\partial \log P_{ij}(\theta_1)}{\partial \theta_1} \\ \sum_{i|t=2} \sum_{j=0}^{m_i} x_{ij} \frac{\partial \log P_{ij}(\theta_1)}{\partial \theta_1} \\ \sum_{i|t=3} \sum_{j=0}^{m_i} x_{ij} \frac{\partial \log P_{ij}(\theta_1)}{\partial \theta_1} \end{bmatrix} - \boldsymbol{\Sigma}^{-1} \boldsymbol{\theta}.$$

Using

$$\mathbf{d} = \begin{bmatrix} \sum_{i|t=1} \sum_{j=0}^{m_i} x_{ij} d_{1ij} \\ \sum_{i|t=2} \sum_{j=0}^{m_i} x_{ij} d_{2ij} \\ \sum_{i|t=3} \sum_{j=0}^{m_i} x_{ij} d_{3ij} \end{bmatrix}$$

this is

$$\mathbf{d} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta} = \mathbf{0}.$$

We use the Newton-Raphson algorithm, so we iterate

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - \mathbf{D}^{-1}\left[\mathbf{d} - \boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}\right],$$

with

$$\mathbf{D} = Diag\left[\sum_{i|t} D_{tij}\right] - \boldsymbol{\Sigma}_{\theta}^{-1}.$$

4
An Application of Person Fit Tests to Multidimensional Personality and Cognitive Tests

In both maximum performance testing (e.g., standardized multiple choice tests of verbal ability) and typical performance testing (e.g., Likert ratings or other preference-type measures; personality questionnaires) there has been a growing interest in using statistical indexes to assess the degree to which formal measurement models accurately represent the relation between trait levels and a person's item response pattern. This research is often referred to as person fit research (e.g., Meijer & Sijtsma, 2001) and the statistical indexes are often referred to as person-fit statistics or scalability indices. Most research studies, however, have been conducted in the context of maximum performance testing using statistics that are designed for unidimensional tests or questionnaires. In practice, however, many tests and questionnaires consists of small subtests that together measure a higher order construct. Well-known examples are the NEO personality inventory, where six subtests measure one higher order construct such as Neuroticism or the Wechsler Intelligence Scales where three subtests together measure Verbal Comprehension. In the present study, we discuss a method that was recently proposed in Chapters 2 and 3 and show how this method can be used to investigate the fit of item response patterns to an item response theory (IRT; van der Linden & Hambleton, 1997) model for tests or questionnaires that consist of multiple related scales. This is done using both maximum performance and typical performance data, although the emphasis is on typical performance data because person fit in this context has been underexposed (Meijer & Sijtsma, 2001).

This study is organized as follows. First, we discuss the potential usefulness of person fit in typical performance testing. Second, we introduce the Lagrange-Multiplier (LM) test in Chapters 2 and 3 that, unlike most scalability indices, can take into account the multidimensional structure of the questionnaire. Third, we study the behavior of this fit statistic using empirical NEO PI-R data. In particular, we study classification of item score patterns using different groupings of items and we try to interpret response patterns that are classified as fitting and misfitting. Fourth, we apply the LM test to cognitive data from a child intelligence test and finally we discuss the results and the pro's and con's of this person fit method as compared to other person-fit statistics to analyze psychological data.

4.1 Person Fit in Typical Performance Testing

From the beginning of personality assessment, authors of questionnaires and inventories are seriously concerned with both measuring and correcting for respondents' tendencies to deceive themselves or others in answering items of, for example, potential pathological significance. Therefore, for several questionnaires, validity scales have been developed. Famous examples are the Lie (L) scale, the Infrequency (F) scale, and the correction (K) scale from the Minnesota Multiphasic Personality Inventory (Hathaway & McKinley, 1993). The L scale, for example, consists of 15 items which express common human weaknesses (e.g., "Once in a while I laugh at a dirty joke") that are denied by very few respondents. Answering too many of these items in the "saintly" direction raises questions concerning the truthfulness of the respondent. Many studies have examined the usefulness of validity scales and, in general, acceptable detection rates have been found. A drawback of a validity scale is, however, that separate scales should be developed which seems only possible for relatively long questionnaires. As an alternative, internal procedures have been proposed that investigate the scalability of an item score pattern. Under individual response pattern scalability we mean the fit of an individual observed response pattern to a measurement model, in this study an IRT model.

In the context of ability testing a number of indices and statistical tests have been proposed to assess the scalability of score patterns. Response pattern scalability measures may also help researchers in personality assessment to identify atypical patterns of responses behavior and as Reise and

Waller (1993) noted they should play a larger role in personality assessment as well. For example, these measures may help to identify persons who do not fit a particular conception of a personality trait. Choca, Shamely, and Van Denburg (1992) discussed that personality traits are not always distributed in a way that makes logical sense or that is internally consistent. Tellegen (1988) used the term "traitedness" to refer to the degree to which a person's behavior is consistent with a dimensional construct. Behavior is not always consistent as would predicted from our personality traits and therefore methods are needed that allow us to investigate the scalability of an individual response pattern. Reise and Waller (1993) investigated the usefulness of a scalability index to the Multidimensional Personality Questionnaire to investigate traitedness. They analyzed 11 unidimensional subscales and although they found low split-half reliabilities for the 11 subscales, data analysis showed that scalability indices can be used to explore the fit of an individual's behavior and a personality construct. Care should be taken, however, in the interpretation of unscalable patterns. Interpreting response pattern scalability as an indicator of traitedness variation is difficult. Possible causes may be response faultiness, misreading, or random responding. Thus, many scalability methods do not allow the recovery of the mechanism that created the deviant item score patterns. For example, the scalability index used by Reise and Waller (1993) is sensitive to many 1 ("agree") scores to items for which most persons get a 0 score ("disagree"). The reason why a person does produce an unlikely response pattern can only be obtained by auxiliary information from earlier testing, personality characteristics, personal history, or observation.

Related research to scalability indices in the context of personality research was conducted by Zickar and Drasgow (1995) to the relation between the power of scalabiliy approach and validity scales. Zickar and Drasgow (1996) evaluated the use of scalability indices for identifying dishonest respondents. A dataset was analyzed in which respondents were instructed either to answer honestly or to "*fake good*". The scalability approach classified a higher number of faking respondents at low rates of misclassification of honest respondents than did a social desirability scale.

Thus far, research on the scalability of individual score patterns has concentrated on unidimensional scales. Both research conducted by Reise and Waller (1993) and Zickar and Drasgow (1996) used unidimensional scales. Personality questionnaires, however, often consists of different small

homogeneous item sets, often called facets that measure a higher hierarchical construct. The individual total scores on these facets are positively correlated and measure a higher-order construct. A well-known example of a questionnaire that consists of facets is, for example, the NEO-PI-R (Costa & McCrae, 1995). Unidimensional scalability indices seem to be most suited to analyze scales with items that consists of fairly homogeneous sets of items, that is, scales that consists of items that covary positively with each other. In this study, we present and apply a method that can be used to investigate scalability for questionnaires that consists of different facet scales. Because a facet scale often consists of a relatively small number of items, the power of scalability indices for unidimensional IRT is low (e.g., Meijer, Molenaar, & Sijtsma, 1995). Taken into account information from other (correlated) scales may enhance the power of a person-fit statistic and as a result improves measurement of individual reponse patterns. Furthermore, we use a method that explicitly tests against specific violations of the assumptions of an IRT model. This method, therefore, may facilitate the interpretation of nonfitting item score patterns.

4.2 The LM Test

Recently, LM tests for IRT models have been proposed by Glas (1998, 1999, 2001), Glas and Suárez-Falcón (2003) and Becher, Verhelst, and Verstralen (2002). The LM test (Aitchison & Silvey, 1958) is equivalent with the efficient score test (Rao, 1947) and the modification index that is commonly used in structural equation modelling (Sörbom, 1989). The purpose of the LM test is to compare two models, say the null-model and some more general model that is derived from the null model by adding parameters. Only the null-model needs to be estimated. Sörbom (1989) showed that the value of the LM statistic is proportional to the expected increase of the conditional likelihood should the additional parameters be estimated. In the score test formulation, the statistic is based on estimation of the null model and performing one Newton-Raphson step for the added parameters. So, the test is based on an estimate that improves the likelihood, but does not completely maximize it under the alternative model. The more common likelihood ratio (LR) test, on the other hand, is based on actual maximization of the likelihood under the alternative model.

When applied to evaluate the fit of IRT models, the null model is the IRT model tested, and the alternative models represent model violations. The reason for considering the LM test, where a LR test is available, is that in complicated models with many parameters (such as IRT models) every item and person may be the source of various model violations. Instead of estimating all the alternatives for all persons and items, and performing a vast number of LR tests, one can perform a number of LM tests using one estimate under the null model only. So the LM test must be seen as a diagnostic tool, and it derives its relevance from the fact that it serves another purpose than the LR test.

4.2.1 LM Test For Multidimensional Data

In many situations, the assumption that an individual's response behavior can be explained by a unidimensional person parameter θ does not hold. In that case the assumption of unidimensionality can be replaced by the assumption of a multidimensional person parameter $\theta_1, ..., \theta_q, ..., \theta_Q$.

For the graded response model (Samejima, 1969) the probability of a response in category j of item i, $P(X_{ij} = 1)$, is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Psi\left(\alpha_i\left(\theta - \beta_{i1}\right)\right) & \text{if } j = 0 \\ \Psi\left(\alpha_i\left(\theta - \beta_{ij}\right)\right) - \Psi\left(\alpha_i\left(\theta - \beta_{i(j+1)}\right)\right) & \text{if } 0 < j < m \\ \Psi\left(\alpha_i\left(\theta - \beta_{im}\right)\right) & \text{if } j = m. \end{cases}$$

where α and β are the item parameters and Ψ is the abbreviation used for the logistic function given by

$$\Psi\left(x\right) = \frac{\exp(x)}{1 + \exp(x)} \ .$$

To assure that the probability $P_{ij}(\theta)$ is positive, the restriction $\beta_{i(j+1)} > \beta_{ij}$, for $0 < j < m$ is imposed. The multidimensional version of the graded response model is defined by replacing the term $\alpha_i\theta$ by $\sum_{q=1}^{Q} \alpha_{iq}\theta_q$. Further, it is assumed that $\theta_1, ..., \theta_q, ..., \theta_Q$ have a Q-variate normal distribution (McDonald, 1997; Reckase, 1997).

