IRT Model Fit from Different Perspectives

Muhammad Naveed Khalid

IRT Model Fit from Different Perspectives

Muhammad Naveed Khalid

INVITATION

You are
cordially invited
to attend the
public defense
of my dissertation
on **Wednesday**
9 DECMBER 2009
at 16.45

[**An introduction**
will start at **16.25**]

in the

SPIEGEL Building,
Collegezaal 2
University of Twente

A **reception** will
follow after the defense

Muhammad Naveed
Khalid

m.n.khalid@utwente.nl
Office: 053-489 3864
Address: Jupiterstraat 55
7521 JJ Enschede

IRT Model Fit from Different Perspectives

Muhammad Naveed Khalid

Promotiecommissie

Promotor
Prof. dr. Cees A.W. Glas

Assistent Promotor
Dr. ir. B.P. Veldkamp

Overige leden
Prof. dr. M.P.F. Berger
Prof. dr. T.J.H.M. Eggen
Prof. dr. R.R. Meijer
Prof. dr. W.B. Verwey

IRT Model Fit from Different Prespectives
M. N. Khalid
Ph.D. thesis
University of Twente
The Netherlands
9 December, 2009

ISBN: 978-90-365-2900-6

IRT Model Fit from Different Perspectives

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday, 9 December, 2009 at 16.45

by

Muhammad Naveed Khalid
born on 28 June, 1977
in Narowal, Pakistan

This dissertation is approved by the following promotores:

Promotor: Prof. dr. Cees A.W. Glas
Assistant Promotor: dr. ir. B.P. Veldkamp

To My Parents

Acknowledgment

Firstly, the researcher owes to the Almighty Allah, the most Merciful and the most Beneficent, Who gave me courage to do this work. This piece of work will never be accomplished without our Allah Almighty with His blessings and His power that work within me and also without the people behind my life for inspiring, guiding and accompanying me through thick and thin.

The presentation of this dissertation is an indicator that an important milestone in my scientific journey has been reached. Now, it is a great opportunity for me to look back on my achievement at the University of Twente and to express my gratitude to all those people who guided, supported and stayed beside me during my PhD study.

First of all, I would like to thank Prof. Dr. Cees A.W. Glas. It has been a great privilege to be under the guidance with such knowledgeable, inspiring, and patient supervisor. His motivation, enthusiasm, insight throughout the research, sympathetic attitude and real help always inspired me to push myself to reach more in science. I appreciated Dr. Bernard Veldkamp for his corrections, help and guidance during my last year of my Ph.D study.

I have very much appreciated the opportunity to complete my Ph.D study in OMD department. Thanks go to all faculty members for their suggestions, comments, and assistance in the development and completion of thesis. Technical guidance especially in the eve of data analysis using SPSS of Wim Tielen is highly acknowledged. I would like to thank all my colleagues for their support and friendly atmosphere in that department, especially: Khurrem Jehangir for his suggestions, criticism, and comments on earlier drafts of thesis; to Hanneke Geerlings for her support to solve bugs in Fortran programmes; to Anke Weekers, Rinke Klein Entink, Iris Egberink and all other Ph.D fellows for help, outings and conversations we had; also I thank secretaries of OMD department Birgit and Lorette for their help.

The generous support of Higher Education Commission (HEC) of Pakistan is greatly appreciated. The researcher also expresses his gratitude to fellows Ch. Shahzad Ahmad Arfan (Late), Mr. Naeem Abdul Rehman, Mr. Muhammad Nawaz Balouch, Mr. Muhammad Asif and Mr. Muhammad Akram Raza for their company, and friendship.

I would like to extend my heartfelt gratitude to all family members especially, Zahid Mehmood, who always gave motivation and encouragement for higher education; to Muhammad Waheed, Muhammad Fareed, my nephews Muhammad Jawwad, Muhammad Uzair, and nieces Hifsa, Hamna, and Fiza. Finally, I would like to dedicate this dissertation to my Parents. Their love and support always cheer me up and motivate me to finish this work. I am very grateful for their support and patience.

Muhammad Naveed Khalid
Enschede, 9 Decmber, 2009

Contents

List of Tables xii

List of Figures xiv

1 Introduction 1

1.1 Overview of the Thesis 6

2 A Step-Wise Method for Evaluation of Differential Item Functioning 9

2.1 Introduction 10

2.2 Detection and Modeling of DIF 12

2.2.1 IRT Models 13

2.2.2 MML Estimation 15

2.2.3 A Lagrange Multiplier Test for DIF 17

2.3 An Empirical Example 19

2.4 Design of Simulation Studies 21

2.5 Results 22

2.5.1 Type I Error Rates 22

2.5.2 Power of the Test 23

2.5.3 DIF and Population Parameters 24

2.5.4 Non-uniform DIF 27

2.6 Discussion and Conclusions 28

3 Assessing Item Fit: A Comparative Study of Frequentist and Bayesian Frameworks 31

3.1 Introduction 32

3.2 Fit to IRT Models 35

3.2.1 Differential Item Functioning 36

3.2.2 Fit of ICC 37

3.2.3 Local Independence 37

3.3 Lagrange Multiplier (LM) Test 38

3.4 Posterior Predictive Checks 40

3.5 Tests for Fit to IRT Models 42

3.6 Simulation Design 45

3.7 Results 45

3.7.1 Differential Item Functioning 45

3.7.2 Local Independence 48
3.7.3 Shape of the ICCs 49
3.8 An Empirical Example 52
3.9 Conclusions 58

4 Assessing Person Fit: A Comparative Study of Frequentist and Bayesian Frameworks 61
4.1 Introduction 62
4.2 Person Fit Statistics 64
4.2.1 IRT Models 64
4.2.2 W Statistic 65
4.2.3 UB Statistic 65
4.2.4 ξ_1 Statistic 66
4.2.5 ξ_2 Statistic 66
4.2.6 l_z Statistic 67
4.2.7 LM Statistic 67
4.3 Snijders’ Correction Procedure 67
4.4 An LM Test for Constancy of Theta 69
4.5 Posterior Predictive Checks 71
4.6 Simulation Design 73
4.7 Results 74
4.7.1 Type I Error Rate 74
4.7.2 Detection Rates for Guessing 75
4.7.3 Detection Rates for Item Disclosure 79
4.8 An Empirical Example 81
4.9 Discussion and Conclusions 84
4.A The Null Distributions of PFS 87

5 A Comparison of Top-down and Bottom-up approaches in the Identification of Differential Item Functioning using Confirmatory Factor Analysis 91
5.1 Introduction 92
5.2 CFA Based DIF Procedure 95
5.2.1 Confirmatory Factor Analysis 95
5.2.2 DIF Procedure 96
5.3 Study Design 97
5.4 Results 99
5.4.1 2PL 99
5.4.2 3PL 102

5.4.3 GRM 104
5.4.4 Multidimensional (between) Models 106
5.4.5 Multidimensional (within) Models 107
5.5 An Empirical Example 109
5.6 Discussion and Conclusions 110

Summary **113**

References **117**

List of Tables

2.1 The results of LM test to evaluate fit of DIF. 21
2.2 The type I error rates under the 2PLM. 23
2.3 The power of test under the 2PLM. 24
2.4 The power of test under the 3PLM. 25
2.5 Estimates of the mean of the ability distribution in the different steps of the purifications procedure (K = 20). 26
2.6 A comparison of the purification process using the LM tests for uniform and non-uniform DIF. 29
3.1 The power and type I error (DIF) in the 2-PNO model. 46
3.2 The power and type I error (DIF) in the 3-PNO model. 47
3.3 The power and type I error (LOC) in the 2-PNO model. 49
3.4 The power and type I error (LOC) in the 3-PNO model. 50
3.5 The power and type I error (ICC) in the 2-PNO model. 51
3.6 The power and type I error (ICC) in the 2-PNO model. 52
3.7 Outcomes of LM tests for ICCs for examination data. 53
3.8 Outcomes of LM tests for ICCs for examination data combined with linking group data. 55
3.9 Outcomes of LM tests for local independence for examination data. 57
3.10 Outcomes of LM tests for local independence for examination data combined with linking group data. 59
4.1 Type I error rates under the bayesian and frequentist frameworks. 76
4.2 Detection rates for guessing simulees under the bayesian and frequentist frameworks. 78
4.3 Detection rates for item disclosure simulees under the bayesian and frequentist frameworks. 80
4.4 Detection rate of fit statistics in percentages. 82
5.1 The power under the 2PL model. 100
5.2 The type 1 error rates under the 2PL model. 100
5.3 The power under the 3PL model. 103
5.4 The type 1 error rates under the 3PL model. 103
5.5 The power under the GRM model. 105
5.6 The type 1 error rates under the GRM model. 105
5.7 The power under the multidimensional (between) model. 106
5.8 The type 1 error rates under the multidimensional (between) model. 107

5.9 The power under multidimensional (within) model. 108
5.10 The type 1 error rates under multidimensional (within) model. 108
5.11 The evaluation of DIF in both approaches. 109

List of Figures

2.1 Response functions and expected item-total score under the GPCM. 15

2.2 Change in the estimates of the means of the ability distribution over iterations. 20

2.3 Distribution of significance probabilities of UB in the bayesian framework. 82

2.4 Distribution of significance probabilities of UB in the frequentist framework. 83

2.5 Distribution of significance probabilities of LM in the frequentist framework. 83

1

Introduction

Item response theory (IRT) provides a framework for modeling and analyzing item response data. The advantages of item response theory over classical test theory for analyzing mental test data are well documented (see, for example, Lord, 1980; Hambleton & Swaminathan, 1985; Hambleton et al., 1991). IRT postulates that an examinee's performance on a test depends on a set of unobservable "latent traits" that characterize the examinee. An examinee's observed score on an item is regressed on the latent traits. The resulting regression model, termed an item response model, specifies the relationship between the item response and the latent traits, with the coefficients of the model corresponding to parameters that characterize the item. It is this item-level modeling that gives IRT its advantages over classical test theory.

IRT is based on strong mathematical and statistical assumptions, and only when these assumptions are met, at least to a reasonable degree, can item response theory methods be implemented effectively for analyzing educational and psychological test data and for drawing inferences about properties of the tests and the performance of individuals. Checking model assumptions and assessing the fit of models to data are routine in statistical endeavors. In regression analysis, numerous procedures are available for checking distributional assumptions, examining outliers, and the fit of the chosen model to data. Some of these procedures have been adapted for use in IRT. The basic problem in IRT is that the regressor, θ, is unobservable; this fact introduces a level of complexity that renders a procedure which is straightforward in regression analysis inapplicable in the IRT context.

IRT models are based on a number of explicit assumptions, so methods for the evaluation of model fit focus on these assumptions. Model fit can be viewed from two perspectives: the items and the respondents. In the first case, for every item, residuals (differences between predictions from the estimated model and observations) and item fit statistics are computed to assess whether the item violates the model. The item fit statistics can be used for evaluating the fit of the so-called manifest IRT model (Holland, 1990). This class comprised statistics developed to be sensitive to specific model violations (Andersen, 1973, Glas, 1988, 1997, 1998, 1999, 2005, Glas & Verhelst, 1989, 1995a, 1995b; Glas & Suárez-Falcón, 2003; Holland & Rosenbaum, 1986; Kelderman, 1984, 1989; Martin Löf, 1973, Mokken, 1971; Molenaar, 1983; Sijtsma & Meijer, 1992; Stout, 1987, 1990). In the second case, residuals and person fit statistics are computed for every person to assess whether the responses to the items follow the model. The person fit statistics focuses on the appropriateness of the stochastic model on the level of the individual. For this reason they are commonly called person-fit statistics. In the IRT context several person fit statistics have been proposed that can be used to detect individual item score patterns that do not fit the IRT model (Drasgow, Levine, & Williams, 1985; Glas & Meijer, 2003; Meijer & van Krimpen-Stoop, 1999; Klauer, 1995; Molenaar & Hoijtink, 1990; Meijer, Molenaar & Sijtsma, 1994; Reise, 1995; Glas & Dagohoy, 2007).

IRT models can be evaluated by Pearson-type statistics, that is, statistics based on the difference between observations and their expectations under the null-model. However, these tests are rather global and give no information with respect to specific model violations. Most of the goodness-of-fit tests that have been proposed in literature are assumed to have an asymptotic χ^2 distribution (Orlando & Thissen, 2000). Glas and Suárez-Falcón (2003) have pointed out that the asymptotic distributions of these χ^2 statistics are not known or questionable. These issues were addressed using LM tests sketched by Glas (1998, 1999) in the marginal maximum likelihood (MML) framework (see, for instance, Bock & Aitkin, 1981; Mislevy, 1986). A related important issue is that fit statistics are sensitive to sample size. They tend to reject the model even for moderate sample sizes. Another issue is the effect size, that is, the severity of the model violation. A related problem is defining a stopping rule for the searching procedure for misfitting items. Effect size use is important to assist with the avoidance of practically trivial but statistically significant results, regardless of the detection method.

However, frequentist estimation methods, as for instance, MML are complicated (in terms of computation) for multilevel and multidimensional psychometric models (Fox & Glas, 2001; Béguin &

Glas, 2001) due to the complex dependency structures of the models. They require the evaluation of multiple integrals to solve the estimation equations for parameters. Further, in the estimations of person parameters, they fail to take into account the uncertainty about item parameters when making inference on examinees' ability parameters. They may underestimate the uncertainty in abilities (Tsutakawa & Soltys, 1988; Tsutakawa & Johnson, 1990).

The above stated problems are avoided in a fully Bayesian framework and now a days it's being widely used for parameter estimation in complex psychometric IRT models. When comparing the fully Bayesian framework with the MML framework the following considerations play a role. First, a fully Bayesian procedure supports definition of a full probability model for quantifying uncertainty in statistical inferences (see, for instance, Gelman, Carlin, Stern, & Rubin, 2004, p. 3). This does involve the definition of priors, which creates some degree of bias, but this can be minimized using of non-informative priors. Second, estimates of model parameters that might otherwise be poorly determined by the data can be enhanced by imposing restrictions on these parameters via their prior distributions. However this can also be done in a Bayes modal framework, which is closely related to the MML framework (Mislevy, 1986). Baker (1998) has investigated the recovery of item parameter estimates using Gibbs sampling (Albert, 1992) and BILOG (Mislevy & Bock, 1989). The item parameter recovery characteristics were comparable for the largest dataset of 50 items and 500 examinees. However, for short tests and sample sizes the item parameter recovery characteristics of BILOG were superior to those of the Gibbs sampling approach.

Recently Sinharay (2005), Sinharay, Johnson and Stern (2006) have applied a popular Bayesian approach posterior predictive checks (PPCs) for the assessment of model violations in unidimensioal IRT models. However, Bayarri and Berger (2000) have showed that PPCs also comes with problems due to twice use of data as a result PPP-values were conservative (i.e., often failed to detect model misfit) and inadequate behavior of posterior p-value. One of the research questions was to compare the frequentist procedures (LM test) and Bayesian procedures (PPCs) for evaluating of item fit in IRT models.

One of the difficulties in the assessment of person fit is the fact that the ability parameter of the examinee is unknown. The literature contains many proposed person fit statistics which are functions of the unknown ability parameter. The use of an estimated rather than the true value of the ability parameter has an effect on the distribution of the person fit statistic Snijders (2001). These estimates usually decrease the asymptotic variance of most statistics proposed in the literature. Therefore, their asymptotic distribution is usually unknown. To address this issue, Snijders (2001) proposed a method for standardization of a specific class of person fit statistics for dichotomous items, such that their asymptotic distribution can be properly derived. Snijders (2001) applied a correction for the mean and variance of l_z and derived an asymptotic normal approximation for the conditional distributions of l_z when $\hat{\theta}$ is used in its calculation. Glas and Dagohoy (2007) proposed an alternative approach which is based on LM test. The one of major question that was to describe and compare the person fit statistics that take in account Snijders' correction, LM tests, and Bayesian procedures (PPCs) for evaluating goodness of fit in the unidimensional dichotomous IRT models.

In recent years, confirmatory factor analytic (CFA) techniques have become the most common method of testing for measurement equivalence/invariance (ME/I) as alternative for IRT based procedures. McDonald (1999) and Raju et al., (2002) have provide a comprehensive review of the methodological similarities and differences among the CFA and IRT. When conducting DIF studies using the CFA and IRT, there is a variation in the way nested models are constructed. Unlike Bottom-up approach, where the baseline model is typically one in which all parameters except the referent are free to vary and an item is studied by additionally constraining its parameters to be equal across groups, many IRT researchers use the opposite approach which is, a Top-down procedure. Specifically, a baseline model is constructed by constraining the parameters for all items to be equal across groups and a series of augmented models is formed by freeing the parameters for the studied item(s), one at a time, and examining the changes in G^2 (e.g., Thissen, 1991; Bolt, 2002). When this approach is used, it is not necessary to specify a referent item for identifying the metric, because, in each comparison, all items except the studied item are constrained. However, it is necessary to anchor the metric by choosing a reference group whose latent mean is set to zero for parameter estimation.

The Bottom-up approach follows is motivated by following argument. According to statistical theory (Maydeu-Olivares & Cai, 2006) for the difference between a baseline model and constrained models to follow a central chi-square distribution under the null hypothesis, the baseline model has to fit the data. If the baseline model contains a number of DIF items, then it might not fit adequately and DIF detection could be adversely affected. Thus, from a statistical standpoint, the approach to comparing nested models in Bottom-up approach is theoretically appropriate, whereas the traditional Top-down approach implemented in the IRT LR approach is not. Given the current variation in the way nested models are constructed (Top-down and Bottom-up), the other research question was to make a comprehensive comparative simulation study and explored the factors that have impact on their performance.

1.1 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, item bias or differential item functioning (DIF) is seen as a lack of fit to an IRT model. It is shown that inferences about the presence and importance of DIF can only be made if DIF is sufficiently modeled. This requires a process of so-called test purification where items with DIF are identified using statistical tests and DIF is modeled using group-specific item parameters. In the present study, DIF is identified using a Lagrange multiplier statistic. The first problem addressed is that the dependency of these statistics might cause problems in the presence of relatively large number DIF items. However, simulation studies show that the power and Type I error rate of a step wise procedure where DIF items are identified one at a time are good. The second problem pertains to the importance of DIF, i.e., the effect size, and related problem of defining a stopping rule for the searching procedure. Simulations show that the importance of DIF and the stopping rule can be based on the estimate of the difference between the means of the ability distributions of the studied groups of respondents. The searching procedure is stopped when the change in this effect size becomes negligible.

Chapter 3 presents the measures for evaluating the most important assumptions underlying unidimensional item response models such as subpopulation invariance, form of item response function, and local stochastic independence. These item fit statistics are studied in two frameworks. In a frequentist MML framework, LM tests for model fit based on residuals are studied. In the framework of LM model tests, the alternative hypothesis clarifies which assumptions are exactly targeted by the residuals. The alternative framework is the Bayesian one. The PPCs is a much used Bayesian model checking tool because it has an intuitive appeal, and is simple to apply. A number of simulation studies are presented that assess the Type I error rates and the power of the proposed item fit tests of in both frameworks. Overall, the LM statistic performs better in terms of power and Type I error rates.

Chapter 4 presents fit statistics that are used for evaluating the degree of fit between the chosen psychometric model and an examinee's item score pattern. Person fit statistic reflects the extent to which the examinee has answered test questions according to the assumptions and description of the model. Frequentist tests as the LM test and tests with Snijders' correction (which take into account the estimation of ability parameter) are compared with PPCs. Simulation studies are carried out using a number of fit statistics in a number of combinations in both frameworks.

In Chapter 5, a method based on structural equation modelling (or, more specifically, confirmatory factor analysis) for examining measurement equivalence is presented. Top-down and Bottom-up approaches were evaluated for constructing nested models. A comprehensive comparative simulation study is carried out to explore the factors that have impact on performance for detecting DIF items.

2

A Step-Wise Method for Evaluation of Differential Item Functioning

Abstract: Item bias or differential item functioning (DIF) has an important impact on the fairness of psychological and educational testing. In this paper, DIF is seen as a lack of fit to an item response (IRT) model. Inferences about the presence and importance of DIF require a process of so-called test purification where items with DIF are identified using statistical tests and DIF is modeled using group-specific item parameters. In the present study, DIF is identified using item-oriented Lagrange multiplier statistics. The first problem addressed is that the dependence of these statistics might cause problems in the presence of a relatively large number of DIF items. A stepwise procedure is proposed where DIF items are identified one or two at a time. Simulation studies are presented to illustrate the power and Type I error rate of the procedure. The second problem pertains to the importance of DIF, i.e., the effect size, and related problem of defining a stopping rule for the searching procedure for DIF. The estimate of the difference between the means and variances of the ability distributions of the studied groups of respondents is used as an effect size and the purification procedure is stopped when the change in this effect size becomes negligible.

This chapter has been submitted for publication as: M. N. Khalid and Cees A. W. Glas, A Stepwise Method for Evaluation of DIF.

2.1 Introduction

Differential item functioning (DIF) occurs when respondents with the same ability but from different groups (say, gender or ethnicity groups) have different response probabilities on an item of a test or questionnaire (Embretson & Reise, 2000). Several statistical DIF detection methods have emerged (Holland & Thayer, 1988; Muthen, 1988; Shealy & Stout, 1993; Swaminathan & Rogers, 1990; Thissen, Steinberg, & Wainer, 1988; Kelderman & Macready, 1990; Finch, 2005; Oort, 1998; Navas-Ara & Gómez-Benito, 2002) and many reviews of DIF methods are provided in the literature (e.g., Camilli & Shepard, 1994; Holland & Wainer, 1993; Millsap & Everson, 1993; Roussos & Stout, 2004; Penfield & Camilli, 2007). Most of the techniques for the detection of DIF that have been proposed are based on evaluation of differences in response probabilities between groups conditional on some measure of ability. We consider two classes, the first class where a manifest score, such as the number-correct score, is taken as a proxy for ability and a second class where a latent ability variable of an IRT model functions as an ability measure.

The most used method in the first class is the Mantel-Haenszel (MH) approach where DIF is evaluated by testing whether the response probabilities, given number-correct scores, differ between the groups. Though the MH test works quite well in practice, Fischer (1993, 1995) points out that its application based on the assumption that the Rasch model holds, and when applying the MH test in other cases, several theoretical limitations arise. Also in the log-linear approach manifest sum scores are used as proxies for ability and the issues raised by Fischer (1993, 1995; see also Meredith & Millsap, 1992) apply here as well. The observed score is nonlinearly related to the latent ability metric (Embretson & Reise, 2000; Lord, 1980), and factors such as guessing may preclude an adequate representation of the probability of correct response conditional on ability. However, in general the correlation between the number-correct scores and ability estimates is quite high, so this is not the most important reason for considering alternative methods. The main problem arises in situations where the number-correct score loses its value as a proxy for ability. One may think of test situations with large amounts of missing data, or of computerized adaptive testing, where every student is administered a virtually unique set of items.

In an IRT model, ability is represented by latent variable θ, and a possible solution to the problem is to apply the MH and log-linear approach using subgroups that are homogenous with respect to an estimate of θ. This, however, introduces the problem that the estimate of θ is subject to estimation error, which is difficult to take into account when forming the subgroups. An alternative is to view DIF as a special case of misfit of an IRT model and to use the machinery for IRT model-fit evaluation to explore DIF. An overview of this approach was given by Thissen, Steinberg, and Wainer (1993). In that overview, evaluation of item parameter invariance over subgroups using Likelihood ratio and Wald statistics was presented as the main statistical tool for detection of DIF. Glas (1998, 1999) argued that the Likelihood ratio and Wald approach are not very efficient because they require estimation of the parameters of the IRT model under the alternative hypothesis of DIF for every single item. Therefore, Glas (1998, 1999) proposed using the Lagrange multiplier (LM) test by Aitchison and Silvey (1958), and the equivalent efficient-score test (Rao, 1948), which do not require estimation of the parameters of the alternative model. Further, this approach supports the evaluation of many more model assumptions such as the form of the response function, unidimensionality and local stochastic independence, both on the level of items (Glas & Falcon, 2003) and the level of persons (Glas & Dagohoy, 2007).

All methods listed above are seriously affected by the presence of high proportions of DIF items in a test and by the inclusion of DIF items in matching variable. Finch (2005) conducted a series of simulation to compare the performance of MIMIC, the Mantel-Haenszel, the IRT likelihood ratio test and the SIBTEST and found that an inflated Type I error rate and deflated power when there were more than 20 % DIF items in the test. To address this issue, many scale purification procedures have been developed (Lord, 1980; Wang & Su, 2004a, 2004b; French & Maller 2007). In the present chapter, an alternative purification method using Lagrange multiplier tests is proposed.

Another issue is the importance of DIF, i.e., the extent to which the inferences made using the test results are biased by DIF. Considering the effect size of DIF is important to avoid complicating inferences by practically trivial but statistically significant results. An example of a method to quantify the effect size is the DIF classification system for use with the MH statistical method developed by the Educational Testing

Service (Camilli & Shepard, 1994; Clauser & Mazor, 1998). In the framework of IRT, the present paper proposes to use an estimate of the difference between the means of the ability distributions of the studied groups of respondents as an effect size. This is motivated by the fact that ability distributions play an important role in most inferences made using IRT, such as in making pass/fail decisions, test equating, and the estimation of linear regression models on ability parameters as used in large scale educations surveys (NEAP, TIMSS, PISA).

This study is organized as follows. First, the modeling of DIF and a concise frame work of LM test is sketched for the identification of misfit items. Next, an example using empirical data is presented to show how the procedure works in practice. Then a number of simulation studies of the Type I error rate and power are presented. Next, the difference between two versions of the LM test, one targeted at uniform DIF and one targeted at non-uniform DIF is shown using a simulated example. Finally, some conclusions are drawn, and some suggestions for further research are given.

2.2 Detection and Modeling of DIF

In IRT models, the influences of items and persons on the observed responses are modeled by different sets of parameters. Since DIF is defined as the occurrence of differences in expected scores conditional on ability, IRT modeling seems especially fit for dealing with this problem. In practice, more than one DIF item may be present, and therefore, a stepwise procedure will be proposed where DIF items are identified one or two at a time. Both the significance of the test statistics and the impact of DIF are taken into account. Below, the following procedure will be outlined. First, marginal maximum likelihood (MML) estimates of the item parameters and the means and variance parameters of the different groups of respondents are made using all items. Then the item is identified with the largest significant value on a Lagrange multiplier (LM) test statistic targeted at DIF. To model the DIF in this item, the item is given group-specific item parameters. That is, in the analysis, the item is split into two virtual items, one that is supposed to be given to the focal group and one that is supposed to be given to the reference group. Then, new MML estimates are made and the impact of DIF in terms for the change in the means and variances of the ability distributions is evaluated. If this change is considered substantial, the next item with DIF is searched for. The process is repeated until no more significant or relevant DIF is found. The assumptions of this procedure are that (1) the item which is mostly affected by DIF will have the largest value of the LM statistic regardless of the bias caused by the other items with DIF, and (2) the change in the means and variances of ability distributions will decrease when the items with the DIF are given group specific item parameters one or two at a time.

2.2.1 IRT Models

In the present study, we both consider dichotomously and polytomously scored items. For dichotomously scored items, the one-parameter logistic model (1PLM) by Rasch (1960), the two-parameter logistic model (2PLM) and the three-parameter logistic model (3PLM) by Birnbaum (1968) will be used. For polytomously scored items, we use the generalized partial credit model (GPCM, Muraki, 1992). However, the methods proposed here also apply to other models for polytomously scored items, such as the PCM by Masters (1982) or the nominal response model by Bock (1972).

In the 3PLM, the item is characterized by a difficulty parameter β_i, a discrimination parameter α_i and a guessing parameter γ_i. Further, θ_n is the latent ability parameter of respondent n. The probability of correctly answering an item (denoted by $\mathrm{X}_{ni}=1$) is given

$$P(\mathrm{X}_{ni}=1 \mid \theta_n) = P_i(\theta_n) = \gamma_i + (1-\gamma_i)\frac{\exp(\alpha_i(\theta_n-\beta_i))}{1+\exp(\alpha_i(\theta_n-\beta_i))} . \quad (2.1)$$

If the guessing parameter γ_i is constrained to zero the model reduces to the 2PLM and if also the discrimination parameter α_i is constrained to one the model reduces to the 1PLM.

DIF pertains to different response probabilities in different groups. Here we consider two groups labeled the reference group and the focal group. The generalization to more than two groups is straightforward. A background variable will be defined by

$$y_n = \begin{cases} 1 & \text{if person n belongs to the focal group,} \\ 0 & \text{if person n belongs to the reference group.} \end{cases}$$

As a generalization of the model defined by equation 2.1 we consider

$$P_i(\theta_n) = \gamma_i + (1-\gamma_i)\frac{\exp(\alpha_i(\theta_n - \beta_i) + y_n(\phi_i(\theta_n - \delta_i)))}{1+\exp(\alpha_i(\theta_n - \beta_i) + y_n(\phi_i(\theta_n - \delta_i)))} \ . \qquad (2.2)$$

This model implies that the responses of the reference population are properly described by the model given by equation 2.1, but that the responses of the focal population need additional location parameters δ_i, additional discrimination parameters ϕ_i, or both given by equation 2.2. The first instance covers so-called uniform DIF, that is, a shift of the item response curve for the focal population, while the later two cases are often labeled non-uniform DIF, that is, the item response curve for the focal population not only shifted, but it also intersects the item response curve of the reference population.