Only the special case of between-items multidimensionality will be considered here. This entails the assumption that the test can be split up into T subtests and every subtest relates to a specific person parameter θ_t. So

this is the special case where all α_{iq} $(q = 1, ..., Q; Q = T)$ are zero except for the proficiency parameter θ_t to which the item relates.

Assuming local independence of responses given θ, the likelihood of a response pattern $\mathbf{x}$ is given by

$$p_{\theta_t}(\mathbf{x}) = \prod_{t=1}^{T} \prod_{i|t} \prod_{j=0}^{m_i} P_{ij}(\theta_t)^{x_{ij}} g\left(\theta_1, \theta_2, ..., \theta_T | \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta\right), \tag{4.1}$$

where $g\left(\boldsymbol{\theta}|\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta\right)$is a multivariate normal density with covariance matrix $\boldsymbol{\Sigma}_\theta$. The mean $\boldsymbol{\mu}_\theta$ is set equal to zero to identify the latent scale.

Similarly, a general alternative model is given by $P(X_{ij} = 1|Y_i = y_i) = P_{ij}(\theta_t + y_i\delta)$. The variable Y_i can be the score on some other item or a function of the score on a set of items in the test (without item i). Testing the null hypothesis that $\delta = 0$ is done using

$$LM = \frac{h_1^2}{\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}^{'} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}}, \tag{4.2}$$

where h_1 is the first order derivative of the log-likelihood of the general alternative model, $\boldsymbol{\Sigma}_{22}$ is the variance of h_1^2 using true parameter and $\boldsymbol{\Sigma}_{21}^{'} \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}$ takes into account the loss of variation due to the estimation of θ.

In Dagohoy and Glas (2005) it is shown that h_1 can be rewritten as the difference between observed and espected values, and as such can be used to interpret person fit in terms of residuals. They furthermore showed using simulated data that taking auxiliary information from other subscales, that is using an estimate of all ability parameters pertaining to all subscales, increases the power as compared using the estimate of θ_t obtained on the focussed subscale alone.The LM statistic is evaluated using the maximum likelihood estimates of the parameters of the special model of the null hypothesis. In this case, the LM statistic has an asymptotic χ^2-distribution with 1 degree of freedom.

4.3 Empirical Study 1

Instrument and Data

To illustrate the usefulness of the LM statistic to detect aberrant reponse patterns, in the first study we used empirical data from the the NEO-PI-R

personality scale (Costa & McCrae 1992a). A sample of 1168 persons was available. The sample consisted of female undergraduates who participated for course credit. The NEO-PI- R consists of 240 items; each item contains a statement such as "I am often down in the dumps" that is scored on a five-point Likert scale ranging from 1 "strongly disagree", 2 "disagree", 3 "neutral ", 4 "agree", and 5 "strongly agree". The items are partitioned into thirty facet scales that contain eight items each. Six facets (thus 48 items) together measure a domain. There are five domains: Neoroticism, Extraversion, Openness, Agreeableness, and Conscientiousness.

Data Analysis

The polytomous IRT graded response model was used to account for test performance. The item parameters for all the NEO items were obtained by calibrating the item responses from the 1168 respondents with Multilog (Thissen, Chen, & Bock, 2003). This was done by estimating a unidimensional model for every domain within every facet. Then, latent covariance parameters were estimated using a procedure by Rubin and Thomas (2001), and multidimensional person parameters were estimated by maximum likelihood. The latter two estimates were computed by dedicated software developed by the authors. Finally, the LM index was computed and score patterns with LM values larger than 3.85 were considered to be misfitting. Table 4.1 gives an overview of the mean α and the range of β within the facets and domains.

We first investigated the agreement between scalability results when using different groupings of items for each domain. Therefore, two analyses were preformed, a detailed and aggregate analysis. First, a detailed analysis was carried out as follows. For each domain, there are six subscales, $t = 1,...,6$, associated with six ability parameters θ_1, θ_2, θ_3, θ_4, θ_5, θ_6 with eight items for each subscale. The null-hypothesis that the ability parameter θ_{1a} of the first 4 items of the first subscale relate to the same ability parameter θ_{1b} as the last 4 items in the same subscale was tested. That is, H_o: $\theta_{1a}=\theta_{1b}$ taking into account the responses from the other remaining 5 subscales θ_2, θ_3, θ_4, θ_5, θ_6. The aggregate analysis was done by making clusters of two subscales with correlations close to each other. That is, subscales with the highest correlation were grouped together. For example, Table 4.2 gives the correlation between subscales for the Openness domain. As can be seen in this table, we grouped subscales Fantasy and Aesthetics together with correlation of 0.44, Feeling and Actions with correlation of 0.38 and,

Table 4.1
Descriptives of the Scales: average α and range of β for five facets and six domains within each facet

Scale	No. items	α	β
A1 Trust	8	1.646	(-2.688, 2.220)
A2 Straightforwardness	8	1.448	(-3.964, 1.985)
A3 Altruism	8	1.382	(-4.240, 0.983)
A4 Compliance	8	1.063	(-2.860, 2.827)
A5 Modesty	8	1.282	(-3.766, 2.343)
A6 Tender-mindedness	8	0.935	(-5.418, 3.281)
C1 Competence	8	1.290	(-4.181, 2.434)
C2 Order	8	1.394	(-3.094, 2.648)
C3 Dutifulness	8	1.001	(-4.812, 1.744)
C4 Achievement striving	8	1.671	(-3.732, 3.181)
C5 Self-discipline	8	1.326	(-3.514, 2.478)
C6 Deliberation	8	1.145	(-4.369, 2.857)
E1 Warmth	8	1.156	(-4.138, 1.109)
E2 Gregariousness	8	1.302	(-3.476, 2.012)
E3 Assertiveness	8	1.491	(-2.451, 2.472)
E4 Activity	8	0.985	(-5.515, 4.223)
E5 Excitement seeking	8	1.202	(3.707, 1.464)
E6 Positive emotions	8	1.553	-(3.303, 1.232)
N1 Anxiety	8	1.409	(-2.961, 2.167)
N2 Angry hostility	8	1.269	(-2.618, 2.880)
N3 Depression	8	1.752	(-2.129, 1.989)
N4 Self conciousness	8	1.163	(-3.147, 2.568)
N5 Impulsiveness	8	1.114	(-3.917, 2.911)
N6 vVulnerability	8	1.246	(-2.756, 3.171)
O1 Fantasy	8	1.370	(-3.543, 1.674)
O2 Aesthetics	8	1.519	(-3.255, 1.137)
O3 Feeling	8	1.375	(-3.868, 0.985)
O4 Actions	8	0.926	(-4.331, 3.811)
O5 Ideas	8	1.597	(-2.995, 1.564)
O6 Values	8	1.034	(-4.241, 1.698)

Table 4.2
Correlation matrix for the six subscales in the Openness Domain in NEO Personality Data

	Fantasy	Aesthetics	Feeling	Actions	Ideas	Values
Fantasy	1.000	**0.442**	0.337	0.253	0.307	0.210
Aesthetics		1.000	0.401	0.311	0.469	0.219
Feeling			1.000	**0.385**	0.277	0.293
Actions				1.000	0.269	0.266
Ideas					1.000	**0.294**
Values						1.000

Ideas and Values with correlation of 0.29. So three subscales, t = 1,...,3, associated with three ability parameters θ_1^*, θ_2^*, θ_3^* of 16 items each were analyzed and we tested the null-hypothesis that the first 8 items in the first cluster of 16 items relate to the same ability parameter as the last 8 items in the same cluster. That is, H_o: θ_{1a}^*=θ_{1b}^* taking into account the responses from the other remaining 2 subscales θ_2^*, θ_3^*. And finally, we cross tabulated persons detected as aberrant for the different types of computations. Table 4.3 gives the result of our cross-tabulation on persons detected as aberrant for the different types of computations.

The "plus " and "minus " signs in all the tables refer to aberrant and normal persons, respectively. Coefficient Kappa was used as a measure of the degree of agreement; the values are given at the bottom of the table. The values of Kappa indicate that the degree of agreement between our detailed and aggregate analysis is close to zero. The expected value was also computed and these are the values inside the parenthesis and we see that the difference between the observed and expected values are also close to zero. That is, the observed and expected values are almost in agreement with each other. We conclude that the agreement between the detailed and aggregate analsysis was small and that the particular configuration of the items taken into account in the person-fit analysis influences the classification of a score pattern as fitting or misfitting. Still, the proportion of respondents that were categorized similarly by the two approaches was between 0.7 and 0.8. Although we should thus be careful in interpreting the item score pattern because of the relatively small number of items per subtest, we were curious to know whether the configuration of the item scores within a pattern differs for score pattern with high and low LM values. In Table 4.4 we depicted four score patterns that were flagged as misfitting by the LM statistic in the Openness domain. This domain consists of 6 facets: Fantasy, Aesthetics, Feelings, Actions. Ideas,

Table 4.3
Cross tabulation tables for the Domains in NEO Personality Data (NEW)

Openness

		6_dims +	6_dims -
3_dims	+	.021 (*.014*)	.131 (*.138*)
	-	.072 (*.079*)	.776 (*.769*)

Kappa = 0.064

Neuroticism

		6_dims +	6_dims -
3_dims	+	.032 (*.035*)	.181 (*.178*)
	-	.131 (*.128*)	.656 (*.659*)

Kappa = -0.020

Extraversion

		6_dims +	6_dims -
3_dims	+	.011 (*.010*)	.122 (*.123*)
	-	.062 (*.063*)	.805 (*.804*)

Kappa = 0.011

Conscientiousness

		6_dims +	6_dims -
3_dims	+	.014 (*.014*)	.178 (*.177*)
	-	.062 (*.061*)	.746 (*.746*)

Kappa = 0.000

Agreeableness

		6_dims +	6_dims -
3_dims	+	.018 (*.024*)	.164 (*.158*)
	-	.114 (*.108*)	.704 (*.710*)

Kappa = -0.045

and Values. We grouped Fantasy and Aesthetics together and compared the coresponding $\widehat{\theta}s$ of these two facets taken into account the $\widehat{\theta}$ values of the remaining subtests. In Table 4.4, we depicted the LM-index values and $\widehat{\theta}$ values for each subscale. Consider, for example, the item score pattern of person 1004. This person obtains $\widehat{\theta} = 1.23$ based on the 16 items of the first two subtests but has relatively many low scores on the first subtest "Fantasy" resulting in $\widehat{\theta} = 0.26$ and relatively high scores on Aesthetics resulting in $\widehat{\theta} = 1.72$. Given the relatively high correlation between these subscores this is unexpected. Another example is person 1064. This person obtains a relatively high value $\widehat{\theta} = 1.62$ on Fantasy and a relatively low value on Aesthetics ($\widehat{\theta} = 0.01$). For both persons it can be questioned whether the conception of the trait that is measured by both subtests namely the overall trait Openness has the same interpretation. Thus for these persons the response behavior is not consistent with the dimensional construct. For item scores fitting the model (Table 4.5) the $\widehat{\theta}$ values on the first and the second subtests are in agreement

4.4 Empirical Study 2

Data and Method

In this second study, we analyzed data from a child intelligence test, the Revised Amsterdam Child Intelligence test (RAKIT, Bleichrodt, Drenth, Zaal, & Resing, 1984). The RAKIT consists of 12 subtests and measures cognitive development of children ranging from age 4 to 11. The test is used for individual diagnosis with respect to, for example, school choice and cognitive developmental retardation. The sample consisted of 408 persons. We used data from two subtests measuring perceptual reasoning, namely, *Figure Recognition* and *Exclusion*. In the subtest Figure Recognition *(*35 items*)* each item presents an incomplete drawing of an everyday object (e.g., a ball or a shoe) and the child has to name the object displayed in the incomplete drawing. In the subtest *Exclusion* (30 items) each item consists of four abstract figures. Three of them share a common rule that is not shared by the fourth one. For example, three figures consist of a triangle and a circle, whereas the fourth figure consist of a triangle and a square. The child has to name the figure that does not obey the common rule shared by the others. This test measures logical reasoning, in particular, inductive reasoning. The items of both tests are dichotomously

Table 4.4
Response pattern of sample persons who are detected as aberrant

Person	LM(δ^*)	Response pattern $[\theta_{1a}$; $\theta_{1b}]$ ($\boldsymbol{\theta}_1$) ($\boldsymbol{\theta}_2$) ($\boldsymbol{\theta}_3$)			
869	5.88(-1.99)	5344444454252344	[1.97 ; 0.59]	(1.13)	4444444425243424(2.25) 3242444442424443(1.26)
1083	3.94(1.52)	5233113142444432	[-0.37 ; 0.90]	(0.44)	2444424314142324(1.09) 2242224443251354(0.21)
1064	9.95(-2.51)	3424444422232433	[1.62 ; 0.01]	(0.55)	5445444424222222(1.13) 4444444424441443(2.47)
1004	2.84(1.42)	5224224244444544	[0.26 ; 1.72]	(1.23)	4444444424244442(2.65) 4444444444444443(2.96)

Table 4.5
Response pattern of sample persons who are NOT detected as aberrant

Person	LM(δ^*)	Response pattern $[\theta_{1a}$; $\theta_{1b}]$ ($\boldsymbol{\theta}_1$) ($\boldsymbol{\theta}_2$) ($\boldsymbol{\theta}_3$)			
15	0.00(-0.05)	4444522244445415	[0.84 ; 0.78]	(0.86)	2442422224234232(0.10) 4444452453344453(1.62)
61	0.05(-0.16)	5545544444143433	[1.11 ; 0.99]	(1.10)	2454443434242443(1.42) 3333454455243443(1.10)
616	0.05(-0.18)	4443345545224443	[1.32 ; 1.00]	(1.19)	5554555524242434(0.45) 4354454544455553(1.18)
720	0.01(0.06)	5542445555554545	[0.64 ; 0.52]	(0.58)	5555555545254544(0.16) 5525255555554554(0.01)
795	0.02(-0.12)	4223235342543234	[0.28 ; 0.40]	(0.37)	3444444234232223(1.41) 2232344444444444(1.06)

scored (correct, incorrect) and are administered in ascending difficulty order. Emons, Sijtsma, and Meijer (2005) analyzed these subtests in an IRT context and found a good fit to the data. They also analyzed the data at the individual level using a nonparametric person fit approach, where they used both a global person-fit statistic to make the binary decision about fit and misfit of a person 's item score pattern and a local person-fit approach to evaluate unexpected parts in the item score pattern . They, however, did not explicitly test against multidimensionality and their person-fit analysis was conducted for each subtest separately, thus not taking auxiliary information from other subtests into account.

We used the two parameter logistic IRT model (2PLM, van der Linden & Hambleton, 1997) to analyze the test data. MML estimates of the item parameters were computed using Multilog. Using these item parameters, MML estimates of the item and latent covariance parameters were again computed using software developed by the authors and also trait level scores were estimated using maximum likelihood using software developed by the authors. Finally, the multidimensional LM index was computed. This was done as follows: we split the score patterns of the Figure Recognition subtest into two parts consisting of the first 18 items and the second 17 items and we used the $\widehat{\theta}$ values on Exclusion as auxilairy information. Given the null hypothesis of a fitting IRT model, we expect similar $\widehat{\theta}$ values.Again, score patterns with LM values larger than 3.85 were considered to be misfitting.

Results

Twenty-three of the 408 score patterns were classified as aberrant. In Table 4.6 nine deviant item score patterns on Figure Recognition are given with their corresponding LM values and $\widehat{\theta}s$ on the first and the second subtest. To illustrate the use of the LM statistic consider the score pattern of person 335. On the first subtest this person obtains $\widehat{\theta} = -0.02$, whereas on the second subtest $\widehat{\theta} = 2.69$. Inspecting the score pattern it can be seen that the second part of the score pattern (containing relatively difficult items) consists of relatively many 1 scores whereas the first part (consisting of the relatively easy items) consists of many zero scores. This is unexpected on the basis of the probabilities nature of the 2PLM. There are too many correct answers on the difficult items and too few correct answers on the easier items. A possible interpretation of this score pattern is that person 335 is someone who was working very fast and who easily skipped items A

Table 4.6
Response pattern of sample persons who are detected as aberrant

Person	LM(δ^*)	Response pattern	$[\theta_{1a}\ ;\ \theta_{1b}]$	$(\boldsymbol{\theta}_1)$	Response pattern	$(\boldsymbol{\theta}_2)$
369	4.95(1.58)	11110111111110111111	[0.40 ; 3.11]	(1.24)	11111111110101111110100111010010	(1.01)
357	5.03(-1.38)	10111111111111111110	[0.21 ; -0.71]	(-0.09)	11111111111111111111111111101111100	(1.65)
335	8.54(1.89)	11110111100011111111	[-0.02 ; 2.69]	(0.83)	11111111111111010000000000000000	(-0.62)
292	7.51(1.68)	11111111111110101000	[-0.48 ; 1.90]	(0.53)	11111111111111011111000000000	(0.42)
269	8.59(2.12)	11110010011001100000	[-1.56 ; 0.45]	(-0.66)	11111101111111111110010110000010	(0.36)
240	9.07(1.91)	11111011110011011000	[-0.44 ; 2.34]	(0.76)	11111110111111111101111010101010	(1.17)
212	3.91(1.35)	11111111110111111011	[0.83 ; 2.69]	(1.07)	11111111010011000000000000000000	(-1.20)
183	6.09(1.50)	11111111010100111010	[-0.39 ; 1.71]	(0.46)	11111111111011111110001000000000	(-0.12)
177	4.21(-1.37)	11111110111111111110	[-0.02 ; -1.07]	(-0.43)	11111111111011010100000000000000	(-0.98)

different type of unexpected item score pattern was generated by person 357. Given an overall $\widehat{\theta} = -0.09$, this person answered 16 out if the 18 items correctly and answered only one item of the most difficult items correctly. This is unexpected on the basis of the probabilistic nature of the 2PLM. We expected fewer correct scores on the easier subtest than were obtained and more correct answers on the second subtest. This pattern is a good example of the answers of someone who is working very slowly but precisely and uses a suboptimal strategy to obtain a maximal score given his or her ability. If he or she had guessed the answers on the more difficult items later in the test, the test score would have been higher. In Table 4.7 we depicted some fitting item score patterns. It is interesting to consider person 13, like person 357 (Table 4.6) this person has a $\widehat{\theta}$ value around zero and also 2 incorrect answers on the first subtest but this person has more correct answers on the second part of Figure Recognition resulting in a fitting item score pattern.

4.5 Discussion

An application of the LM statistic is as a check of the interpretability of the trait level score. This requires no assumptions regarding the source of the low scalability. By definition, if an examinee has a high value on the LM statistic his or her pattern of item responses is not scalable on the trait dimension used to scale individual differences. Interesting is that earlier researchers in the personality context (e.g., Reise & Waller, 1993) used the standardized log likelihood statistic to investigate scalability. These type of statistics are often presented as statistics to investigate the general fit of an item score pattern. However, note that the log-likelihood statistic is sensitive to a specific type of misfit to the IRT model, namely violations to Guttman patterns (see Molenaar & Hoijtink, 1990). Denote the number-correct score as X_+ and assume that the test consists of k items, then when the items are ordered from easy to difficult, an item score pattern with correct responses in the first X_+ positions and incorrect responses in the remaining $k - X_+$ positions is called a Guttman pattern because it meets the requirements of the Guttman (1950) model. As we discussed above, a perfect Guttman pattern may result in violation of multidimensionality because the subtest score on the first subtest is too high and the subtest score on the second subtest is too low given the assumptions of