For polytomous items, the GPCM by Muraki (1992) will be used. The probability of a student *n* scoring in category *j* on item *i* (denoted by $\mathrm{X}_{ni} = 1$) is given by

$$P(\mathrm{X}_{nij} = 1 \mid \theta_n) \ = \ P_{ij}(\theta_n) \ = \ \frac{\exp(j\alpha_i\theta_n - \beta_{ij})}{1+\sum_{h=1}^{M_i}\exp(h\alpha_i\theta_n - \beta_{ih})} \ , \qquad (2.3)$$

for *j* = *1,…, M_i*. An example of the category response functions $P_{ij}(\theta_n)$ for an item with four ordered response categories is given in Figure 2.1. Further, the graph shows the expected item-total score

$$E(T_i \mid \theta) \ = \ \sum_{j=1}^{M_i} jE(X_{ij} \mid \theta) \ = \ \sum_{j=1}^{M_i} jP_{ij}(\theta) \ . \qquad (2.4)$$

where the item-total score is defined as $T_i = \sum_{j=1}^{M_i} jX_{ij}$. Note that the expected item-total score increases as a function of θ. The generalization of equation 2.2 can be easily extended for equation 2.3.

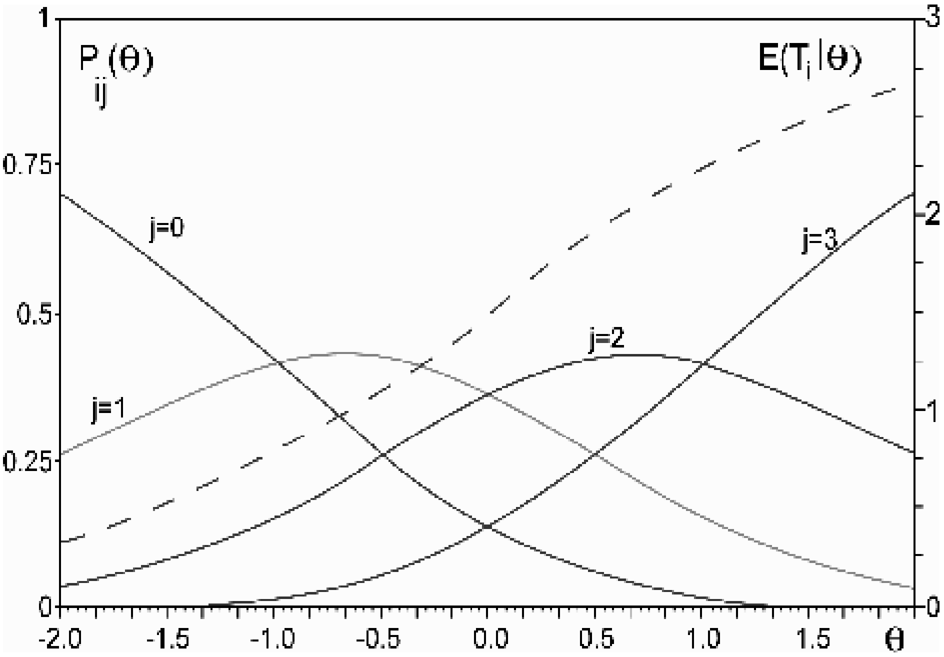

Figure 2.1. Response functions and expected item-total score under the GPCM.

2.2.2 MML Estimation

The LM test for DIF will be implemented in an MML estimation framework. To describe the statistic, MML estimation will be outlined first. MML estimation was developed by Bock and Aitkin (1981; see also Bock & Zimowski, 1997; Rigdon & Tsutakawa, 1983; Mislevy, 1984, 1986). In the MML frame work adopted here, it is assumed that the respondents belong to groups, and that ability parameters of the respondents with in a group have a normal distribution indexed by a group specific-mean and variance parameter. Let $g(\theta_n;\boldsymbol{\lambda}_{y(n)})$ be the density of ability distribution of group y, with parameters $\boldsymbol{\lambda}_{y(n)}$ where $y(n) = y_n$, i.e., the index of the group to which respondent n belongs. Usually, to identify the model, the mean and variance of one of the groups are set to zero and unity, respectively. Further, let ξ be a vector that contains all the item parameters. Finally, $\boldsymbol{\eta}$ is the vector of all item parameters ξ and the parameters λ of the ability distributions. The log likelihood function of $\boldsymbol{\eta}$ can be written as

$$\log L(\boldsymbol{\eta}) = \sum_{n=1}^{N} \log \int p(\boldsymbol{x}_n \mid \theta_n, \xi) g(\theta_n; \boldsymbol{\lambda}_{y(n)}) d\theta_n \quad . \tag{2.5}$$

where $p(\boldsymbol{x}_n \mid \theta_n, \xi)$ is the probability of response pattern $\boldsymbol{x}_n$ of respondent n (n = 1,…, N). The estimation equations that maximize the log-likelihood are found by setting the first-order derivatives of equation 2.5 with respect to $\boldsymbol{\eta}$ equal to zero. Glas (1999) shows that expressions for the first-order derivatives can be derived using Fischer's identity (Efron, 1977; Louis, 1982):

$$\frac{\partial}{\partial \boldsymbol{\eta}} \log L(\boldsymbol{\eta}) = \sum_{n} E\left[\omega_n(\boldsymbol{\eta}) \mid \boldsymbol{x}_n; \eta\right] \tag{2.6}$$

with

$$\omega_n(\boldsymbol{\eta}) = \frac{\partial}{\partial \boldsymbol{\eta}} \log \left[p(\boldsymbol{x}_n \mid \theta_n, \xi) g(\theta_n; \boldsymbol{\lambda}_{y(n)}) \right] \ .$$

The expectation in equation 2.6 is with respect to the posterior distribution $p(\theta_n \mid \boldsymbol{x}_n; \xi, \lambda_{y(n)})$. That is, the first order derivatives are equal to the posterior expectations of the first order derivatives of a likelihood function where the ability parameters are treated as observations. This grossly simplifies the derivations of the likelihood equations because $\omega_n(\boldsymbol{\eta})$ is very simple to derive. As an example we derive the MML estimate for the mean of the ability distribution of the focal group, that is, the group of respondents where y_n = *1*. The distribution of the ability parameters is normal, so if the values of θ_n would be known, the estimation equation $\sum_n \omega_n(\boldsymbol{\eta}) = 0$ would be equivalent to

$$\mu \ = \ \frac{\sum_{n=1}^{N} y_n \theta_n}{\sum_{n=1}^{N} y_n} \quad .$$

By Fisher's identity as given in equation 2.6, the MML estimation equation becomes

$$\mu \ = \ \frac{\sum_{n=1}^{N} y_n E\left[\theta_n \mid \mathbf{x}_n; \eta\right]}{\sum_{n=1}^{N} y_n} \quad . \tag{2.7}$$

Below, this identity will prove very helpful in the interpretation of the LM test for DIF.

2.2.3 A Lagrange Multiplier Test for DIF

In IRT, test statistics with a known asymptotic distribution are very rare. The advantage of having such a statistic available is that the test procedure can be easily generalized to a broad class of IRT models. Therefore, in the present article, the testing procedure will be based on the Lagrange multiplier test. In 1948, Rao introduced a testing procedure based on the score function as an alternative to likelihood ratio and Wald tests. Silvey (1959) rediscovered the score test as the Lagrange multiplier (LM) test. The LM test (Aitchison & Silvey, 1958) is equivalent with the efficient-score test (Rao, 1948) and with the modification index that is commonly used in structural equation modeling (Sörbom, 1989). Applications of LM tests to the framework of IRT have been described by Glas (1998, 1999), Glas and Falcon (2003), Jansen and Glas (2005), and Glas and Dagohoy (2007).

To identify DIF as defined by the model given in equation 2.2, we test the null hypothesis $\phi_i = 0$ and $\delta_i = 0$ using the statistic given by

$$\mathrm{LM} = \boldsymbol{h}' \boldsymbol{W}^{-1} \boldsymbol{h} \ , \tag{2.8}$$

where $\boldsymbol{h}$ is a 2-dimensional vector with as elements the first order derivatives of the likelihood function with respect to ϕ_i and δ_i, respectively. $\boldsymbol{W}$ is the 2 x 2 covariance matrix of $\boldsymbol{h}$**.** The statistic is evaluated in the point $\phi_i = 0$ and $\delta_i = 0$ using MML estimates under the null model, that is, using the MML estimates of the 2PLM or 3PLM. The idea of the test is that if the absolute values of these derivatives are large, the parameters fixed to zero will change if they are set free. In that case, the test becomes significant and the IRT model under the null hypothesis is rejected because of the presence of DIF. If the absolute values of these derivatives are small, the fixed parameters will probably show little change should they be set free. It means that the test is not significant and the IRT model under the null hypothesis is adequate.

For the null hypothesis $\phi_i = 0$ and $\delta_i = 0$, LM has an asymptotic chi-square distribution with two degrees of freedom. Details about the computation of $\boldsymbol{W}$ can be found in Glas (1998). The advantage of using the LM test instead of the analogous likelihood ratio or Wald tests is that only the null model, that is the 2PLM or 3PLM, has to be estimated, and using these estimates, a whole range of model violations can be evaluated, including DIF for all items, violations of local independence, multidimensionality and the form of the response functions (Glas, 1999).

As a special case, consider the alternative model given by equation 2.2, in the 2PLM version, that is, with $\gamma_i = 0$, and with $\phi_i = 0$. Then the probability of a correct response becomes

$$P_i(\theta_n) = \frac{\exp(\alpha_i(\theta_n - \beta_i) + y_n\delta_i)}{1 + \exp(\alpha_i(\theta_n - \beta_i) + y_n\delta_i)} \ . \tag{2.9}$$

If we treat α_i, β_i and θ_n as known constants this is an exponential family model with parameter δ_i. It is well known that the first order derivative of an exponential family likelihood is the difference between the sufficient statistic and its expectation (see, for instance, Andersen, 1980). The parameter δ_i in equation 2.9 is an item difficulty parameter pertaining to the subgroup with $y_n = 1$. The sufficient statistic for an item difficulty parameter is the number-correct score. So conditional on θ_n the first order derivative is

$$\sum_{n=1}^{N} y_n x_{ni} - \sum_{n=1}^{N} y_n P_i(\theta_n) \ ,$$

and using Fishers identity as given in equation 2.6 results in

$$\sum_{n=1}^{N} y_n x_{ni} - \sum_{n=1}^{N} y_n E[P_i(\theta_n) \mid \mathbf{x}_n; \boldsymbol{\eta}] \ .$$

So the statistic is based on residuals, that is, on the difference between the number-correct score in the focal group and its posterior expected value.

A DIF statistic for polytomously scored items based on residuals can be constructed analogously. To create a test based on the differences between item-total scores in subgroups and their expectations, a model is defined where the item-total score is a sufficient statistic, that is,

$$P_{ij}(\theta_n) = \frac{\exp(j\alpha_i\theta_n - \beta_{ij} + y_n j\delta_i)}{1+\sum_{h=1}^{M_i}\exp(h\alpha_i\theta_n - \beta_{ih} + y_n h\delta_i)} . \tag{2.10}$$

Note that $T_i = \sum_{j=1}^{M_i} y_n jX_{ij}$ is a sufficient statistic for δ_i. Therefore, an LM test for the null hypothesis $\delta_i = 0$ will be based on the residuals

$$\sum_{n=1}^{N}\sum_{j=1}^{M_i} y_n jX_{ij} - \sum_{n=1}^{N}\sum_{j=1}^{M_i} y_n jE(P_{ij} \mid \mathbf{x}_n; \boldsymbol{\eta}) . \tag{2.11}$$

An example will be given in the next section.

2.3 An Empirical example

The example pertains to the scale for ‘Attitude towards English Reading’ which consisted of 50 items with five response categories of each. The scale was administered to 8th grade students in a number of elementary schools of Pakistan. The respondents were divided into two groups on the basis of gender. The sample consisted of 1080 boys and 1553 girls. The item parameters were estimated by MML assuming standard normal distributions for the θ-parameters of both groups.

Table 2.1 gives the results for the LM test of the hypothesis $\delta_i = 0$. The results pertain to the first 14 items plus the 6 items with the most significant results in the remaining 36 items. The column labeled ‘LM’ gives the values of the LM-statistics; the column labeled ‘Prob’ gives the significance probabilities. The statistics have 1 degree of freedom. 10 of the 50 LM-tests were significant at a 5% significance level. The observed item-total scores (first term in equation 2.11) and expected item-total scores (second term in equation 2.11) averaged over the two groups are shown under the headings ‘Obs’ and ‘Exp’, respectively. To get an impression of the effect size of the misfit, the mean absolute difference between the observed and expected item-total scores are given under the heading “Abs.Diff”. The observed and expected values were quite close: the mean absolute difference was approximately .02 and the largest absolute difference was .19. This analysis was the starting point for the iterative procedure of identification and modeling of DIF. The item with the largest LM value, Item 37, was split into two virtual items, one that was supposed to be given to the boys and one that was supposed to be given to the girls, new MML estimates were made and the next item with the largest DIF was identified. Figure 2.2 gives the history of the procedure over iterations in terms of the difference between the estimates of the means of the ability distributions of the boys and girls as obtained using the MML estimates. The mean of the ability distribution of the girls was set equal to zero to identify the model, so the values displayed in Figure 2 are the averages for the boys, together with a confidence interval. Note that the initial change is quite large and the change decreases over iterations. The change of the variance of the ability distributions over iterations was very small.

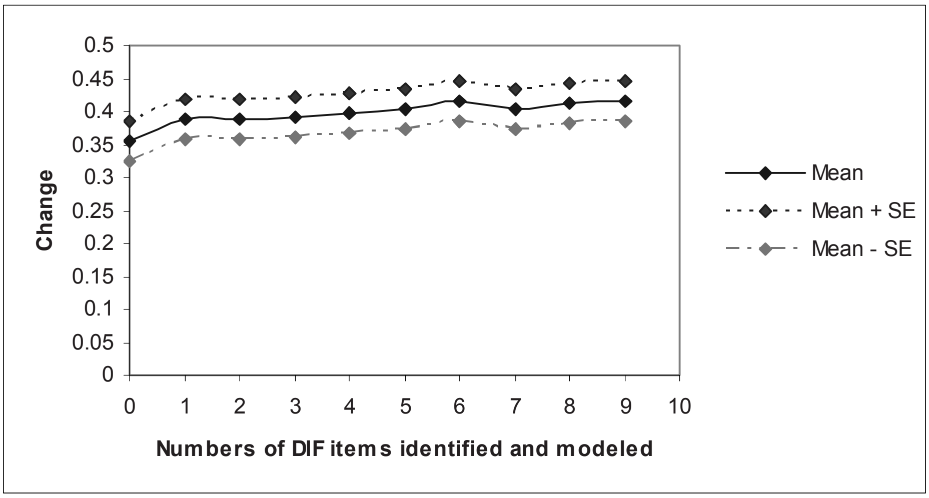

Figure 2.2. Change in the estimates of the means of the ability distribution over iterations.