Table 4.7
Response pattern of sample persons who are NOT detected as aberrant

Person	LM(δ^*)	Response pattern		$[\theta_{1a}\ ;\ \theta_{1b}]$	$(\boldsymbol{\theta}_1)$	Response pattern	$(\boldsymbol{\theta}_2)$
5	0.07(-0.20)	111111110010011010	10001000000000000	[-0.88 ; -0.51]	(-0.64)	11111111001100100011010010100 1001	(-0.20)
13	0.05(0.13)	111011111111101100	11010010000000000	[-0.25 ; 0.27]	(-0.08)	11111111111111010011000000000 0000	(-0.51)
24	0.00(-0.01)	111111111111111001	11001011000000000	[0.13 ; 0.42]	(0.07)	11101110110110110110000000000 0000	(-0.92)
25	0.07(0.20)	111110110110100110	10100000000000000	[-0.93 ; -0.38]	(-0.75)	11111001111010010000000000000 0000	(-1.73)
28	0.00(0.01)	11011101111010011	100110000000000000	[-0.60 ; -0.10]	(-0.35)	11111111111101110110110000000 0000	(-0.02)
38	0.03(-0.19)	11110110011000000	000000010000000000	[-1.98 ; -1.14]	(-1.51)	11111111111111111111011011011 00	(0.90)
60	0.00(-0.02)	111011110001010000	10000010000000000	[-1.35 ; -0.70]	(-0.97)	11111111101111111011000111000 0000	(0.35)
85	0.04(-0.12)	1111110101010111111	1001001000000000	[-0.19 ; 0.22]	(0.00)	11011110110011111010111100101 0000	(0.64)
92	0.00(0.02)	1111111111111001110	1101001000000000	[0.09 ; 0.41]	(0.06)	11111111111101001000000000000 0000	(-0.93)

the stochastic IRT model. Thus different statistics are sensitive to different types of aberrant response behavior and depending on the type of violations the researcher is interested in, different statistics can be used. This is important to realize, especially in personality measurement where the question is whether scalability indices can provide information regarding the relation between between an individual's behavior and a nomothetic trait construct. It is clear that well designed trait measures associated with thought-out and empirically defended trait constructs may reduce the possibility of psychologically informative deviations in scalability.

Because behavior is not always consistent as our personality trait measurement models predict it to be, we require techniques that allow us to assess the congruency between someone's self-presentation behavior and the formal operationalization of an individual difference trait construct (the measurement model). Because aberrant response behavior may have multiple causes and hence may be ambiguous as to its source, it is very important to understand that different statistics are sensitive to different types of aberrant behavior. In certain cases, such as those shown in our analysis of response to the NEO subscales, we believe that the information from person-fit test statistics can be used as a foundation for arguing against the blind use of total scores on the basis of 48 items.

In this study we used auxiliary information from other subscales, in personality measurement and intelligence testing this may improve the power of person-fit statistic. For example it has been suggested that in a personality context statistics like the log-likelihood statistic are only useful when applied to very long personality scales. Using auxiliary information from other subscales may reduce this problem and seems to fit the existing practice of using many short scales to measure different aspects of a higher order construct.

We recognize that other sources of data such as informant descriptions are needed to validate such judgments, however. A possible strategy may be to ask a respondent for the reasons for misfit. For example, in intelligence testing, a psychologist may ask whether a respondent was extremely nervous at the start of the test, or in personality testing , it may be possible to check by means of directed questions whether the personality trait being measured was applicable to the respondent.

5
A Bayesian Approach to Evaluation of Person Fit to Polytomous IRT Models

Two new person fit statistics for item response theory (IRT) models for polytomous items were introduced in the previous chapters. The first test is focused on shifts in ability, the second is focused on violation of local independence. Both are Lagrange multiplier tests. It was shown that the derivation of the distribution of Lagrange multiplier statistics can take the effects of estimation of the item and person parameters into account. Naive test statistics that ignore the effects of estimation of the persons' ability parameters result in incorrect Type I error rates and a marked decrease of power. In this chapter, an alternative approach to detection of aberrant persons taking into account the uncertainty caused by parameter estimation based on a Bayesian framework and posterior predictive checks using Markov chain Monte Carlo (MCMC) methods will be presented. More information about posterior predictive checks can be found in Meng (1994) Gelman, Carlin, Stern and Rubin (1995); and Gelman, Meng and Stern (1996). Although this approach applies to parametric IRT model, we will only focus on the sequential model and graded response model.

The development of powerful sampling-based estimation techniques have stimulated the application of Bayesian methods. Compared to the traditional frequentist approach, this Bayesian approach has several advantages. First, there is no need to derive the theoretical sampling distribution

of the statistic, which sometimes may be very difficult, if not impossible. Second, the person-fit statistic may depend on unknown quantities such as the item and person parameters. This uncertainty is explicitly taken into account. The third advantage pertains to generality of the procedure. Simulation studies have shown that a fully Bayesian approach to estimation of the parameters in simple IRT models (say one- or two-parameter models) are generally not superior to estimates obtained by a maximum marginal likelihood (MML) procedure or a Bayes modal procedure (see, for instance, Baker, 1998; or Kim, 2001). However, the Bayesian approach also applies to complicated IRT models, where MML or Bayes modal approaches pose important problems. Recently, the fully Bayesian approach has been adopted to the estimation of IRT models with multiple raters, multiple item types, missing data (Patz & Junker, 1997, 1999), testlet structures (Bradlow,Wainer, &Wang, 1999;Wainer, Bradlow, & Du, 2000), latent classes (Hoijtink & Molenaar, 1997), models with a multilevel structure on the ability parameters (Fox & Glas, 2001) and the item parameters (Janssen, Tuerlinckx, Meulders, & de Boeck, 2000), and multidimensional IRT models (Béguin & Glas, 2001). The motivation for this recent interest in Bayesian inference and MCMC estimation procedures is that the complex dependency structures in the mentioned models require the evaluation of multiple integrals to solve the estimation equations in an MML or a Bayes modal framework (Patz & Junker, 1999). These problems are easily avoided in an MCMC framework.

The principle of the approach is as follows. To decide whether an item score pattern fits an IRT model, a sampling distribution under the null model, that is, the IRT model, is needed. Let t be the observed value of a person fit statistics T. Then the significance probability or probability of an exceedance is defined as the probability that the value of the test statistics is equal or smaller than the observed value, that is, $p=P(T\leq t)$ or equal or larger than the observed value, $p=P(T\geq t)$, depending on whether low or high values of the statistic indicates aberrant score patterns. The statistics T are closely related to the statistic discussed in the two previous chapters.

This article is organized as follows. First, some often used IRT models and person-fit statistics are introduced. Second, the principles of MCMC methods to sample the posterior distribution of a person-fit statistic will

be discussed. Third, a simulation study is presented to examine the effectiveness of several person-fit statistics.

5.1 IRT Models and Person Fit

5.1.1 Models for polytomous items

Only the graded response model (GRM, Samejima, 1969) and sequential model (SM, Tutz, 1997; Verhelst, Glas & de Vries, 1997) will be considered in this paper. The reason is that a well-proved Bayesian estimation procedure for these two models based on the normal ogive representation combined with data augmentation is already in existence (Albert, 1992; Johnson & Albert, 1999; Baker, 1998). This approach is not applicable to the generalized partial credit model (GPCM, Muraki, 1992, also see Masters, 1982). The estimation procedure for the GPCM relies on the logistic representation (Patz & Junker, 1999; Maris & Maris, 2002). To avoid confounding by the estimation framework, only the GRM and SM will be compared.

Consider items $i = 1, ..., k$, with categories $j = 0, ..., m_i$. We will drop the index i of m_i for convenience. A response pattern is coded as $x = (x_1, ..., x_i$
$, ..., x_k)$, a response on an item i as $\mathbf{x}_i = (x_{i0}, x_{i1}, ..., x_{ij}, ..., x_m)$, and $x_{ij} = 1$ if a response was given in category j, and zero otherwise. We will use an abbreviation for the normal ogive function given by

$$\Phi(x) = \int_{-\infty}^{\alpha_i(\theta-\beta_{i1})} \frac{1}{(2\pi)^{1/2}} \exp\left(-t^2/2\right) dt \ .$$

The Graded Response Model

In the graded response model (Samejima, 1969) the probability of a response in category j of item i, $P(X_{ij} = 1)$, is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Phi\left(\alpha_i\left(\theta - \beta_{i1}\right)\right) & \text{if } j = 0 \\ \Phi\left(\alpha_i\left(\theta - \beta_{ij}\right)\right) - \Phi\left(\alpha_i\left(\theta - \beta_{i(j+1)}\right)\right) & \text{if } 0 < j < m \\ \Phi\left(\alpha_i\left(\theta - \beta_{im}\right)\right) & \text{if } j = m \ . \end{cases} \tag{5.1}$$

To assure that the probabilities $P_{ij}(\theta)$ are positive, the restriction $\beta_{i(j+1)} > \beta_{ij}$, for $0 < j < m$ is imposed.

The sequential model

In the sequential model (Tutz, 1990) the probability of a response in category j of item i is given by

$$P_{ij}(\theta) = \begin{cases} 1 - \Phi\left(\alpha_i\left(\theta - \beta_{i1}\right)\right) & \text{if } j = 0 \\ \prod_{h=1}^{j} \Phi\left(\alpha_i\left(\theta - \beta_{ih}\right)\right)\left[1 - \left(\Phi(\alpha_i(\theta - \beta_{i(j+1)})\right)\right] & \text{if } 0 < j < m \\ \prod_{h=1}^{m} \Phi\left(\alpha_i\left(\theta - \beta_{ih}\right)\right) & \text{if } j = m\ . \end{cases} \tag{5.2}$$

Verhelst, Glas and de Vries (1997) note that every item in the SM can be viewed as a sequence of virtual dichotomous items. These dichotomous items are considered to be presented as long as a correct response is given, and the presentation stops when an incorrect response is given. An important consequence of this conceptualization of the response process is that estimation and testing procedures for the 2PL model with incomplete data can be directly applied to the SM.

5.1.2 Person Fit Tests

To investigate the goodness of fit of item score patterns, the two new person fit statistics for IRT models for polytomous items introduced in Chapter 2 and Chapter 3 will be used. The first test is focused on shifts in ability, the second is focused on violation of local independence. Both are Lagrange multiplier (LM) tests. For the case of dichotomously scored items, the formulation of this statistic is almost similar to the formulation of the UB-statistic by Smith (1985, 1986). The main difference is that Smith multiplies with a factor $1/G$, while the LM tests include a correction term in the denominator that accounts for the estimation of θ. For polytomously scored items, a generalization would result in

$$UB = \frac{1}{G}\sum_{g=0}^{G} \frac{\left[\sum_{i \in A_g}\left[y_i - E_\theta(Y_i)\right]\right]^2}{\sum_{i \in A_g}^{k}\sum_{j=0}^{m_i}\alpha_{ij}P_{ij}(\theta)\left[\alpha_{ij} - E_\theta(Y_i)\right]}, \tag{5.3}$$

where A_g (g=1,2,...,G) is a partitioning of the items into disjoint subsets; usually G=2. Further, $y_i = \sum_{j=0}^{m_i} x_{ij}\alpha_{ij}$, that is, it is the weighted score on item i, and $E_\theta(Y_i)$ is its expectation. In a Bayesian framework, Glas and Meijer (2003) found that the UB-statistic for dichotomous items by

Smith (1985, 1986) had an acceptable Type I error rate when simulating the distribution for these statistics using an MCMC method. Type I error rate indicates the number of incorrectly rejected null hypotheses based on the statistical test.