A conservative conclusion is to stop the modeling DIF after six items because the impact on the estimates of the ability distribution, and inferences made using these distributions, such as norming and equating, became negligible.

Table 2.1. The results of LM test to evaluate fit of DIF.

			Boys		Girls		
Item	LM	Prob	Obs	Exp	Obs	Exp	Abs.Diff
1	1.09	0.30	2.75	2.70	2.49	2.52	0.04
2	0.95	0.33	3.28	3.25	3.05	3.07	0.03
3	2.70	0.10	3.23	3.18	2.94	2.98	0.04
4	6.20	0.01	3.26	3.19	2.91	2.96	0.06
5	2.45	0.12	2.70	2.76	2.65	2.60	0.05
6	3.40	0.07	3.27	3.21	2.97	3.01	0.05
7	1.02	0.31	3.13	3.16	2.97	2.95	0.02
8	2.88	0.09	2.93	2.98	2.76	2.72	0.05
9	0.40	0.53	3.11	3.13	2.91	2.89	0.02
10	0.03	0.86	2.99	2.98	2.79	2.79	0.01
11	0.20	0.65	2.67	2.65	2.44	2.46	0.02
12	0.68	0.41	3.05	3.08	2.91	2.90	0.02
13	3.28	0.07	3.32	3.27	3.00	3.03	0.04
14	2.81	0.09	2.78	2.84	2.71	2.67	0.05
25	8.50	0.00	3.02	3.11	2.95	2.88	0.08
30	8.26	0.00	3.32	3.23	2.96	3.02	0.07
33	4.51	0.03	3.14	3.08	2.81	2.85	0.06
37	20.18	0.00	1.87	2.09	2.01	1.86	0.19
41	14.21	0.00	2.30	2.48	2.41	2.28	0.15
50	5.13	0.02	3.44	3.38	3.15	3.20	0.06

2.4 Design of Simulation Studies

The first simulation studies presented concern the LM test targeted at uniform DIF, that is, the test for the null-hypothesis $\delta_i = 0$. The LM test targeted at non-uniform DIF, that is the test for the null hypothesis $\phi_i = 0$ and $\delta_i = 0$ will be treated in a next section. The simulations pertain to the 1PLM, the 2PLM and the 3PLM for dichotomous items. The Ability parameters were drawn from a standard normal distribution. For the 3PLM studies, data were generated using guessing parameters fixed at 0.2. The item discrimination parameters were drawn from a log-normal distribution with a mean equal to 1.0 and a standard deviation equal to 0.5 and the item difficulty parameters were drawn from standard normal distribution, except for the items with DIF. For the latter items, the discrimination and difficulty parameters were fixed to one and zero, respectively. This was done to prevent extreme parameter values when the effect size δ_i was added. Effect sizes were $\delta_i = 0$, $\delta_i = 0.5$ and $\delta_i = 1.0$. Test length was varied as K = 10, K = 20, and K = 40 and the sample sizes were N = 100, N = 400, and N= 1000 per group. The number of DIF items was varies as 0%, 10%, 20%, 30% and 40% of the test length. 100 replications were made in each condition of study. In all studies a nominal significance level of 5 % was used. The Type I error rates were evaluated by the proportion of times in the course of 100 replications a DIF-free item was mistakenly identified as exhibiting DIF. The power of the test was determined by the proportion of times in the course of 100 replications a DIF item was correctly identified. In the present example, the stepwise procedure consisted of four steps where two significant items (if present) were given group-specific item parameters in each step, so the changes in the means and variances of ability distributions were considered here as a stopping rule. The changes will be studied in the next section.

2.5 Results

2.5.1 Type I Error Rates

Table 2.2 summarizes the performance of LM test as function of sample size, test length, effect size, and the number of misfit items. The column labeled 0% gives the Type I error rate when no DIF items are present. The Type I error rate approached the nominal significance level in all settings of a sample size of N = 400 and N = 1000 for the test lengths K =20 and K = 40. In the presence of DIF items, the control of Type I error rate deteriorated for a test length of 10 items with 30% or 40% DIF items. It must be noted that 40% items with DIF is very high. If this percentage were equal to 50%, it cannot even be logically decided which one of the two parts of the test has DIF. So the conclusion is that the control of Type I error is good for reasonable test lengths (K = 20 and K = 40) combined with a reasonable sample size (say, 400 or more), or for a short test length (K = 10) with less than 20% DIF items. The results for the 1PLM and the 3PLM were analogous.

2.5.2 Power of the Test

Table 2.3 and 2.4 show results of the estimated power of test in the same simulation as in the previous section, for the 2PLM and the 3PLM, respectively. The results for the 1PLM are not shown, because they where very close to and not statistically different from the results for the 2PLM. Note that the tables show the expected main effects of sample size, test length, and effect size of DIF. If we disregard combinations of test length and sample size that have already been disqualified in the Type I error study reported above, it can be seen that the power of the procedure was high and for some combinations equal to 1.0. The power for the 3PLM was substantially lower than the power for the 2PLM.

Table 2.2. The Type I error rates by test length, effect size and sample size under the 2PLM.

			Percentage of Items with DIF				
K	δ	N	0%	10%	20%	30%	40%
10	0.5	100	0.06	0.07	0.08	0.09	0.13
		400	0.06	0.04	0.06	0.09	0.20
		1000	0.04	0.05	0.05	0.08	0.32
	1.0	100		0.08	0.08	0.16	0.34
		400		0.04	0.05	0.12	0.47
		1000		0.05	0.04	0.11	0.55
20	0.5	100	0.08	0.09	0.08	0.10	0.09
		400	0.05	0.06	0.05	0.07	0.06
		1000	0.06	0.06	0.06	0.05	0.03
	1.0	100		0.08	0.08	0.08	0.07
		400		0.06	0.05	0.05	0.04
		1000		0.05	0.06	0.05	0.03
40	0.5	100	0.13	0.15	0.15	0.15	0.15
		400	0.06	0.07	0.07	0.07	0.06
		1000	0.06	0.06	0.04	0.06	0.04
	1.0	100		0.15	0.14	0.11	0.09
		400		0.07	0.06	0.05	0.05
		1000		0.05	0.06	0.05	0.04

Table 2.3. The Power of test by test length, effect size and sample size under the 2PLM.

			Number of Item with DIF			
K	δ	N	10%	20%	30%	40%
10	0.5	100	0.33	0.28	0.21	0.17
		400	0.81	0.85	0.70	0.52
		1000	1.00	1.00	0.96	0.63
	1.0	100	0.81	0.77	0.60	0.40
		400	1.00	1.00	0.91	0.45
		1000	1.00	1.00	0.93	0.37
20	0.5	100	0.42	0.40	0.38	0.39
		400	0.89	0.84	0.83	0.84
		1000	1.00	0.99	1.00	0.99
	1.0	100	0.84	0.89	0.87	0.87
		400	1.00	1.00	1.00	1.00
		1000	1.00	1.00	1.00	1.00
40	0.5	100	0.54	0.52	0.47	0.48
		400	0.88	0.87	0.86	0.87
		1000	1.00	1.00	1.00	1.00
	1.0	100	0.94	0.92	0.94	0.89
		400	1.00	1.00	1.00	1.00
		1000	1.00	1.00	1.00	1.00

The results show that the proposed method compares favorably with alternative scale purification methods. Hulin, Lissak, and Drasgow (1982) conclude that scale purification procedures suffer when there is more than 20% DIF contamination in the test. In line with their findings, samples of 100 are insufficient for conducting a test with reasonable power and Type I error rate characteristics.

2.5.3 DIF and Population Parameters

The second aim of the study was to address the issue of importance of DIF, i.e., the effect size, and related problem of defining a stopping rule for the searching procedure. The associated formal test of model fit based on a statistic with a known (asymptotic) distribution is only relevant for moderate sample sizes; for large sample sizes, these tests become less interesting, because their power then becomes so large that even the smallest deviations from the model become significant. In these cases, the effect size becomes more important than the significance probability of the test. The location of the latent scale can be identified by setting the mean of the ability distribution of the reference population equal to zero. In addition, to identify the 2PLM and 3PLM, the variance of the reference population can be set equal to 1.0.

Table 2.4. The Power of test by test length, effect size and sample size under the 3PLM.

			Number of Item with DIF			
K	δ	N	10%	20%	30%	40%
10	0.5	100	0.18	0.10	0.05	0.05
		400	0.80	0.58	0.48	0.30
		1000	1.00	0.98	0.68	0.44
	1.0	100	0.72	0.50	0.29	0.12
		400	1.00	1.00	0.70	0.35
		1000	1.00	1.00	0.83	0.37
20	0.5	100	0.25	0.13	0.11	0.09
		400	0.80	0.76	0.70	0.62
		1000	1.00	1.00	0.97	0.89
	1.0	100	0.78	0.62	0.58	0.52
		400	1.00	1.00	0.99	0.95
		1000	1.00	1.00	1.00	1.00
40	0.5	100	0.30	0.20	0.20	0.20
		400	0.86	0.76	0.77	0.76
		1000	1.00	1.00	1.00	1.00
	1.0	100	0.75	0.65	0.59	0.56
		400	1.00	1.00	1.00	1.00
		1000	1.00	1.00	1.00	1.00

In the stepwise procedure defined above an identified DIF item is given group specific item parameters and new MML estimates of the item parameters and the parameters of the ability distribution are made. In the present case, the relevant ability distribution parameters are those of the focal population. It is assumed that the change in the estimates between steps gives an indication of the importance of the identified DIF.

Table 2.5. Estimates of the mean of the ability distribution in the different steps of the purifications procedure (test length K = 20).

δ	N	DIF items	Step 0	Step 1	Step 2	Step 3	Step 4
0.5	100	10%	-0.033	-0.025			
		20%	-0.036	-0.031	-0.037		
		30%	-0.067	-0.051	-0.063	-0.055	
		40%	-0.085	-0.075	-0.072	-0.079	-0.066
	400	10%	-0.015	0.001			
		20%	-0.051	-0.027	-0.009		
		30%	-0.054	-0.030	-0.013	0.002	
		40%	-0.090	-0.069	-0.048	-0.028	-0.010
	1000	10%	-0.023	0.001			
		20%	-0.043	-0.019	0.001		
		30%	-0.069	-0.044	-0.021	0.000	
		40%	-0.094	-0.069	-0.044	-0.020	0.000
1.0	100	10%	-0.035	-0.000			
		20%	-0.096	-0.055	-0.016		
		30%	-0.136	-0.091	-0.061	-0.026	
		40%	-0.150	-0.103	-0.056	-0.017	0.012
	400	10%	-0.026	0.017			
		20%	-0.095	-0.046	-0.004		
		30%	-0.137	-0.088	-0.043	-0.003	
		40%	-0.214	-0.163	-0.113	-0.065	-0.023
	1000	10%	-0.046	-0.002			
		20%	-0.102	-0.056	-0.013		
		30%	-0.129	-0.083	-0.038	0.005	
		40%	-0.194	-0.145	-0.098	-0.051	-0.005

Average standard errors for the estimates: N = 100 : Se(Mean) = 0.180, N = 400 : Se(Mean) = 0.075, N = 1000 : Se(Mean) = 0.055

Table 2.5 gives the change in the estimate of the mean of the ability distribution of the focal ability distribution for one of the settings of the simulations reported above. The table pertains to the 2PLM and a test length of 20 items. The estimates are averages over 100 replications. The average standard errors of the estimates over 100 replications are reported at the bottom of the table for all three sample sizes. In every step, items identified with DIF were given group specific item-parameters two at a time.

The column labeled ‘Step 0’ gives the estimates of the means in the initial MML analysis, where no items were treated yet. The true means were all equal to zero, so it can be seen that there was a clear main-effect of the percentage of DIF items present. Further, it can be seen that in the final step of the procedure the estimates approach the true value of zero. In practice, the true value is of course not known, and therefore the convergence of the procedure must be judged from the differences in the estimates between steps. In the present example, only uniform DIF was generated and as a consequence, there was no systematic trend in the estimates of the variances of the ability distributions. All estimates were sufficiently close to the true value of 1.0. As will become clear in the next section, this no longer holds when non-uniform DIF is present.

2.5.4 Non-uniform DIF

In the previous sections, the focus was on uniform DIF. In the present section, a simulated example of non-uniform DIF is presented. In non-uniform DIF, usually both the difficulty and discrimination parameters differ between groups. Using same setup as in the previous simulations, a dataset of 20 items was simulated using the 2PLM. DIF was imposed on the first 6 items of the test by choosing $\phi_i = -0.50$ and $\delta_i = 0.50$. So in the focal group the discrimination parameters of the DIF items were lowered from 1.0 to 0.5 and the item difficulties rose from 0.0 to 0.5. This might reflect the situation where the respondents of the focal group were less motivated to make an effort on these items, which resulted in a lower probability of a correct response and an attenuated relation between the responses and the latent ability dimension. One of the questions of interest was the relation between the test targeted at uniform DIF (null-hypothesis $\delta_i = 0$) and test targeted at non-uniform DIF (null-hypothesis $\phi_i = 0$ and $\delta_i = 0$). The results are shown in Table 2.6. The columns 3 to 5 pertain to the first MML analysis where none of the items were given group-specific item parameters yet, the columns 6 to 9 pertain to the situation after the third step when 6 items where identified as DIF items. Note that all 6 items were correctly identified, even though also the test for item 7 was highly significant in the first analysis. The columns under the label ‘df = 1’ concern the test for $\delta_i = 0$, which has one degree of freedom; the columns under the label ‘df = 2’ refer to the test for $\phi_i = 0$ and $\delta_i = 0$, which has two degrees of freedom. Note that overall the test with one degree of freedom seems to have a higher power: in 19 cases its significance probability is lower than the significance probability of the test with two degrees of freedom. The latter test has the lowest significance probability in 8 cases. So in practice, the test with two-degrees of freedom will not add much information over the test with one degree of freedom.

Finally, at the bottom of the table, the estimates of the mean and standard deviation of the ability distribution of the focal group are given, together with the standard errors. It can be seen that in the initial analysis (Step 0) both the estimate of the mean and the variance were biased. However, after three steps the estimate of the variance is very close to its true value of 1.0 and the estimate of the mean is clearly within the confidence region around 0.0. So in this case, the change in both parameters must be considered to judge the convergence of the procedure.

2.6 Discussion and Conclusion

IRT is widely applied in the field of educational and psychological testing for such topics as the evaluation of the reliability and validity of tests, optimal item selection, computerized adaptive testing, developing and refining exams, maintaining item banks, and equating the difficulty of successive versions of examinations. However, these applications assume that the IRT models used hold. The presence of misfitting items may potentially threaten the realization of the advantages of IRT models. Therefore, over the course of the past decades the topic of model-fit has become of more and more interest to test developers and measurement practitioners. DIF is one of the most important threats to IRT model fit. A method for the analysis of DIF was proposed that addresses two issues. The first issue is that the presence of a large number of items with DIF biases statistical search procedures for DIF. Therefore, a stepwise purification procedure was introduced that consisted of alternating between identifying DIF using an LM test and modeling DIF using group-specific item parameters. The second issue is the importance of DIF and the related issue of when to stop searching for DIF and modeling DIF. It was argued that many applications of IRT entail inferences about the latent ability distribution. One may think of norming and standard setting, linking and equating, the estimation of group differences and linear regression models on ability parameters as used in large scale educations surveys. Therefore, the importance of DIF was related to ability distributions and it was suggested to monitor the purification procedure using the change of the estimates of the parameters of the ability distributions over the steps of the procedure.

Table 2.6. A comparison of the purification process using the LM tests for uniform and non-uniform DIF.

Item	Start Purification Procedure (Step 0)				End Purification Procedure (Step 3)			
	df = 1		df = 2		df = 1		df = 2	
	LM	Prob	LM	Prob	LM	Prob	LM	Prob
1	5.46	.02	8.22	.02	-	-	-	-
2	6.51	.01	9.65	.01	-	-	-	-
3	6.71	.01	10.59	.01	-	-	-	-
4	7.89	.00	11.84	.00	-	-	-	-
5	2.39	.12	6.00	.05	-	-	-	-
6	14.34	.00	20.23	.00	-	-	-	-
7	7.37	.01	9.56	.01	3.09	.08	3.36	.19
8	0.11	.74	0.19	.91	1.89	.17	2.13	.34
9	2.20	.14	3.46	.18	0.09	.77	0.09	.95
10	0.20	.65	8.02	.02	0.17	.68	3.87	.14
11	2.43	.12	2.60	.27	0.26	.61	0.61	.74
12	0.07	.79	0.47	.79	1.44	.23	1.47	.48
13	1.19	.28	1.19	.55	0.01	.94	0.50	.78
14	0.12	.73	0.48	.79	1.52	.22	1.54	.46
15	3.02	.08	3.54	.17	0.79	.37	0.79	.67
16	0.97	.32	1.97	.37	0.00	.95	0.08	.96
17	0.64	.42	0.66	.72	0.05	.82	1.68	.43
18	2.10	.15	3.51	.17	0.29	.59	0.47	.79
19	2.11	.15	2.13	.34	0.12	.73	0.65	.72
20	0.43	.51	4.94	.08	0.02	.89	1.48	.48

Mean	-.237	Mean	-.111
SE(Mean)	0.078	SE(Mean)	0.084
SD	0.823	SD	0.985
SE (SD)	0.061	SE (SD)	0.080

Simulation studies were presented to assess the Type I error rate and power of the procedure. It was concluded that the procedure worked well for sample sizes from 400 respondents and test lengths from 20 items. For a test length of 10 items, the procedure only worked well when the proportion of DIF items was 10% and 20%. In all situations, the power decreased with the proportion of DIF items. The power for the 3PLM was less than the power for the 2PLM. Further, for the case of uniform DIF, it was shown that DIF biased the estimates of the means of the ability distributions, but this bias vanished in the course of the stepwise purification procedure when DIF was modeled by the introduction of group-specific item parameters. In the case of non-uniform DIF, both the means and variances of the ability distributions were biased, but also this bias could be removed with group-specific item parameters. Further, the simulation studies also showed that the LM test targeted at uniform DIF was sufficiently sensitive to a combination of uniform and non-uniform DIF, and the inferences did not change when the LM test for non-uniform DIF was used.

3

Assessing Item Fit: A Comparative Study of Frequentist and Bayesian Frameworks

Abstract: Item fit indices for item response theory (IRT) models in a frequentist and Bayesian framework are compared. The assumptions that are targeted are differential item functioning (DIF), local independence (LI), and the form of the item characteristics curve (ICC) in the one-, two-, and three parameter logistic models. It is shown that Lagrange multiplier (LM) tests, which is a frequentist based approach, can be defined in such a way that the statistics are based on the residuals, that is, differences between observations and their expectations under the model. In a Bayesian framework, identical residuals are used in posterior predictive checks. In a Bayesian framework, it proves convenient to use normal ogive representation of IRT models. For comparability of the two frameworks, the LM statistics are adapted from the usual logistic representation to normal ogive representation. Power and Type I error rates are evaluated using a number of simulation studies. Results show that Type I error rate and power are conservative in the Bayesian framework and that there is more power for the fit indices in a frequentist framework. An empirical data example is presented to show how the frameworks compare in practice.

This chapter has been submitted for publication as: Cees A. W. Glas and M.N. Khalid, Assessing Item Fit: A Comparative Study of Frequentist and Bayesian Frameworks.

3.1 Introduction

Psychometric theory is the statistical framework for measurement in many fields of psychology and education. These measurements may concern abilities, personality traits, attitudes, opinions, and achievement. Item response theory (IRT) models play a prominent role in psychometric theory. In these models, the properties of a measurement instrument are completely described in terms of the properties of the items, and the responses are modeled as functions of item and person parameters. While many of the technical challenges that arise when applying IRT models have been resolved (e.g., model parameter estimation), the assessment of model fit remains a major hurdle for effective IRT model implementation (Hambleton & Han, 2005).

Model checking, or assessing the fit of a model, is an important part of any data modeling process. Before using the model to make inferences regarding the data, it is crucial to establish that the model fits the data well enough according to some criteria. In particular, the model should explain aspects of the data that influence the inferences made using the IRT model. Otherwise, the conclusions obtained using the model might not be relevant. IRT models are based on a number of explicit assumptions, so the method for the evaluation of model fit focus on these assumptions. The most important assumptions underlying these models are subpopulation invariance (DIF), the form of the ICC, local stochastic independence, and item score pattern. Researchers have proposed a significant number of fit statistics for assessing fit of IRT models. These statistics were developed to be sensitive to specific model violations (Andersen, 1973; Glas, 1988, 1999; Jansen & Glas, 2005; Glas & Verhelst, 1995a, 1995b; Glas & Suárez-Falcón, 2003; Holland & Rosenbaum, 1986; Kelderman, 1984, 1989; Maydeu-Olivares & Joe, 2005, 2006; Mokken, 1971; Molenaar, 1983; Sijtsma & Meijer, 1992; Stout, 1987, 1990). An essential feature of these statistics is that they are based on information that is aggregated over persons; therefore they will be refer to as aggregate test statistics.

To date most of the research to IRT model fit procedures has been done in a frequentist framework. Chi-square statistics are natural tests of the discrepancy between the observed and expected frequencies or proportions computed in a residual analysis. Both Pearson and likelihood ratio statistics have been proposed; these statistics have been standard tools for assessing model fit since the earliest applications of IRT.

A number of problems arise in using chi-square statistics as tests of model data fit in the IRT context. Principal among them is whether the statistics have the chi-square distribution claimed and if so, whether the degrees of freedom are correctly determined. Glas and Suárez-Falcón (2003) note that the standard theory for chi-square statistics does not hold in the IRT context because the observations on which the statistics are based do not have a multinomial or Poisson distribution. Simulation studies (Yen, 1981; McKinley & Mills, 1985; Orlando & Thissen, 2000, 2003) have shown that the fit statistics in common use do generally appear to have an approximate chi-square distribution; however, the number of degrees of freedom remains at issue. Orlando and Thissen (2000) argued that because the definition of the observed proportions correct are based on model-dependent trait estimates, the degrees of freedom may not be as claimed. Stone and Zhang (2003) agreed with the assessment of Orlando and Thissen (2000) and further noted that when the expected frequencies depend on unknown item and ability parameters, and when these are replaced by their estimates, the distribution of the chi-square statistic is grossly affected. Glas and Suárez-Falcón (2003) have also criticized these procedures along the same lines for failing to take into account the stochastic nature of the item parameter estimates. The model fit indices which are based on the likelihood ratio and Wald statistics are also problematic (computational intensive) because every alternative model for every model violation for every person and each item would have to be estimated (Glas, 1999).

To address the above mentioned issues Glas (1999) has proposed procedures based on the Lagrange multiplier (LM) statistic by Aitchison and Silvey (1958). The LM statistics estimate the IRT model only once and produce a number of tables of residuals that are informative with respect to specific model violations. An advantage of the use of LM tests is the necessity to formulate specific parametric alternatives to the assumptions targeted by test statistics. Glas have sketched the approach of LM test in the marginal maximum likelihood (MML) frame work (see, for instance, Bock & Aitkin, 1981; Mislevy, 1986) which is the standard procedure for parameter estimation in IRT. However, MML frame work may be less efficient (in terms of computation) for multilevel and multidimensional psychometric models (Fox & Glas, 2001; Béguin &

Glas, 2001) due to complex dependency structures of models and require the evaluation of multiple integrals to solve the estimation equations for parameters.

These computational problems are avoided in a fully Bayesian framework and now-a-days this framework is widely used for parameter estimation in complex psychometric IRT models. When comparing the fully Bayesian framework with the MML framework the following considerations play a role. First, a fully Bayesian procedure supports definition of a full probability model for quantifying uncertainty in statistical inferences (see, for instance, Gelman, Carlin, Stern, & Rubin, 2004, p. 3). This does involve the definition of priors, which creates some degree of bias, but this can be minimized using of non-informative priors. Second, estimates of model parameters that might otherwise be poorly determined by the data can be enhanced by imposing restrictions on these parameters via their prior distributions. However this can also be done in a Bayes modal framework, which is closely related to the MML framework (Mislevy, 1986).

Recently Sinharay (2005), Sinharay, Johnson and Stern (2006) have applied the popular Bayesian approach of posterior predictive checks (PPCs) to the assessment of model violations in unidimensioal IRT models. However, Bayarri and Berger (2000) have showed that PPCs also comes with problems due to twice use of data as a result posterior p-value were conservative (i.e., often failed to detect model misfit) and inadequate behavior of posterior p-value.

The advantage of a Bayesian approach, particularly when implemented through Markov chain Monte Carlo (MCMC) sampling from the posterior distribution, is the easy calculation of the posterior distribution of any function of the estimates. However, the frequentist approach has a long standing, more rigorously developed tradition of statistical test for model fit. The purpose of this study is to introduce analogous frequentist procedures (LM test) and Bayesian procedures (PPCs) and to compare their Type I error rate and power.

This chapter is organized as follows. First, the model violations, assumptions targeted by item fit statistics, that examined in this study are presented. The second section introduces the description of LM statistics and PPCs. The third section outlines the design of simulation studies.

Next, results from a study comparing empirical Type I error rates and power for the above frameworks are presented. Then, both frameworks are applied in the context of an empirical example. Finally, some conclusions are drawn, and some suggestions for further research are given.

3.2 Fit to IRT Models

The fit of a model, or the correspondence between model predictions and observed data, is generally regarded as an important property of model-based procedures like IRT. When a model does not fit the data, valid use of estimated parameters is compromised. IRT models are based on a number of explicit assumptions which can be viewed from two perspectives: the items and respondents. In the first case, for every item, residuals (differences between predictions from the estimated model and observations) and item fit statistics are computed to assess whether item violates the model. In the second case, residuals and person fit statistics are computed for every person to assess whether the responses to the items follow the model.

For unidimensional IRT models, a number of item fit statistics may be of interest, depending on the context of the problem. These models assume item parameters invariance, a specific shape of the ICC, local independence, fit of response pattern, and normality of the ability distribution, and each of these assumptions should be checked using suitable fit measures. The first assumption entails that the item responses can be described by the same parameters in all possible subpopulations. The shape of ICC describes the relation between the latent variable and the observable responses to items. Evaluation is usually done by comparing observed and expected item response frequencies given some measure of latent trait level. The third assumption, local independence, assumes that responses to different items are independent given the latent trait variable value. The important assumption evaluated from the perspective of person fit is the invariance of the ability parameter over sub-tests.

In a Bayesian framework, the normal-ogive representation of IRT models has a number of important computational advantages (Albert, 1992). Since the objective of this article is to compare the Bayesian and the frequentist likelihood-based framework, we adopt the normal-ogive representation and also apply it to the likelihood-based framework. In the 1- 2-, and 3-parameter models, it is assumed that the proficiency level of a respondent (indexed *n*) can be represented by one dimensional proficiency parameter θ_n. In the 3PNO model the probability of correct response to item *i*, denoted by $X_{ni} = 1$, as a function of θ_n is given by

$$\begin{aligned} P(X_{ni}=1\,|\,\theta_n)=P_i(\theta_n) &= c_i+(1-c_i)\int_{-\infty}^{a_i(\theta_n-b_i)}\frac{1}{\sqrt{2\pi}}\exp\left[\frac{-t^2}{2}\right]\partial t \\ &= c_i+(1-c_i)\Phi(a_i(\theta_n-b_i))\ . \end{aligned} \tag{3.1}$$

Note that $\Phi(.)$ is the cumulative standard normal distribution. The three item parameters a_i, b_i, c_i are called the discrimination, difficulty and guessing parameter, respectively. The 2PNO model follows upon setting the guessing parameter c_i equal to zero, and 1PNO model follows upon introducing the additional constraint $a_i = 1$.

3.2.1 Differential Item Functioning

Differential item functioning (DIF) is a difference in item responses between equally proficient members of two ore more populations. For the 3PNO model, the difference in the response probability between two or more groups can be modeled by defining an alternative model of equation 3.1 as follows. Consider the case of two groups. A background variable will be defined by

$$y_n=\begin{cases}1 & \text{if a person n belongs to the focal group,}\\ 0 & \text{if a person n belongs to the reference group.}\end{cases}$$

As a generalization of the model defined by equation 3.1, define

$$P_i(\theta_n)=c_i+(1-c_i)\Phi(a_i(\theta_n-b_i)+y_n\delta_i))\ . \tag{3.2}$$

This model implies that the responses of the reference group are properly described by formula 3.1, but that the responses of the focal group need additional location parameter δ_i given by equation 3.2. The additional parameter can be interpreted as a shift in the item difficulty parameter. In principle, changes could also be present in the discrimination parameter, but that case is not considered here.

3.2.2 Fit of the ICCs

The form of ICC describes the relation between a latent variable, say proficiency, and observable responses to items. Evaluation of the appropriateness of the ICC is usually done by comparing observed and expected item response frequencies given some measure of the latent trait level. The score range is partitioned into G subsets to form subgroups of respondents. The item targeted should be left out of the total score for technical reasons (see, Glas, 1999). Under the alternative model, the probability of a correct response is given by

$$P_i(\theta_n) = c_i + (1 - c_i)\Phi(a_i(\theta_n - b_i) + \delta_{ig}). \tag{3.3}$$

Under the null model, which is the 3PNO model, the additional parameter δ_{ig} is equal to zero. In the alternative model, δ_{ig} acts as a shift in item difficulty for subgroup g = 2 , …, G. The first group is used as a baseline.