Another statistic that might be viewed as the generalization of the UB-statistic to a test for local independence is given by

$$UD = \frac{\left[\sum_i y_{k(i)} \left[y_i - E_\theta(Y_i)\right]\right]^2}{\sum_i \sum_{j=0}^{m_i} y_{k(i)}^2 \alpha_{ij} P_{ij}(\theta) \left[\alpha_{ij} - E_\theta(Y_i)\right]}, \tag{5.4}$$

where $y_{k(i)}$ is the response on any other item except item i. Usually $k(i) = i - 1$.

A third statistic that will be considered here is the well-known log-likelihood statistic

$$l = \sum_{i,j} x_{ij} \log P_{ij}\left(\theta\right), \tag{5.5}$$

which was first proposed by Levin and Rubin (1979). It was further developed in Drasgow, Levine, and Williams (1985), and Drasgow, Levine, and McLaughlin (1991).

Drasgow et al. (1985) proposed a standardized version l_z of l which is claimed to have an asymptotic standard normal distribution. The person-fit statistic l_z is often used, but Molenaar and Hoijtink (1990), and Van Krimpen-Stoop and Meijer (1999) showed that the distribution of l_z is negatively skewed. This skewness influences the differences between nominal and empirical Type I error rates for small Type I error values. They found that increasing the item discrimination resulted in a distribution that was increasingly negatively skewed.

In an MCMC framework, the distribution of a statistic is simulated, so we will only consider the person-fit statistic l instead of l_z. Analogously, we will study the LM statistic disregarding the correction term for the estimation of θ. This is the so called "naive" LM statistic, which is equivalent to the UB and UD statistic.

The statistics given by (5.5), (5.3) and (5.4) will be compared in the simulation studies.

5.1.3 Evaluating the fit of an item score pattern.

To evaluate the fit of an item score pattern a norm distribution is needed for classifying an item score pattern as fitting or misfitting. This norm distribution can be obtained using a theoretical distribution (e.g., a normal distribution) or it can be simulated. In this paper, we will simulate the norm distribution because often the theoretical distributions proposed in the literature are not in the agreement with the empirical distributions (Meijer & Sijtsma, 2001). Also, the error in the item and person parameters can be taken into account when we simulate the norm distribution. Recently, Glas and Meijer (2003) used a Bayesian approach. Their approach has the advantage that they take into account the uncertainty of the parameters in the IRT model. In this Bayesian method, the posterior distribution of the parameters of the 3PNO model, say $p(\xi|y)$, where ξ are the item and person parameters in the model and y is the observed data, is simulated using an MCMC method. Person fit is then evaluated using a posterior predictive check (Gelman, Meng, & Stern, 1996) based on an index $T(y,\xi)$ where T refers to the person-fit statistics given above. When the Markov chain has converged, draws from the posterior distribution can be used to generate model-conform data y^{rep} and to compute the Bayes p-value defined by

$$\text{Bayes } p\text{-value} = \Pr\left(T\left(y^{rep},\xi\right) \geq T\left(y,\xi\right)|y\right).$$

The Bayes p-value is defined as the probability that the replicated data are more extreme than the observed data. Posterior predictive checks are performed by inserting the person-fit statistics l, UB , UD into the equation. After the burn-in period, when the Markov Chain has converged, in every n-th iteration ($n \geq 1$), using the current draw of the item and person parameters, a person-fit statistic $T(y,\xi)$ is computed, a new model conform response pattern is generated, and a value $T(y^{rep},\xi)$ is computed. Finally, a Bayes p-value is computed as the proportion of iterations in which $T(y^{rep},\xi) \geq T(y,\xi)$.

Detection rates of the different statistics are investigated for different model violations and test lengths using parameter estimates obtained combining normal and aberrant simulees and using normal persons as a calibration sample is investigated.

5.2 Bayesian Estimation of the Models

Bayesian estimation provides a rigorous approach for estimating the probability distribution of unknown variables by utilizing all the available knowledge and data. It considers all the parameters to be stochastic variables and determines the distribution of the variables to be estimated, θ , given the data, y.

In this study, an MCMC procedure will be used to generate the posterior distributions of interest. The MCMC chains will be constructed using the Gibbs sampler (Gelfand & Smith, 1990). To implement the Gibbs sampler, the parameter vector is divided into a number of components, and each successive component is sampled from its conditional distribution given sampled values for all other components. This sampling scheme is repeated until the sampled values form stable posterior distributions.

For the procedure for estimation of the parameters of the GRM we refer to Johnson and Albert (1999, Sec. 6.9). For the SM, due to the consequence of the conceptualization of the response process, the estimation procedure presented by Albert (1992; see also Baker, 1998) applying Gibbs sampling to estimate the parameters of the well-known 2PNO model (e.g., Lord & Novick, 1968) with missing data will be used. For application of the Gibbs sampler, it is important to create a set of partial posterior distributions that are easy to sample from. This often involves data augmentation, that is, the introduction of additional latent variables that lead to a simple set of posterior distributions. In the Gibbs sampling algorithm, these latent variables are sampled along with the variables of interest. The present procedure is based on an augmentation step that transforms the discrete responses to continuous responses. All priors were non-informative.

The convergence of MCMC procedure was investigated by comparing the between and within sequence variance. Based on these analyses, it was concluded that a burn-in period of 1000 iterations was sufficient and the sampled parameter values converged within 4000 iterations. The first 1000 iterations were discarded. In the remaining 3000 iterations, $T\left(y^{rep}, \xi\right)$ and $T\left(y, \xi\right)$ were computed every fifth iteration. So the posterior predictive checks were based on 600 draws. For the statistics that use a partitioning of the items into subtests, two subtests of equal size were formed.

5.3 Simulation Studies

Two sets of simulation studies will be reported. In the first set of simulations, Type I error rate of tests for detection of changes in ability and detection of local independence will be studied. In the second set of simulations, the power of the test for detection of changes in ability and local independence will be studied. These studies also address the false alarm rate and the specificity of the tests, that is, the extent to which tests are sensitive to other model violations than the one they are targeted at. In both simulation studies, two types of parameter estimation procedures are presented, that is, using estimates obtained combining normal and aberrant simulees and using only normal simulees as a calibration sample.

Sample size of $N = 1000$ was crossed with test lengths of $K = 10$, $K = 20$ and $K = 30$. Item parameters were equal to the item parameters in Chapter 3 and the ability parameters were drawn from a standard normal distribution. All parameters were transformed from the logistic to the normal ogive representation. The number of replications in each branch of the study was equal to 100.

5.3.1 Type I error rate

The aim of this section is to compare the Type I error rate of l, UB , UD. The results are shown in Table 5.1. The first column labeled 'Generating Model' gives the model used for generating the data. The estimation model is given in the second column labelled 'Estimation Model'. The third column labeled 'K' gives the number of items and L, UB and UD give the proportions of rejections at a 5% significance level for the log-likelihood statistic and tests for constancy of theta and violation of local independence, respectively.

The results show that the Type I error rate for the combination of data generation and estimation with the SM and the combination of the data generation and estimation with the GRM were below the nominal 5%-level. The l-test was closest, while the UD was lowest (around 0.003). This pattern was replicated when the data were generated with the SM and estimated with the "wrong model", in this case the GRM. For the combination of the GRM as a generating model and the SM as an estimation model, the Type I error rates were grossly inflated. This means that obtaining global model fit before evaluation of person fit is essential.

Table 5.1
Type I error rate

Generating Model	Estimation Model	K	L	UB	UD
SM	SM	10	.04	.02	.00
		20	.04	.02	.00
		30	.04	.02	.00
	GRM	10	.04	.02	.00
		20	.04	.02	.00
		30	.04	.02	.00
GRM	SM	10	.08	.02	.03
		20	.17	.13	.14
		30	.27	.21	.34
	GRM	10	.04	.02	.00
		20	.04	.02	.00
		30	.05	.02	.00

5.3.2 Power of the Tests

The next set of simulation studies pertains to the power of the l, UB , UD-test to detect model violations in aberrant persons. Changes in ability were generated by shifting the person parameter θ with an amount δ. The parameter δ will be referred to as the effect size. Local independence was violated by cumulatively adding δy_{i-1} to the person parameter θ in the probability of the response y_i. So the alternative model can be viewed as a response-dependent learning model.

The estimation procedure was run using parameter estimates obtained combining normal and aberrant simulees and using only normal simulees as a calibration sample. In the later case, for every 5th draw of the item parameters, an importance sample (Tanner, 1993) of the parameter θ of an aberrant person was obtained, and $T\left(y^{rep}, \xi\right)$ and $T\left(y, \xi\right)$were computed using this draw. In all simulations, 10% of the simulees were aberrant. Two simulation studies will be reported, pertaining to detection of changes in ability and detection of violation of local independence. The simulation design consisted of crossing test lengths of 10, 20 and 30, two effect sizes δ, and the two calibration setups.

The results are shown in Tables 5.2, 5.3, 5.4 and 5.5, for detection of changes in ability using estimates obtained combining normal and aberrant simulees and using normal persons as a calibration sample, and detection of local independence using estimates obtained combining normal and aberrant simulees and using normal simulees as a calibration sample, respectively. The second column labeled 'Generating Model' gives the model used

Table 5.2
Detection of changes in ability using parameter estimates obtained combining normal and aberrant simulees

No. of Items	Generating Model	Estimation Model	δ	UB		UD		Lkd	
				False Alarms	Hits	False Alarms	Hits	False Alarms	Hits
10	SM	SM	0.5	.02	.05	.00	.00	.04	.04
			1.0	.02	.13	.00	.00	.04	.08
		GRM	0.5	.02	.05	.00	.00	.04	.04
			1.0	.02	.13	.00	.00	.04	.08
	GRM	SM	0.5	.02	.01	.04	.03	.08	.07
			1.0	.03	.01	.04	.03	.07	.08
		GRM	0.5	.02	.04	.00	.00	.04	.06
			1.0	.02	.10	.00	.00	.04	.07
20	SM	SM	0.5	.02	.08	.00	.00	.04	.05
			1.0	.02	.29	.00	.00	.04	.09
		GRM	0.5	.02	.08	.00	.00	.04	.05
			1.0	.02	.29	.00	.00	.03	.09
	GRM	SM	0.5	.11	.19	.13	.13	.17	.17
			1.0	.10	.31	.12	.13	.16	.20
		GRM	0.5	.02	.07	.00	.00	.04	.06
			1.0	.02	.23	.00	.00	.04	.10
30	SM	SM	0.5	.02	.12	.00	.00	.04	.06
			1.0	.02	.45	.00	.00	.04	.11
		GRM	0.5	.02	.12	.00	.00	.04	.07
			1.0	.02	.44	.00	.00	.03	.12
	GRM	SM	0.5	.20	.23	.33	.28	.26	.26
			1.0	.19	.37	.33	.26	.25	.28
		GRM	0.5	.02	.08	.00	.00	.05	.06
			1.0	.02	.39	.00	.00	.04	.12

Table 5.3
Detection of changes in ability
using normal persons as a calibration sample

No. of Items	Generating Model	Estimation Model	δ	UB		UD		Lkd	
				False Alarms	Hits	False Alarms	Hits	False Alarms	Hits
10	SM	SM	0.5	.02	.28	.00	.18	.04	.21
			1.0	.02	.46	.00	.22	.04	.31
		GRM	0.5	.02	.06	.00	.00	.04	.06
			1.0	.02	.15	.00	.00	.04	.11
	GRM	SM	0.5	.03	.12	.04	.11	.08	.22
			1.0	.03	.14	.04	.15	.08	.24
		GRM	0.5	.02	.06	.00	.01	.04	.06
			1.0	.02	.13	.00	.01	.04	.10
20	SM	SM	0.5	.02	.42	.00	.22	.04	.26
			1.0	.02	.76	.00	.27	.04	.43
		GRM	0.5	.02	.12	.00	.00	.04	.06
			1.0	.02	.40	.00	.01	.04	.18
	GRM	SM	0.5	.13	.40	.13	.20	.18	.40
			1.0	.12	.60	.12	.27	.18	.51
		GRM	0.5	.02	.09	.00	.00	.04	.07
			1.0	.02	.33	.00	.01	.05	.15
30	SM	SM	0.5	.02	.55	.00	.24	.04	.29
			1.0	.02	.91	.00	.31	.04	.53
		GRM	0.5	.02	.17	.00	.01	.04	.07
			1.0	.02	.63	.00	.01	.04	.19
	GRM	SM	0.5	.21	.39	.34	.24	.27	.51
			1.0	.21	.60	.33	.30	.26	.63
		GRM	0.5	.02	.16	.00	.01	.05	.07
			1.0	.02	.51	.00	.01	.04	.19

Table 5.4
Detection of violation of local independence
using parameter estimates obtained combining normal and aberrant simulees

No. of Items	Generating Model	Estimation Model	δ	UB		UD		Lkd	
				False Alarms	Hits	False Alarms	Hits	False Alarms	Hits
10	SM	SM	0.2	.03	.11	.03	.14	.03	.13
			0.4	.02	.12	.03	.16	.03	.13
		GRM	0.2	.02	.04	.00	.08	.03	.35
			0.4	.02	.05	.00	.10	.05	.11
	GRM	SM	0.2	.03	.15	.03	.14	.03	.15
			0.4	.03	.14	.03	.13	.03	.15
		GRM	0.2	.02	.12	.00	.05	.03	.14
			0.4	.01	.11	.00	.05	.04	.16
20	SM	SM	0.2	.04	.21	.02	.23	.04	.19
			0.4	.03	.26	.03	.27	.03	.23
		GRM	0.2	.02	.13	.00	.12	.05	.13
			0.4	.02	.16	.00	.15	.05	.16
	GRM	SM	0.2	.04	.24	.02	.24	.03	.15
			0.4	.04	.25	.02	.23	.04	.14
		GRM	0.2	.02	.22	.00	.11	.05	.27
			0.4	.03	.21	.00	.13	.04	.25
30	SM	SM	0.2	.04	.34	.03	.31	.04	.24
			0.4	.03	.36	.02	.35	.04	.26
		GRM	0.2	.02	.22	.00	.28	.04	.19
			0.4	.02	.23	.00	.30	.05	.22
	GRM	SM	0.2	.04	.32	.03	.32	.04	.25
			0.4	.04	.32	.03	.32	.04	.25
		GRM	0.2	.02	.30	.00	.22	.05	.28
			0.4	.03	.31	.00	.33	.05	.33

Table 5.5
Detection of violation of local independence
using normal persons as a calibration sample

No. of Items	Generating Model	Estimation Model	δ	UB		UD		Lkd	
				False Alarms	Hits	False Alarms	Hits	False Alarms	Hits
10	SM	SM	0.2	.02	.59	.57	.00	.04	.50
			0.4	.02	.59	.54	.00	.04	.56
		GRM	0.2	.02	.33	.01	.00	.04	.34
			0.4	.02	.41	.00	.00	.05	.41
	GRM	SM	0.2	.02	.40	.40	.04	.08	.49
			0.4	.02	.55	.56	.04	.07	.55
		GRM	0.2	.02	.42	.00	.00	.03	.34
			0.4	.01	.41	.00	.00	.04	.36
20	SM	SM	0.2	.02	.66	.61	.00	.04	.42
			0.4	.02	.73	.68	.00	.04	.55
		GRM	0.2	.02	.43	.00	.00	.05	.43
			0.4	.02	.47	.01	.00	.04	.45
	GRM	SM	0.2	.08	.59	.58	.13	.17	.48
			0.4	.11	.59	.37	.09	.19	.46
		GRM	0.2	.02	.53	.01	.00	.04	.44
			0.4	.02	.58	.03	.00	.04	.48
30	SM	SM	0.2	.03	.80	.75	.00	.03	.80
			0.4	.03	.87	.81	.00	.03	.82
		GRM	0.2	.02	.46	.00	.00	.04	.56
			0.4	.02	.53	.00	.00	.03	.58
	GRM	SM	0.2	.04	.62	.02	.27	.04	.55
			0.4	.04	.62	.02	.20	.04	.55
		GRM	0.2	.02	.63	.00	.00	.04	.55
			0.4	.03	.64	.00	.00	.05	.56

for generating the data. The estimation model is given in the third column labelled 'Estimation Model'. The columns labeled 'False Alarms' pertain to the proportion of incorrectly flagged respondents in the sample of 900 non-aberrant simulees, the columns labeled 'Hits' refer to the proportion of correctly flagged simulees in the sample of the 100 aberrant simulees. UB , UD and Lkd stands for LM tests for constancy of theta and violation of local independence and the log-likelihood test, respectively.

As in the study of Type I error rate, the false alarm rate of UD was much lower than the nominal significance level used that was 5% and the false alarm rate of Lkd and UB were relatively close to the nominal significance level though we again see that this is not the case for the generating model GRM with estimation model SM. The power of the test for changes in ability and violation of local independence were much higher for estimates obtained using normal simulees than combining normal and aberrant simulees. From inspection of Table 5.2 we see that the power to detect changes of ability of the UB-test becomes reasonable for test-lengths of 20 and 30 items combined with an effect size of 1.0. If we compare the result for 20 items with the analogous results for the LM-test in Table 3.3 in Chapter 3, we see that the power in the Bayesian framework is slightly less than the power in the likelihood-based framework. Interestingly, also the finding that the power for the GRM is slightly less than the power for the SM is replicated here. In Table 5.3, it can be seen that the power increases when the aberrant simulees are not included in the calibration sample. The results with respect to the detection of violations of local independence are analogous the results with respect to detection of shifts in ability. When we compare the power with the power in a frequentist framework, such as is displayed in Table 2.9 in Chapter 2, we see that the power is slightly less in a Bayesian framework. Also in this case, as can be seen in Table 5.5, the power increases when no aberrant simulees are present in the calibration sample.

5.4 Discussion

This study investigated person-fit assessment using the tests UB , UD and l. The approach was generalized to the Bayesian framework and was applied to a specific IRT model, that is, the graded response model and the sequential model. Whereas procedures for conventional frequentist statisti-

cal inference focus on point estimates and their standard errors, Bayesian methods seek to characterize the posterior distribution of the parameters. This can done by an MCMC method, which produces samples from the joint posterior density of model parameters that may be then summarized to estimate θ (Jackman, 2000).

Type I error rates and the detection rates of three person-fit measures were compared. Simulations showed that, except for UD , all person-fit measures have reasonable Type I error rates. We found that Type I error rates were substantially greater than the nominal significance level when we generate under the GRM and estimate with SM. Furthermore, the detection rate largely depended on the number items, the effect size, the estimation procedure and the type of aberrant response behavior. Person-fit measures UB and l stood out with respect to detection rates. Thus, a general preference for person-fit measure in this case would be either the UB or l. As shown by the results in Glas and Meijer (2003), the detection rates for UB and l in the case of dichotomous items were non-existent.

Overall, the power is not high. The power of the test for detection of changes in ability was highest for k=30 and δ=1.0 and the power of the test for detection of local independence was highest for k=30 and δ=0.4. Moreover, we can only detect the most severe cases and depending on the applicability this is not too serious.

6
Summary

Person-fit analysis is used to identify item-score patterns on a test that are unlikely given a test model assumed to describe the data. Evaluation of Person-fit seeks to identify persons with aberrant response patterns, for example caused by 'sleeping' or 'fumbling' behavior, where a person of high ability answers easy items incorrectly or 'guessing' behavior where a person of low ability correctly answers more difficult items.