3.2.3 Local Independence

The assumption entails that responses to different items are independent given the latent trait value. So the proposed latent variable completely describes the responses and no additional variables are necessary to describe the responses. For the 3PNO model, the dependency between the items *i* and *l* can be modeled as

$$P_i(\theta_n) = c_i + (1 - c_i)\Phi(a_i(\theta_n - b_i) + x_{il}\delta_{il}) . \tag{3.4}$$

Note the parameter δ_{il} models the associations between two items. A suitable item fit statistic can be used to test the special model, where $\delta_{il} =$ 0, against the alternative model, where $\delta_{il} \neq 0$.

3.3 Lagrange Multiplier (LM) Test

In 1948, Rao introduced a fundamental principle of testing based on the score function as an alternative to likelihood ratio and Wald tests. Silvey (1959) rediscovered the score test as a Lagrange multiplier (LM) test. Applications of LM tests of model-fit to the framework of IRT have been described by Glas and Verhelst (1995), Glas (1998, 1999), Glas and Falcón (2003), Jansen and Glas (2005), and Glas and Dagohoy (2007). The LM test is based on evaluating a quadratic function of partial derivatives of the log-likelihood function of the general model evaluated at the maximum likelihood estimates of the special model. The vector of the first order derivatives of the special model is equal to zero because their values originate from solving the likelihood equations. The magnitude of the elements of the vector of the first order derivatives corresponding with special parameters determines the value of the statistic: the closer they are to zero, the better the model fits.

More formally, the principle of the LM tests is developed as follows. Consider a null hypothesis about a model with parameters ϕ_0. This model is a special case of a general model with parameters ϕ. In the present case the special model is derived from the general model by fixing one or more parameters to known constants. It will be assumed that these parameters are not fixed at points on a boundary of parameter space. Let ϕ_0 be partitioned as $\phi'_0 = (\phi'_{01}, \phi'_{02}) = (\phi'_{01}$, c), where c is a vector of postulated constants. Let $h(\phi)$ be the partial derivatives of the log-likelihood of the general model, so $h(\phi) = (\partial / \partial\phi) \ln L(\phi)$. This vector of partial derivatives gauges the change of the log-likelihood as the function of local changes in ϕ. Let $H(\phi, \phi)$ be defined as – $(\partial^2 / \partial\phi \; \partial\phi') \ln L(\phi)$. [Note that H(., .) is used as a generic symbol for a matrix of the opposite of second order derivatives of the log-likelihood function and the variables with respect to which derivatives are taken are the arguments of H(., .). An analogous definition is used for h (.)].

Then LM statistics is given by

$$LM = h(\phi_0)' \; H(\phi_0,\phi_0)^{-1} \; h(\phi_0). \tag{3.5}$$

If the LM statistic is evaluated using the ML estimate of ϕ_{01} and the postulated values of c, it has an asymptotic chi-square distribution with degrees of freedom equal to the number of parameters fixed. An important computational aspect of the procedures is that at the point of the ML estimates $\phi^{\wedge}{}_{01}$, the free parameters have partial derivatives equal to zero. Therefore, (3.5) can be computed as

$$LM(c) = h(c)' W^{-1} h(c) \tag{3.6}$$

with

$$W = H(c,c) - H(c,\phi^{\wedge}{}_{01}) H(\phi^{\wedge}{}_{01},\phi^{\wedge}{}_{01})^{-1} H(\phi^{\wedge}{}_{01},c).$$

Note that $H(\phi^{\wedge}{}_{01},\phi^{\wedge}{}_{01})$ also plays a role in the Newton-Raphson procedure for solving the estimation equations and in computation of the observed information matrix or standard error. So its inverse will generally be available at the end of the estimation procedure. Further, if the validity of the model of the null-hypothesis is tested against various alternative models, the computational work is reduced because the inverse of $H(\phi^{\wedge}{}_{01},\phi^{\wedge}{}_{01})$ is already available and the order of **W** is equal to the number of parameters fixed, which must be small to keep the interpretation of the outcome tractable.

The interpretation of the test is supported by observing that the value of equation 3.6 depends on the magnitude of $h(c)$ that is the first order derivatives with respect to the parameters ϕ_{02} evaluated in c. If the absolute values of these derivatives are large, the fixed parameters are bound to change once they are set free. It means that the test is significant, that is, the special model is rejected. If the absolute values of these derivatives are small, the fixed parameters will probably show little change should they be set free. It means that the test is not significant, that is, the special model is adequate.

The model violations that are described above can be evaluated using LM test in a marginal maximum likelihood (MML) framework. Details will be given below. Under the null model, the 3PNO model, additional parameters are set to zero that designate specific model violations under alternative models such as DIF, local independence, and form of ICC. The vector $h(c)$ gauges the change in the additional parameters which is the difference between the observed proportion correct and its posterior expectation for a raw-score group, computed at the MML. The covariance matrix of $h(c)$, which is **W**, is available at the end of estimation procedure that is needed to derive the asymptotic distribution of the statistic.

3.4 Posterior Predictive Checks

Bayesian statistics has received considerable attention in statistics over the past decade. There has been a recent surge in the use of Bayesian estimation in IRT. Albert (1992), Patz and Junker (1999a, 1999b), Bradlow, Wainer, and Wang (1999), Janssen, Tuerlinckx, Meulders, and De Boeck (2000), Béguin and Glas (2001), Fox and Glas (2001, 2003), Bolt, Cohen, and Wollack (2002), Sinharay, Johnson, and Williamson (2003), and Wallack, Cohen, and Wells (2003) are only some of the recent examples of application of Bayesian estimation and Markov chain Monte Carlo (MCMC) algorithms (e.g., Gelman, Carlin, Stern, & Rubin, 2003) to fit complicated psychometric models. There are numerous instances in which classical methods fail, but the Bayesian approach offers a feasible method for assessing model fit. However, little attention has been given to assessing fit of IRT models from a Bayesian perspective.

In Bayesian statistics, a model can be checked in at least three ways: (1) examining sensitivity of inferences to reasonable changes in the prior distribution and the likelihood; (2) checking that the posterior inferences are reasonable, given the substantive context of the model; and (3) checking that the model fits the data. We address the third of these concerns using the posterior predictive distribution for a discrepancy, an extension of classical test statistics to allow dependence on unknown (nuisance) parameters. Posterior predictive assessment was introduced by Guttman (1967), applied by Rubin (1981), and given a formal Bayesian definition by Rubin (1984).

The posterior predictive checking (PPC) method is a popular Bayesian model checking tool because of its simplicity. The method primarily consists of comparing the observed data with replicated reference data (those predicted by the model) using a number of test statistics. The idea of PPC is to generate simulated values from the posterior predictive distributions of replicated data and to compare these samples to the observed data via the test statistic. If the replicated data and the observed data differ systematically, it is an indication of a potential model misfit. The framework of PPC can be defined as follows.

Letting X be the observed data and ϕ be the vector of all the parameters in the model, we define $p(X|\phi)$ as the likelihood and $p(\phi)$ as the prior distribution on the parameters. The PPC method suggests checking a model by examining whether the observed data appear extreme with respect to the posterior predictive distribution of replicated data, which is obtained by

$$p(X^{rep} \mid X) = \int p(X^{rep} \mid \phi)p(\phi \mid X)d\phi \tag{3.7}$$

Usually, discrepancy measures or a test quantities $T(X,\phi)$ are defined, and the posterior distribution of $T(X,\phi)$ is compared to the posterior predictive distribution of $T(X^{rep},\phi)$, with substantial difference between them indicate model fit. The PPC method allows a reasonable summary of such comparisons with the posterior p-value:

$$\begin{aligned}&\Pr(T(X^{rep},\phi) \geq T(X,\phi) \mid X) = \\ &\int_{T(X^{rep},\phi) \geq T(X,\phi)} p(X^{rep} \mid \phi)p(\phi \mid X)dX^{rep}d\phi\end{aligned} \tag{3.8}$$

Extreme posterior p-values are indicative of model misfit. Because of the difficulty in dealing with equation 3.7 and 3.8 analytically for all but simple problems, Rubin (1984) suggested simulating replicate data sets using MCMC from the posterior predictive distribution in practical applications of the PPC method. The standard methods of MCMC estimation include the Gibbs sampler and the more general Metropolis-Hastings (MH) algorithm. The Gibbs sampler is a special case of the MH algorithm in which parameters are sampled from their full conditional distributions (rather than a relatively arbitrary proposal distribution). The prescribed algorithm generates a Markov chain of iterates that constitute a random walk over the posterior distribution. The results are equivalent to integrating the density over the data to obtain a sample from the posterior distribution for the parameters.

One generates Q draws $\phi^1,\phi^2,\phi^3,......,\phi^Q$ from the posterior distribution $P(\phi \mid X)$ of ϕ, and draws $X^{rep,q}$ from the likelihood distribution $P(X \mid \phi^q)$, $q = 1, 2, . . . , Q$. The process results in Q replicated data sets. Equation 3.8 implies that the proportion of the Q replications for which $T\ (X^{rep,q},\phi)$ exceeds $T(X,\phi)$ provides an estimate of the posterior p-value. Extreme posterior p-values (close to 0, or 1, or both, depending on the nature of the test statistic) indicate model misfit.

The important applications of the PPC method in educational measurement and IRT context include Rubin and Stern (1994), Hoijtink and Molenaar (1997), Scheines, Hoijtink and Boomsma, 1999, Albert and Gosh (2000), Janssen, et al. (2000), van Onna (2003), Glas and Meijer (2003), and Fox and Glas (2001, 2003).

3.5 Tests for Fit of IRT Models

Because the objective of this study is to compare the Bayesian and the frequentist likelihood-based framework, the statistics used for the evaluation of differential item functioning, the shape of the response models and local independence will be generalizations of LM tests to PPCs. We proceed as follows. If the value of θ_n was known, the likelihood of a response x_{ni} would be given by

$$\log L_{ni} = x_{ni} \log P_i(\theta_n) + (1 - x_{ni}) \log(1 - P_i(\theta_n)) \ . \tag{3.9}$$

The first-order derivative of some item parameter ϕ_i, say $d(\theta_n,\phi_i)$, is then given by

$$d(\theta_n,\phi_i) \ = \ \frac{\left(x_{ni} - P_i(\theta_n)\right) P_i^{'}(\theta_n)}{P_i(\theta_n)(1 - P_i(\theta_n))} \ , \tag{3.10}$$

where $P_i^{'}(\theta_n)$ is the first-order derivative with respect to ϕ_i. In a marginal maximum likelihood framework (MML, Bock & Aitkin, 1981), it is assumed that the θ_n-parameters are stochastic variables with a normal distribution, and the likelihood in equation 3.9 is marginalized with respect to this distribution. Further, the first order derivatives as in equation 3.10 then become posterior expectations, that is,

$$E\left(d(\theta_n,\phi_i)\,|\,\boldsymbol{x}_n\right) \ , \tag{3.11}$$

where $\boldsymbol{x}_n$ is a vector of all responses of respondent n (see, for instance Glas, 1999). These first order derivatives can be used in the evaluation of equation 3.5.

For the evaluation of model fit using a PPC, $T(X,\phi)$ can be based on the test statistics that are defined analogous to the fit statistics for the MML framework. It will be shown that $T(X,\phi)$ can then be defined as a Pearson-type statistic, that is, as the squared difference between the observed and expectation divided by the standard deviations of the difference. One of the advantages of PPCs is that there is no need to derive the distribution of test statistics under the null model. The distribution is implicitly generated when running the MCMC procedure. This is also the explanation for the fact that there is no need to implicate a covariance matrix such as the covariance matrix **W** in the LM-statistic, where accounting for the dependence between the vectors $h(c)$ is essential for the derivation of the asymptotic distribution of the statistics.

The expressions for a test for DIF are derived as follows. The alternative model for DIF is given by equation 3.2. The additional parameters is δ_i. Using equation 3.10, it is easily verified that the first-order derivative with respect to δ_i, evaluated at $\delta_i = 0$ is given by

$$d(\theta_n, \delta_i) = \begin{cases} \dfrac{\left(x_{ni} - P_i(\theta_n)\right)(1 - c_i)\phi(a_i(\theta_n - b_i))}{P_i(\theta_n)(1 - P_i(\theta_n))} & \text{if } y_n = 1 \\ \\ 0 & \text{if } y_n = 0, \end{cases} \qquad (3.12)$$

where $\phi(a_i(\theta_n - b_i))$ is the normal density function evaluated at $a_i(\theta_n - b_i)$. Combining this with equation 3.11 and equation 3.6 gives the desired LM statistic. Note that $x_{ni} - P_i(\theta_n)$ can be seen as a residual. It is a quadratic form depending on squared residuals with a covariance matrix as a matrix of weights. The covariance matrix is essential in the derivation of the asymptotic distribution of the statistic. However, for an analogous PPC this is not necessary because the distribution of the statistic is generated. Therefore, we only take into account simply weighted the residuals and define

$$T(X, \phi) = \sum_g \frac{\left[\sum_{n|g} x_{ni} - P_i(\theta_n)\right]^2}{\sum_{n|g} P_i(\theta_n)(1 - P_i(\theta_n))}, \qquad (3.13)$$

where the summations are over groups g and respondents n in group g, respectively.

The alternative model to test the ICCs is given by equation 3.3. The additional parameters is δ_{ig} and $d(\theta_n, \delta_{ig})$ is non-zero if the total score (disregarding the responses on item i) is in the range g. For the PPC, the discrepancy measure given by equation 3.13 can be used, except that the summation is now over score ranges g. Analogously, the alternative model to test local independence is given by equation 3.4 and $d(\theta_n, \delta_{il})$ is non-zero if $x_{il} = 1$. The discrepancy measure for a PPC becomes

$$T(X,\phi) = \frac{\left[\sum_n x_{nl}\left(x_{ni} - P_i(\theta_n)\right)\right]^2}{\sum_n x_{nl} P_i(\theta_n)(1 - P_i(\theta_n))} \,. \tag{3.14}$$

3.6 Simulation Design

Two characteristics of testing applications were manipulated: the number of items (10, 20, and 40) and the sample size (100, 400, and 1000). To simulate item responses, an ability parameter was randomly drawn from a standard normal distribution. The difficulty and discrimination parameters for fitting items were drawn from standard normal and log-normal distributions, respectively. For the 3PNO model, the guessing parameter was fixed at 0.20. The misfit was introduced by generating responses using non-zero values for the δ-parameter. The δ-parameter had values 0.5 or 1.0. To prevent unrealistic parameter values, the discrimination and difficulty parameters for the misfitting items were fixed at 1.0 and 0.0, respectively. The number of items showing model violations varied from 10% to 20% of the test length. For every simulated data set MML and Bayesian estimates were computed. The MML estimates were computed as proposed by Bock and Aitkin (1981) and the Bayesian estimates were computed using the method by Albert (1992) with non-informative priors where the discrimination parameters were constrained to be positive. The procedure had a run length of 4,000 iterations with a burn-in period of 1,000 iterations. Finally, for every condition, 100 replications were simulated, and for every statistic the proportion of replications with a *p*-value less than .05 was determined under both frameworks.