This thesis is concerned with person-fit analysis in the context of polytomous item responses. Most existing person-fit statistics have been proposed in the context of dichotomous item responses. In Chapter 2, a general class of person fit tests is presented for IRT models for polytomous items (with dichotomous items as a special case) based on the Lagrange multiplier (LM) test. The tests take the effects of parameter estimation into account and can be considered an alternative to the approach proposed by Snijders (2001). Simulation studies are conducted to investigate the Type I error rates of three types of tests: (1) naive tests that do not take estimation effects into account, (2) tests that take the effects of ability estimation into account, and (3) tests that take both the effects of estimation of the item and person parameters into account. Furthermore, Type I error rate and power of the tests are studied in the framework of polytomous items. Results showed that naive test statistics that ignore the effects of estima-

tion of the ability parameters result in incorrect Type I error rates and a marked decrease of power. Incorporating a correction to account for the effects of estimation of the ability parameters results in acceptable Type I error rates and power. Incorporating a correction for the estimation of the item parameters had very little additional effect.

In Chapter 3, we generalized the approach presented in Chapter 2 to the Generalized partial credit model (GPCM) and two alternatives for the GPCM: the sequential model (SM, Tutz, 1990) and the graded response model (GRM, Samejima, 1969). A general formulation for the person fit tests for the three models is presented for between-item multidimensionality. Although the rationales underlying the models are different, the models are hard to distinguish because their item-category response curves are similar (Verhelst, Glas, & de Vries, 1997). Therefore, it was investigated whether the three models can be distinguished using person-fit tests. The first simulation study pertaining to the unidimensional version of the models resulted in an acceptable Type I error rate even when a "wrong model" (a model not used to generate the data) was used. The power of the tests to detect model violations of constancy of the latent trait and local independence was reasonable. Furthermore, different results were found for the three models. In the second simulation study, using the multidimensional version of the models, the special case of between-items multidimensionality was considered. Results showed that using auxiliary information from other subscales did not increase the power of the tests. Furthermore, results showed that there was a main effect of the effect size for the model violations against constancy of the latent trait and local independence. Data from the NEO PI-R were used to get an impression of the degree of agreement between the three IRT models in a real testing situation. Results based on the Kappa coefficient showed that the degree of agreement betweeen the three models was moderate. The degree of agreement between the unidimensional and multidimensional versions of the models was highest for GPCM.

In Chapter 4, the multidimensional version of the person fit test proposed in Chapter 3 was applied to a personality inventory and a cognitive test to illustrate how we can interpret response patterns that are classified as fitting or misfitting. Because aberrant response behavior may have multiple causes and hence may be difficult to interpret, it is important to study score patterns that are classified as aberrant in more detail. In this

chapter we show that person-fit test statistics may add information to a person's estimated latent trait value or total score.

In Chapter 5, the person fit tests introduced in Chapter 3 are generalized to a Bayesian framework. Compared to the traditional frequentist approach, this Bayesian approach has several advantages. First, there is no need to derive the theoretical sampling distribution of the statistics, which sometimes may be very difficult, if not impossible. Second, the person-fit statistic may depend on unknown quantities such as the item and person parameters. This uncertainty is explicitly taken into account. A simulation study was presented to examine the effectiveness of several person-fit statistics using different estimation methods, person-fit statistics, model violations and test lengths. In this simulation study, all person-fit measures had reasonable Type I error rates except for UD which had a low Type I error rate. Further, the detection rate largely depended on the number items, the effect size, the estimation procedure and the type of aberrant response behavior. The three models make no difference but there are serious exception where the Type I error rate is grossly inflated. Therefore, classification decisions cannot be justified if the model does not fit. In general, the power of the Bayesian version of the test was comparable with the power of the frequentist version of the test.

7
Samenvatting (Summary in Dutch)

In dit proefschrift staat het identificeren en analyseren van niet-passende response patronen op een psychologische of onderwijskundige test of vragenlijst centraal. In het Engels wordt deze analyse aangeduid als person-fit analyse. In het vervolg houden we deze terminologie aan. Het doel van person fit is om antwoordpatronen die onwaarschijnlijk zijn gegeven het veronderstelde item response model te detecteren. Mogelijke verklaringen voor het niet passen van een response patroon bij het veronderstelde model zijn bijvoorbeeld verschillende vormen van bedrog zoals afkijken of voorkennis. Een ander verklaring is gisgedrag op een aantal of bijna alle items. In de literatuur zijn verschillende statistische toetsen voorgesteld om afwijkende response patronen te detecteren. De meeste tests zijn gebaseerd op dichotome (0/1) data. In dit proefschrift staan polytome data centraal.

In hoofdstuk 2 wordt een algemene klasse van person fit toetsen gepresenteerd die zijn gebaseerd op Lagrange multiplier toetsen. Deze toetsen houden rekening met het effect van het schatten van de parameters en vormen een alternatief voor de benadering zoals voorgesteld door Snijders (2001). Door middel van simulatie studies wordt de Type I fout onderzocht van drie verschillende toetsen: (1) naïeve toetsen die geen rekening houden met het schatten van de parameters, (2) toetsen die rekening houden met het schatten van de vaardigheidsparameter en (3) toetsen die

rekening houden met het schatten van de item en vaardigheidsparameters. Resultaten laten zien dat naïeve toetsen die geen rekening houden met het schatten van de vaardigheidsparameters resulteerde in een incorrecte Type I fout en een geringer onderscheidend vermogen. Een correctie voor het schatten van de vaardigheidsparameter leidde tot een acceptabele Type I fout, terwijl een correctie voor het schatten van de item parameters weinig effect had.

In Hoofdstuk 3 wordt de benadering zoals gevolgd in Hoofdstuk 2 gegeneraliseerd naar het "generalized partial credit model" (GPCM) en twee alternatieven voor het GPCM: het "sequential model" (Tutz, 1990) en het "graded response model" (GRM Samejima, 1969). Er wordt een algemene formulering gegeven voor de drie modellen voor "between item" multidimensionaliteit. Hoewel de rationale voor de drie modellen verschillend is, zijn de modellen moeilijk van elkaar te onderscheiden omdat de categorie response curves eenzelfde vorm hebben. Er wordt daarom door middel van person fit toetsen onderzocht of de responsepatronen van elkaar te onderscheiden zijn voor de verschillende modellen. De eerste simulatiestudie die betrekking had op de eendimensionele versie van de modellen resulteerde in een acceptabele Type I fout, zelfs wanneer een model werd gebruikt dat niet was gebruikt om de data te simuleren. Het onderscheidend vermogen was acceptabel, hoewel er verschillende resultaten werden gevonden voor de verschillende modellen. In de tweede simulatiestudie werd een multidimensionele versie van de modellen gebruikt waarbij additionele informatie van andere subschalen werd gebruikt. De resultaten laten zien dat het onderscheidend vermogen niet werd beïnvloed door de correlatie tussen de subschalen. Bovendien laten de resultaten zien dat er een hoofdeffect is van modelschendingen tegen de assumptie van een constante latente trek en locale onafhankelijkheid. Veder werden data van de NEO-PI-R gebruikt om een idee te krijgen van de mate van overeenkomst tussen de verschillende modellen bij empirische data. De mate van overeenkomst tussen de verschillende modellen was matig. De mate van overeenkomst tussen de eendimensionele en multidimensionele versies van de modellen was het grootst voor het GPCM.

In hoofdstuk 4 wordt een multidimensionele versie van de person fit toetsen gepresenteerd zoals voorgesteld in hoofdstuk 3 en deze toetsen worden toegepast op persoonlijkheidsdata en cognitieve data om het gebruik van de toetsen te illustreren. De oorzaak van de afwijkende scorepatronen

kan verschillen en daarom is het nuttig om de patronen in meer detail te bestuderen. Hoofdstuk 4 laat zien dat person fit toetsen nuttig kunnen zijn om de soort van afwijkendheid verder te onderzoeken.

In hoofdstuk 5 wordt een Bayesiaanse benadering gepresenteerd voor de person fit toetsen die zijn gepresenteerd in Hoofdstuk 3. Deze Bayesiaanse benadering heeft verschillende voordelen: (1) Er hoeft geen theoretische steekproevenverdeling te worden afgeleid, hetgeen soms erg lastig is; (2) de onzekerheid waarmee de item en persoonsparameters worden geschat wordt expliciet meegenomen. Door middel van een simulatiestudie wordt het onderscheidend vermogen onderzocht voor de verschillende toesten, verschillende schattingsmethoden en verschillende testkenmerken. Het onderscheidend vermogen bleek gevoelig te zijn voor het aantal items in de test en de soort van afwijkendheid die werd gesimuleerd. In het algemeen waren de resultaten vergelijkbaar voor de verschillende modellen die werden gebruikt. De power van de Bayesiaanse versie van de toetsen was vergelijkbaar dan de frequentistische versie van de toetsen.

References

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

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

[3] Andersen, E.B. (1973). A goodness of for test for the Rasch model. *Psychometrika, 38,* 123-140.

[4] Andersen, E.B. (1977). Sufficient statistics and latent trait models. *Psychometrika, 42,* 69-81.

[5] Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-680.

[6] Andrich, D. (1988). *Rasch models for Measurement.* Sage University Paper Series on Quantitative Applications in the Social Sciences 07-068. Beverly Hills: Sage Publications.

[7] Baker, F. B. (1998). An investigation of item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement 22*, 153-169.

[8] Becher, T. M., Verstralen, H. H. F. M., & Verhelst, N. D. (2002). Equivalent linear logistic test models. *Psychometrika, 67,* 123-136.

[9] Béguin, A. A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 541-562.

[10] Bleichrodt, N., Drenth, P.J.D., Zaal, J.N. & Resing, W.C.M. (1984). Revisie Amsterdamse Kinder Intelligentietest. Lisse: Swets & Zeitlinger.

[11] Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37,* 29-51.

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

[13] Bock, R.D., Gibbons, R.D., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement 12,* 261-280.

[14] Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

[15] Buse, A. (1982). The likelihood ratio, Wald, and Lagrange multiplier tests: an expository note. *The American Statistician, 36,* 153-157.

[16] Choca, J.P., Shanely, L.A.& and Van Denburg, E. (1992). *Interpretive guide to the Millon Clinical Multiaxial Inventory.* Washington DC: American Psychological Association.

[17] Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: *The NEO Personality Inventory. Psychological Assessment*, *4*, 5-13.

[18] Costa, P.T., Jr. & McCrae, R.R. (1992a). Four ways five factors are basic. *Personality and Individual Differences*, *13*, 653-665.

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

[20] Drasgow, F., & Levine, M. V. (1986). Optimal detection of certain forms of inappropriate test scores. *Applied Psychological Measurement, 10*, 59-67.

[21] Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Detecting inappropriate test scores with optimal and practical appropriateness indices. *Applied Psychological Measurement, 11*, 59-79.

[22] Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1991). Appropriateness measurement for some multidimensional test batteries. *Applied Psychological Measurement, 15*, 171-191.

[23] 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.

[24] Emons, W. H. M., Sijtsma, K., & Meijer, R.R, (2005). Global, local, and graphical person-fit analysis using person-response functions.*Psychological Methods*, *10*, 101-119.

[25] Fischer, G.H., & Scheiblechner, H.H. (1970). Algorithmen und Programme für das probabilistische Testmodell von Rasch. *Psychologische Beiträge, 12,* 23-51.

[26] Fox, J.P., & Glas, C.A.W. (2001). Bayesian Estimation of a Multilevel IRT Model using Gibbs Sampling. *Psychometrika 66*, 271-288.

[27] Gelfand, A. E., & Smith, A. F. M. (1990). Samplingbased approaches to calculating marginal densities. *Journal of the American Statistical Association 85*, 398-409.

[28] Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). Bayesian data analysis. *London: Chapman and Hall.*

[29] Gelman, A., Meng, X.-L., & Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733-807.

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

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

[32] Glas, C.A.W. (2001). Differential item functioning depending on general covariates. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory* (pp.131-148). New York, NJ: Springer.

[33] Glas, C.A.W. & Dagohoy, A.V.T.(in press). Person fit tests for IRT models for polytomous items with estimated person and item parameters. *Psychometrika.*

[34] Glas, C. A. W., & Meijer, R.R. (2003). A Bayesian Approach to Person Fit Analysis in Item Response Theory Models. *Applied Psychological Measurement, 27*, 217-233.

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

[36] Guttman, L. (1950). The basis for scalogram analysis. In S.A. Star, & J.A. Clausen (Eds.), *Measurement and prediction* (pp. 60-90). Princeton: Princeton University Press

[37] Hathaway, S.R. & McKinley, J.C. (1993). *The Minnesota Multiphasic Personality Inventory Manual.* New York: Psychological Corporation.

[38] Hoijtink, H., & Molenaar, I.W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs Sampler and posterior predictive checks. *Psychometrika, 62,* 171-189.

[39] Jackman, S. (2000). Estimation and inference via Bayesian simulation: An introduction to Markov Chain Monte Carlo. *American Journal of Political Science, 44,* 375-404.

[40] Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika, 51*, 357-373.

[41] Janssen, R., Tuerlinckx, F., Meulders, M. & de Boeck, P. (2000). A hierarchical IRT model for criterion-referenced measurement. *Journal od Educational and Behavioral Statistics, 25,* 285-306.

[42] Johnson, V.E., & Albert, J.H. (1999). *Ordinal data modeling.* New York, NJ: Springer.

[43] Kelderman, H. (1984). Loglinear RM tests. *Psychometrika,49,* 223-245.

[44] Kiefer, J., & Wolfowitz, J. (1956). Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Annals of Mathematical Statistics, 27,* 887-903.

[45] Kim, S.-H. (2001). An evaluation of a Markov Chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25,* 163-176.

[46] Lawley, D.N. (1943). On problems connected with item selection and test construction. *Proceedings of the Royal Society of Edinburgh, 61,* 273-287.

[47] Lawley, D.N. (1944). The factorial analysis of multiple test items. *Proceedings of the Royal Society of Edinburgh, 62-A,* 74-82.

[48] Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269-290.

[49] Li M.F., Olejnik S. (1997) The power of Rasch person-fit statistics in detecting unusual response patterns. *Applied Psychological Measurement 21*, 215-231.

[50] Lord, F.M. (1952). A theory of test scores. *Psychometric Monograph 7.*

[51] Lord, F.M. (1953a). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. *Psychometrika, 18,* 57-75.

[52] Lord, F.M. (1953b). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-548.

[53] Lord, F.M. and Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

[54] Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47,* 149-174.

[55] McDonald, R.P. (1997). Normal-ogive multidimensional model. In W.J.van der Linden and R.K.Hambleton (Eds.), *Handbook of modern item response theory.* (pp.257-269). New York, NJ: Springer.

[56] Meijer, R. R. (1994). The number of Guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement,* 18, 311-314.

[57] Meijer, R. R., Sijtsma, K., & Molenaar, I. W. (1995). Reliability estimation for single dichotomous items based on Mokken's IRT model. Applied Psychological Measurement, 19, 323-335.

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

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

[60] Mellenbergh, G.J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement, 19,* 91-100.

[61] Meng, X.L. (1994). Posterior predictive p-values. *Ann. Statist., 22,* 1142-1160.

[62] Mislevy, R.J. (1984). Estimating latent distributions. *Psychometrika, 49,* 359-381.

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

[64] Molenaar, I.W. (1983). Some improved diagnostics for failure in the Rasch model. *Psychometrika, 48,* 49-72.

[65] Molenaar, I.W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, 55, 75-106.

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

[67] Nering, M. L. (1995). The Distribution of person fit statistics using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121-129.

[68] Nering, M. L. (1996). The effects of person misfit in computersized adaptive testing (Doctoral dissertation, University of Minnesota). *Dissertation Abstracts International, 57*, 04B.

[69] Neyman, J., and Scott, E.L. (1948). Consistent estimates, based on partially consistent observations. *Econometrica, 16,* 1-32.

[70] Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement, 24*, 50-64.

[71] 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.

[72] 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.

[73] 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.

[74] Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research.

[75] Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J.van der Linden and R.K.Hambleton (Eds.), *Handbook of modern item response theory.* (pp.271-286). New York, NJ: Springer.

[76] Reise, S. P. (1995). Scoring method and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement*, *19*, 213-229.

[77] Reise, S.P., & Due A.M. (1991). The influence of test characteristics on the detection of aberrant response patterns. *Applied Psychological Measurement 15,* 217-226.

[78] Reise, S.P., & Waller, N.G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, 65, 143-151.

[79] Reise, S.P., & Waller, N.G. (1993). How Many IRT Parameters Does It Take to Model Psychopathology Items? *Psychological Methods*, 8, 164-184.

[80] Rigdon S.E., & Tsutakawa, R.K. (1983). Parameter estimation in latent trait models. *Psychometrika, 48,* 567-574.

[81] Rogers, H.J., & Hattie, J.A., (1987). A Monte Carlo investigation of several person and item fit statistics for item response models. *Applied Psychological Measurement, 11*, 47-57.

[82] Rubin, D.B., & Thomas, N. (2001). Using parameter expansion to improve the performance of the EM algorithm for multidimensional IRT population-survey models. In A. Boomsma, M. A.J. van Duijn, & T.A.B. Snijders (Eds.): *Essays on Item Response Theory* (pp.193-204). New York, NJ: Springer.

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

[84] Samejima, F. (1972). A general model for free response data. *Psychometrika, Monograph Supplement, No. 18.*

[85] Samejima, F. (1973). Homogeneous case of the continuous response model. *Psychometrika, 38,* 203-219.

[86] Smith, R. M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and psychological measurement*, *45*, 433-444.

[87] Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement*, *46*, 359-372.

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

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

[90] Tanner, M.A. (1993). *Tools for statistical inference.* New York, NJ: Springer.

[91] Tatsuoka, K. K., & Tatsouka, M.M. (1982). Detection of aberrant response patterns and their effect on dimensionality. *Journal of Educational Statistics, 7*, 215-231.

[92] Tatsuoka, K. K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95-110.

[93] Tellegen (1988). The analysis of consistency in personality assessment. J*ournal of Personality*, 56, 621-663.

[94] Thissen D. (1982). Marginal maximum likelihood estimation for the one-parameter logistic model.*Psychometrika, 47,* 175-186.

[95] Thissen, D., Chen, W-H., and Bock, R.D. (2003). *Multilog.* Lincolnwood, IL, Scientific Software International.

[96] Tsutakawa, R. K., & Johnson, J. C. (1990). The effect of uncertainty of item parameter-estimation on ability estimates. *Psychometrika, 55,* 371-390.

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

[98] van der Linden, W. J., & Hambleton, R. K. (Eds.) (1997). *Handbook of Modern Item Response Theory.* New York, NJ: Springer Verlag.

[99] van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23,* 327-345.

[100] van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (2002). Detection of person misfit in computerized adaptive tests with polytomous items. *Applied Psychological Measurement, 26,* 164-180.

[101] Verhelst, N.D., Glas, C.A.W., & de Vries, H.H. (1997). A steps mopdel to analyze partial credit. In W.J.van der Linden and R.K.Hambleton (Eds.), *Handbook of modern item response theory.* (pp.123-138). New York, NJ: Springer.

[102] Wainer, H, Bradlow, E.T., & Du, Z. (2000). Testlet Response Theory: an Analogue for the 3-PL Useful in Testlet-Based Adaptive Testing. In W.J. van der Linden & C.A.W.Glas (Eds.), *Computer Adaptive Testing: Theory and Practice* (pp. 245-269). Boston: Kluwer-Nijhoff Publishing.

[103] Wilson, M., & Masters, G.N. (1993). The partial credit model and null categories. *Psychometrika, 58,* 85-99.

[104] Wright, B.D., & Linacre, J.M. (1992). *BIGSTEPS.* (Computer Software). Chicago, IL: MESA Press.

[105] Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis.* (Computer Software). Chicago, IL: Mesa Press.

[106] Wright, B.D., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological Measurement, 29,* 23-48.

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

[108] Yen, W.M. (1981). Using simultaneous results to choose a latent trait model. *Applied Psychological Measurement, 5,* 245-262.

[109] Yen, W.M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125-145.

[110] Zickar, M. J., & Drasgow, F. (1996). Detecting faking on a personality instrument using appropriateness measurement. *Applied Psychological Measurement, 20*, 71-87.

[111] Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *Bilog MG: Multiple-Group IRT Analysis and Test Maintenance for Binary Items.* Chicago, IL: Scientific Software International, Inc.