3.7 Results

3.7.1 Differential Item Functioning

The tables discussed in this section show how the power of the Bayesian and frequentist procedure fluctuated due to the combinations of test length (denoted by K), effect size (denoted by δ), percentage of DIF items (varied as 10% to 20%) and sample size (denoted by N). Tables 3.1 and 3.2 shows the power and Type I error rate of PPCs and LM procedures for the 2PNO model and the 3PNO model, respectively. The columns labeled “10%” and “20%” shows the proportion of replications for which the test on the differential item functioning was significant at the 5% level. So these columns give an estimate of the power of test under both frameworks. The columns under the label “Power” give the proportion of replications with a *p*-value less than .05 for misfitting items, that is, for items which were correctly flagged, the columns under the label “Type I error rate” give the proportion of replications with a *p*-value less than .05 for fitting items, that is, for items which were incorrectly flagged.

Table 3.1. The Power and Type I error by test length, effect size and sample size under the 2-PNO model.

			Power				Type I error rate			
			PPC		LM		PPC		LM	
			Number of Items with DIF				Number of Items with DIF			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	100	0.11	0.08	0.42	0.32	0.01	0.02	0.07	0.07
		400	0.78	0.61	0.89	0.71	0.02	0.04	0.06	0.09
		1000	1.00	1.00	0.99	0.99	0.03	0.07	0.05	0.07
	1.0	100	0.81	0.63	0.84	0.76	0.01	0.03	0.11	0.15
		400	1.00	1.00	1.00	1.00	0.03	0.10	0.07	0.08
		1000	1.00	1.00	1.00	1.00	0.06	0.08	0.07	0.09
20	0.5	100	0.24	0.20	0.50	0.42	0.02	0.02	0.08	0.07
		400	0.86	0.73	0.89	0.83	0.02	0.02	0.06	0.06
		1000	1.00	1.00	1.00	0.99	0.02	0.03	0.05	0.08
	1.0	100	0.68	0.73	0.95	0.86	0.02	0.02	0.08	0.09
		400	1.00	1.00	1.00	1.00	0.02	0.04	0.06	0.08
		1000	1.00	1.00	1.00	1.00	0.03	0.06	0.08	0.09
40	0.5	100	0.21	0.20	0.48	0.47	0.02	0.02	0.11	0.14
		400	0.83	0.78	0.89	0.88	0.02	0.02	0.06	0.07
		1000	0.99	0.99	1.00	1.00	0.02	0.04	0.06	0.08
	1.0	100	0.74	0.73	0.95	0.92	0.02	0.02	0.11	0.12
		400	1.00	1.00	1.00	1.00	0.02	0.04	0.06	0.09
		1000	1.00	1.00	1.00	1.00	0.03	0.07	0.06	0.10

In general, for the power we see the expected main effects of effect size, test length and sample size. The power of the LM test was higher than the power of the PPCs. Note that the LM test had the largest power; in many instances the power was equal to 1.00.

Comparing Table 3.1 and table 3.2, it can be seen that the detection rates were generally higher in all combinations under 2PNO model than under the 3PNO model. The reason is that latter model has more item parameters which have to be estimated than the 2PNO model. This leads to a loss in precision of estimates and power. Note further that samples of 100 were insufficient to generate the necessary statistical power. Generally, N = 500 is the minimum sample size recommended for estimating two and three parameter logistic models (Hulin, Lissak, & Drasgow, 1982).

Table 3.2. The Power and Type I error by test length, effect size and sample size under the 3-PNO model.

			Power				Type I error rate			
			PPC		LM		PPC		LM	
			Number of Items with DIF				Number of Items with DIF			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	100	0.10	0.06	0.18	0.09	0.01	0.02	0.06	0.09
		400	0.50	0.35	0.79	0.60	0.02	0.03	0.06	0.07
		1000	0.88	0.79	1.00	0.99	0.02	0.05	0.05	0.06
	1.0	100	0.27	0.15	0.72	0.51	0.01	0.01	0.11	0.15
		400	0.97	0.86	1.00	1.00	0.02	0.06	0.08	0.10
		1000	1.00	1.00	1.00	1.00	0.05	0.07	0.07	0.08
20	0.5	100	0.10	0.08	0.25	0.12	0.02	0.02	0.06	0.08
		400	0.53	0.40	0.84	0.75	0.02	0.01	0.05	0.04
		1000	0.94	0.96	1.00	1.00	0.02	0.03	0.04	0.05
	1.0	100	0.44	0.29	0.78	0.60	0.01	0.02	0.07	0.08
		400	1.00	0.97	0.99	0.99	0.02	0.03	0.05	0.06
		1000	1.00	1.00	1.00	1.00	0.03	0.05	0.05	0.06
40	0.5	100	0.08	0.08	0.32	0.20	0.02	0.02	0.09	0.12
		400	0.52	0.45	0.86	0.76	0.03	0.03	0.05	0.07
		1000	0.94	0.91	1.00	1.00	0.02	0.03	0.05	0.06
	1.0	100	0.37	0.34	0.79	0.68	0.02	0.03	0.12	0.14
		400	0.99	0.99	1.00	1.00	0.02	0.03	0.06	0.08
		1000	1.00	1.00	1.00	1.00	0.03	0.07	0.05	0.05

For LM test the control of Type I error rate, that is, the proportion of false alarms remained generally close to the nominal significance level. The degrees of misfit and sample size to some extent have moderate effect on the Type I error rates. The main exceptions occurred for the large effect size with a short test and small sample. The explanation is that in these cases the imposed model violation was such that every combination led to a global violation affecting all items. While for PPCs, the Type I error rates were quite conservative with the exceptions of few ones. The possible explanation may be due to twice use of data as a result posterior p-values were conservative (i.e., often failed to detect model misfit) and inadequate behavior of posterior p-value. It should be noted that in simulation studies without misfitting items (not shown here) the Type I error rate was always close to a nominal significance level of 5% for the LM test and conservative for the PPCs.

3.7.2 Local Independence (LI)

To evaluate the detection rate of local independence, a number of simulation studies were carried out. These studies generally had the same setup as the DIF study. Test statistics were computed in the same way as previous ones. Tables 3.3 and 3.4 show the power and Type I error rate of LM and PPCs as function of sample size, test length, degree of misfit, and the number of misfit items. The optimal condition for the detection of LI was a large sample size, large effect size and a large test length. The main overall trend was that the detection rate decreased with small sample size, short test and small effect size. For LM test the power is uniformly higher and approach to unity in most of the combinations. For the PPCs the proportions of hits were comparable but lower than the LM test. The detection rates of LI were generally higher in all the combinations under 2PNO than 3PNO model.

There was also tendency that the detection rates for DIF were higher than LI. The detection rates decreased as the number of misfit items increased under both frameworks. The Type I error rates are conservative and quite below than the nominal significance level for PPCs. The false alarm rates for the LM test, generally, close to the significance level. The inflation occurred with the increase of misfit items and large effect size.

Table 3.3. The Power and Type I error by test length, effect size and sample size under the 2-PNO model.

			Power				Type I error rate			
			PPC		LM		PPC		LM	
			Number of Items with LOC				Number of Items with LOC			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	100	0.10	0.09	0.22	0.23	0.01	0.00	0.06	0.06
		400	0.12	0.11	0.60	0.47	0.01	0.02	0.06	0.08
		1000	0.59	0.24	0.96	0.85	0.01	0.02	0.06	0.08
	1.0	100	0.17	0.13	0.71	0.52	0.01	0.01	0.10	0.12
		400	0.72	0.13	1.00	0.90	0.01	0.02	0.08	0.10
		1000	0.99	0.73	1.00	0.97	0.01	0.02	0.05	0.07
20	0.5	100	0.17	0.11	0.54	0.35	0.01	0.02	0.06	0.07
		400	0.21	0.14	0.89	0.75	0.01	0.02	0.08	0.09
		1000	0.77	0.67	1.00	0.97	0.01	0.02	0.05	0.06
	1.0	100	0.22	0.10	0.86	0.83	0.01	0.01	0.09	0.09
		400	0.93	0.84	1.00	0.99	0.01	0.02	0.09	0.10
		1000	1.00	1.00	1.00	1.00	0.01	0.03	0.06	0.08
40	0.5	100	0.12	0.13	0.34	0.36	0.01	0.02	0.05	0.05
		400	0.28	0.25	0.48	0.44	0.01	0.03	0.06	0.05
		1000	0.81	0.78	0.98	0.86	0.01	0.02	0.05	0.06
	1.0	100	0.23	0.16	0.49	0.52	0.01	0.02	0.06	0.07
		400	0.94	0.89	1.00	1.00	0.01	0.02	0.06	0.06
		1000	1.00	1.00	1.00	1.00	0.01	0.03	0.05	0.06

3.7.3 Shape of the ICCs

The detection rates and error rates for shape of ICC were shown in tables 3.5 and 3.6. The study had the same setup as the DIF and LI studies. The large sample size and test length was the optimal condition for the detection of items that violates the assumption of ICC. Again, there were clear main effects of the effect size, the sample size and the test length. The main overall trend was that the detection rate decreased in the combinations where sample size and number of items were small. For the LM test the power is uniformly higher and approach to unity in most of the combinations as compared to PPCs. The detection rates of ICC were generally higher in the combinations under the 2PNO than under the 3PNO model.

Table 3.4. The Power and Type I error by test length, effect size and sample size under the 3-PNO model.

			Power				Type I error rate			
			PPC		LM		PPC		LM	
			Number of Items with LOC				Number of Items with LOC			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	100	0.08	0.06	0.31	0.25	0.01	0.00	0.05	0.07
		400	0.10	0.09	0.65	0.32	0.01	0.02	0.05	0.07
		1000	0.35	0.17	0.84	0.66	0.01	0.02	0.06	0.07
	1.0	100	0.15	0.11	0.65	0.49	0.01	0.01	0.10	0.12
		400	0.48	0.10	0.98	0.76	0.01	0.02	0.09	0.11
		1000	0.98	0.49	1.00	0.89	0.01	0.02	0.06	0.07
20	0.5	100	0.14	0.10	0.41	0.31	0.01	0.02	0.06	0.07
		400	0.14	0.12	0.62	0.51	0.01	0.02	0.05	0.07
		1000	0.60	0.50	0.97	0.84	0.01	0.02	0.06	0.09
	1.0	100	0.16	0.03	0.93	0.73	0.01	0.01	0.08	0.09
		400	0.87	0.63	1.00	0.94	0.01	0.02	0.09	0.09
		1000	1.00	0.99	1.00	0.98	0.01	0.03	0.08	0.09
40	0.5	100	0.10	0.10	0.36	0.34	0.01	0.02	0.07	0.08
		400	0.18	0.15	0.40	0.33	0.01	0.03	0.05	0.05
		1000	0.67	0.59	0.91	0.82	0.01	0.02	0.05	0.05
	1.0	100	0.11	0.09	0.52	0.49	0.01	0.02	0.08	0.09
		400	0.80	0.69	1.00	0.92	0.01	0.02	0.05	0.06
		1000	1.00	0.99	1.00	0.98	0.01	0.03	0.05	0.05

There was also tendency that the detection rates for ICC were lower than LI and DIF. Analogous to the simulations of local independence, the Type I error rates are conservative and quite below than the nominal significance level for the PPCs and false alarm rates for the LM test, generally, close to the significance level.

Table 3.5. The Power and Type I error by test length, effect size and sample size under the 2-PNO model.

			Power				Type I error rate			
			PPC		LM		PPC		LM	
			Number of Items with ICC				Number of Items with ICC			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	100	0.06	0.04	0.55	0.40	0.01	0.00	0.05	0.06
		400	0.14	0.10	0.53	0.48	0.01	0.02	0.06	0.07
		1000	0.59	0.24	0.77	0.61	0.01	0.02	0.05	0.05
	1.0	100	0.17	0.13	0.97	0.77	0.01	0.01	0.05	0.06
		400	0.72	0.23	0.97	0.78	0.01	0.02	0.05	0.05
		1000	0.90	0.73	1.00	0.98	0.01	0.02	0.05	0.06
20	0.5	100	0.17	0.11	0.65	0.61	0.01	0.02	0.05	0.05
		400	0.31	0.24	0.88	0.71	0.01	0.02	0.05	0.05
		1000	0.77	0.67	0.81	0.68	0.01	0.02	0.05	0.06
	1.0	100	0.20	0.10	0.97	0.90	0.01	0.01	0.06	0.06
		400	0.83	0.74	0.95	0.87	0.01	0.02	0.05	0.05
		1000	1.00	1.00	1.00	0.98	0.01	0.03	0.05	0.06
40	0.5	100	0.21	0.11	0.47	0.40	0.01	0.02	0.05	0.06
		400	0.38	0.25	0.98	0.92	0.01	0.03	0.06	0.07
		1000	0.81	0.78	1.00	0.96	0.01	0.02	0.05	0.05
	1.0	100	0.23	0.16	0.53	0.49	0.01	0.02	0.06	0.06
		400	0.94	0.89	0.99	0.96	0.01	0.02	0.05	0.06
		1000	1.00	1.00	1.00	1.00	0.01	0.03	0.05	0.05

Table 3.6. The Power and Type I error by test length, effect size and sample size under the 3-PNO model.

			Power				Type I error rate			
			PPC		LM		PPC		LM	
			Number of Items with ICC				Number of Items with ICC			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	100	0.08	0.05	0.49	0.38	0.01	0.00	0.06	0.05
		400	0.12	0.07	0.44	0.19	0.01	0.02	0.07	0.04
		1000	0.35	0.14	0.70	0.17	0.01	0.02	0.06	0.04
	1.0	100	0.13	0.06	0.75	0.47	0.01	0.01	0.06	0.06
		400	0.48	0.15	0.93	0.34	0.01	0.02	0.07	0.06
		1000	0.88	0.34	0.98	0.45	0.01	0.02	0.06	0.06
20	0.5	100	0.14	0.12	0.81	0.67	0.01	0.02	0.05	0.06
		400	0.24	0.15	0.38	0.20	0.01	0.02	0.05	0.06
		1000	0.60	0.50	0.95	0.54	0.01	0.02	0.06	0.05
	1.0	100	0.26	0.17	0.94	0.86	0.01	0.01	0.06	0.06
		400	0.77	0.27	0.84	0.43	0.01	0.02	0.05	0.06
		1000	0.90	0.49	0.97	0.65	0.01	0.03	0.06	0.05
40	0.5	100	0.11	0.10	0.47	0.40	0.01	0.02	0.05	0.06
		400	0.88	0.65	0.98	0.92	0.01	0.03	0.06	0.05
		1000	0.97	0.69	1.00	0.96	0.01	0.02	0.05	0.05
	1.0	100	0.11	0.09	0.83	0.79	0.01	0.02	0.06	0.07
		400	0.80	0.69	0.99	0.96	0.01	0.02	0.06	0.06
		1000	1.00	0.99	1.00	1.00	0.01	0.03	0.06	0.05

3.8 An Empirical Example

The main purpose of this example is to show the type of outcome to expect in a well-fitting data set. The example pertains to data of the central examination in Secondary Education (MAVO-D level) of language proficiency in German in the Netherlands. This centralized examination is a high-stakes test. The data were collected in 1995. The examination consisted of 50 items with 2 response categories of each. The total sample used here consisted of 2021 students. The item parameters were estimated by MML and MCMC assuming standard normal distributions for the θ-parameters. Only the MML analysis will be shown in detail, because the results for MCMC were comparable. We chose to present the complete tables with the statistics on all 50 items to give a realistic impression of the output to expect.

Table 3.7 gives the results for the LM test of the ICCs obtained using the 1PNO model and 2PNO model, respectively. The test was based on a partition of the score range into three subsets. The column labeled ‘LM’ gives the values of the LM-statistics; the column labeled ‘Prob’ gives the significance probabilities. The statistics have 2 degrees of freedom. It can be verified that 27 and 7 of the 50 LM-tests were significant at a 5% significance level for 1PNO and 2PNO model, respectively. If the model holds, the outcomes should be (approximately) uniform and the number of significant tests at a 5% significance level should be approximately 2.5. So the conclusion is that the 1PNO did not fit well and the 2PNO fitted reasonably well.

Table 3.7. Outcomes of LM tests for ICCs for examination data.

	1 PNO Model		2 PNO Model		Group 1		Group 2		Group 3	
Item	LM	Prob.	LM	Prob.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
1	2.44	.29	0.64	.73	.41	.40	.52	.53	.67	.66
2	11.74	.00	2.17	.34	.82	.83	.88	.87	.90	.90
3	0.86	.65	1.45	.49	.52	.52	.66	.67	.80	.79
4	2.56	.28	2.48	.29	.25	.25	.35	.37	.53	.52
5	0.63	.73	7.19	.03	.79	.79	.90	.89	.94	.95
6	0.26	.88	5.09	.08	.15	.16	.25	.25	.39	.38
7	0.77	.68	0.28	.87	.87	.87	.92	.93	.96	.96
8	1.72	.42	0.85	.65	.94	.93	.96	.96	.97	.98
9	2.08	.35	0.82	.66	.62	.62	.78	.78	.88	.88
10	18.91	.00	0.74	.69	.85	.85	.94	.94	.98	.98
11	2.95	.23	2.08	.35	.85	.84	.91	.92	.97	.96
12	165.18	.00	13.19	.00	.84	.84	.95	.96	.00	.99
13	0.28	.87	0.11	.95	.35	.34	.49	.50	.65	.65
14	7.15	.03	8.02	.02	.76	.74	.84	.86	.94	.93
15	14.95	.00	0.74	.69	.90	.90	.96	.97	.99	.99
16	7.56	.02	3.23	.20	.42	.43	.59	.60	.77	.75
17	11.81	.00	8.15	.02	.75	.75	.84	.83	.87	.89
18	11.23	.00	3.38	.18	.92	.92	.98	.98	.99	.99
19	42.29	.00	11.38	.00	.19	.17	.16	.20	.25	.24

Table 3.7. Continued

20	51.77	.00	9.26	.01	.29	.27	.32	.33	.38	.39
21	24.76	.00	3.66	.16	.47	.45	.54	.55	.63	.64
22	3.11	.21	0.34	.84	.37	.37	.50	.50	.62	.63
23	32.83	.00	3.17	.20	.80	.80	.92	.92	.98	.97
24	2.05	.36	0.67	.72	.61	.62	.75	.74	.82	.83
25	1.52	.47	0.92	.63	.47	.47	.60	.61	.74	.73
26	17.17	.00	3.77	.15	.66	.65	.80	.82	.93	.91
27	11.02	.00	1.76	.42	.29	.28	.44	.45	.65	.65
28	1.48	.48	1.16	.56	.47	.47	.61	.62	.77	.76
29	32.53	.00	2.28	.32	.76	.76	.90	.91	.97	.97
30	3.81	.15	2.89	.24	.33	.32	.45	.48	.65	.64
31	12.82	.00	0.13	.94	.45	.45	.56	.55	.66	.66
32	3.82	.15	3.36	.19	.46	.45	.56	.59	.73	.72
33	7.37	.03	1.24	.54	.44	.43	.53	.55	.67	.66
34	18.24	.00	4.92	.09	.27	.25	.31	.34	.45	.44
35	11.26	.00	10.47	.01	.67	.65	.84	.84	.92	.93
36	2.39	.30	2.77	.25	.49	.51	.69	.67	.79	.80
37	4.73	.09	0.23	.89	.78	.78	.86	.85	.90	.90
38	6.45	.04	10.10	.01	.67	.67	.80	.78	.84	.87
39	10.61	.00	2.94	.23	.60	.62	.80	.79	.89	.89
40	21.78	.00	0.92	.63	.44	.43	.51	.52	.63	.62
41	2.17	.34	1.19	.55	.50	.51	.62	.63	.75	.74
42	51.63	.00	2.20	.33	.59	.60	.82	.82	.93	.93
43	2.27	.32	3.00	.22	.47	.47	.60	.62	.76	.74
44	44.74	.00	0.81	.67	.75	.75	.91	.91	.97	.97
45	4.62	.10	0.10	.95	.52	.52	.64	.64	.75	.75
46	36.93	.00	0.21	.90	.54	.54	.77	.77	.89	.90
47	3.98	.14	0.46	.79	.41	.41	.60	.59	.74	.75
48	15.54	.00	4.37	.11	.53	.55	.76	.74	.86	.87
49	3.96	.14	0.65	.72	.40	.40	.51	.52	.65	.65
50	25.31	.00	3.05	.22	.76	.75	.89	.90	.97	.97

Table 3.8. Outcomes of LM tests for ICCs for examination data combined with linking group data.

	1 PNO Model		2 PNO Model		Group 1		Group 2		Group 3	
Item	LM	Prob.	LM	Prob.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.
1	7.86	.02	4.56	.10	.41	.39	.52	.52	.67	.64
2	18.50	.00	2.89	.24	.82	.83	.88	.86	.90	.90
3	0.15	.93	0.14	.93	.52	.52	.66	.66	.80	.78
4	1.53	.46	2.93	.23	.25	.24	.35	.36	.53	.51
5	18.45	.00	48.08	.00	.79	.77	.90	.87	.94	.93
6	0.92	.63	0.46	.79	.15	.16	.25	.24	.39	.37
7	20.06	.00	17.62	.00	.87	.86	.92	.91	.96	.94
8	18.00	.00	18.17	.00	.94	.92	.96	.95	.97	.97
9	1.12	.57	0.09	.95	.62	.62	.78	.77	.88	.88
10	9.18	.01	0.50	.78	.85	.85	.94	.94	.98	.98
11	0.37	.83	3.85	.15	.85	.83	.91	.92	.97	.96
12	30.76	.00	5.21	.07	.84	.84	.95	.96	.00	.99
13	3.98	.14	7.51	.02	.35	.36	.49	.51	.65	.66
14	3.75	.15	12.57	.00	.76	.73	.84	.86	.94	.93
15	3.69	.16	3.54	.17	.90	.89	.96	.97	.99	.99
16	11.16	.00	7.65	.02	.42	.44	.59	.61	.77	.76
17	7.76	.02	4.65	.10	.75	.74	.84	.83	.87	.89
18	6.06	.05	8.47	.01	.92	.91	.98	.98	.99	.00
19	31.69	.00	48.76	.00	.19	.20	.16	.24	.25	.29
20	15.65	.00	16.73	.00	.29	.25	.32	.37	.38	.52
21	11.08	.00	10.78	.00	.47	.49	.54	.56	.63	.64
22	6.07	.05	76.84	.00	.37	.44	.50	.53	.62	.62
23	5.16	.08	0.99	.61	.80	.80	.92	.92	.98	.97
24	12.06	.00	6.70	.04	.61	.59	.75	.73	.82	.84
25	13.68	.00	21.28	.00	.47	.43	.60	.60	.74	.74
26	8.99	.01	10.40	.01	.66	.66	.80	.82	.93	.91
27	15.30	.00	21.47	.00	.29	.31	.44	.47	.65	.64
28	4.76	.09	19.22	.00	.47	.50	.61	.64	.77	.76
29	18.20	.00	0.81	.67	.76	.76	.90	.90	.97	.96
30	0.71	.70	2.35	.31	.33	.32	.45	.46	.65	.62
31	5.07	.08	1.46	.48	.45	.45	.56	.56	.66	.67
32	4.09	.13	8.29	.02	.46	.44	.56	.58	.73	.71

Table 3.8. Continued

33	5.37	.07	1.09	.58	.44	.43	.53	.55	.67	.66
34	14.92	.00	4.99	.08	.27	.25	.31	.34	.45	.45
35	7.48	.02	29.83	.00	.67	.64	.84	.83	.92	.93
36	3.37	.18	2.92	.23	.49	.51	.69	.67	.79	.80
37	1.80	.41	1.75	.42	.78	.79	.86	.86	.90	.90
38	5.56	.06	10.48	.01	.67	.66	.80	.78	.84	.87
39	11.16	.00	3.01	.22	.60	.61	.80	.78	.89	.89
40	12.74	.00	3.23	.20	.44	.44	.51	.53	.63	.63
41	3.78	.15	13.08	.00	.50	.52	.62	.65	.75	.76
42	26.29	.00	4.53	.10	.59	.58	.82	.81	.93	.93
43	2.14	.34	2.34	.31	.47	.47	.60	.62	.76	.75
44	63.92	.00	120.11	.00	.75	.70	.91	.86	.97	.94
45	5.79	.06	1.34	.51	.52	.51	.64	.64	.75	.75
46	91.61	.00	46.13	.00	.54	.58	.77	.79	.89	.91
47	11.82	.00	8.08	.02	.41	.43	.60	.60	.74	.76
48	13.29	.00	12.42	.00	.53	.53	.76	.73	.86	.86
49	2.33	.31	4.72	.09	.40	.41	.51	.53	.65	.65
50	4.79	.09	16.08	.00	.76	.73	.89	.89	.97	.96

The last six columns give the observed average item scores and expected average item scores in the three sub-groups under the headings 'Obs' and 'Exp', respectively. Such information may be helpful in assessing the severity of the model violation in case of a significant test result.

Since the data fitted the 2PNO model quite well, the next question addresses is whether that fit could be aggravated. Fortunately, data from a linking study was also available, where the responses of 1033 test takers collected in a non-high stakes situation. It was expected that these students might be less motivated and that this would lead to a lesser model fit. The results are displayed in Table 3.8. The numbers of significant item tests rose to 29 and 25 for the 1PNO model and 2PNO model, respectively. So the expectation of a decrease in model fit was confirmed.

Table 3.9 and 3.10 give the analogous results for the LM test of the local dependence. The tests are targeted at the dependence between an item *i* and the previous item *i*-1. The columns under the labels "X=0" and "X=1" give observed and expected average score on item *i* given a incorrect or correct score on an item *i*-1, respectively. The numbers of significant tests were 14 (1PNO) and 5 (2PNO) for the examination data and 15 (1PNO) and 16 (2PNO) for the examination data combined with the liking group data. Again, the 2PNO model fitted best, and the introduction of the linking group data decreased the model fit.

Table 3.9. Outcomes of LM tests for local independence for examination data.

		1 PNO Model		2 PNO Model		X = 0		X = 1	
Item i	Item i-1	LM	Prob.	LM	Prob.	Obs.	Exp.	Obs.	Exp.
2	1	0.03	.85	1.32	.25	.85	.86	.88	.88
3	2	2.23	.14	3.59	.06	.56	.61	.67	.66
4	3	0.00	.98	0.02	.89	.32	.32	.41	.41
5	4	0.22	.64	0.88	.35	.86	.86	.90	.90
6	5	0.18	.67	0.25	.62	.19	.20	.28	.28
7	6	0.84	.36	1.81	.18	.91	.91	.95	.94
8	7	0.90	.34	1.50	.22	.92	.94	.96	.96
9	8	0.45	.50	0.22	.64	.67	.69	.76	.76
10	9	4.54	.03	0.68	.41	.87	.88	.94	.94
11	10	0.99	.32	2.68	.10	.88	.84	.91	.91
12	11	3.04	.08	0.01	.91	.86	.86	.94	.94
13	12	0.00	.99	0.34	.56	.34	.32	.51	.51
14	13	7.07	.01	5.34	.02	.80	.82	.89	.87
15	14	1.53	.22	0.02	.90	.92	.92	.96	.96
16	15	1.12	.29	0.07	.79	.40	.41	.60	.60
17	16	0.66	.42	3.12	.08	.77	.79	.85	.84
18	17	11.85	.00	7.16	.01	.92	.94	.97	.97
19	18	0.01	.91	1.95	.16	.11	.16	.21	.20
20	19	1.13	.29	0.08	.78	.33	.33	.35	.35
21	20	0.81	.37	0.06	.80	.53	.54	.57	.56
22	21	2.89	.09	0.81	.37	.48	.47	.51	.52
23	22	7.16	.01	2.24	.13	.87	.88	.93	.92
24	23	0.73	.39	1.18	.28	.60	.63	.74	.74
25	24	1.10	.30	0.64	.42	.56	.54	.61	.62
26	25	0.05	.83	0.63	.43	.76	.75	.82	.82
27	26	0.67	.41	0.31	.58	.35	.34	.49	.49

Table 3.9. Continued

28	27	1.18	.28	0.72	.40	.55	.56	.68	.67
29	28	0.01	.91	2.13	.14	.85	.84	.89	.90
30	29	2.58	.11	1.24	.26	.32	.34	.50	.49
31	30	0.24	.62	0.77	.38	.52	.53	.60	.59
32	31	4.64	.03	2.75	.10	.57	.55	.59	.61
33	32	0.06	.81	1.46	.23	.49	.50	.58	.57
34	33	0.36	.55	3.96	.05	.30	.32	.38	.37
35	34	5.19	.02	3.03	.08	.78	.79	.86	.85
36	35	5.82	.02	4.53	.03	.51	.55	.69	.69
37	36	0.55	.46	3.07	.08	.80	.82	.87	.86
38	37	0.16	.69	0.87	.35	.71	.73	.78	.78
39	38	8.14	.00	5.72	.02	.66	.70	.80	.78
40	39	10.38	.00	3.28	.07	.51	.47	.54	.55
41	40	4.52	.03	2.00	.16	.61	.60	.63	.64
42	41	4.73	.03	0.40	.53	.72	.73	.82	.82
43	42	0.58	.45	0.32	.57	.49	.50	.64	.64
44	43	1.34	.25	0.13	.72	.84	.84	.90	.90
45	44	1.21	.27	0.26	.61	.54	.53	.65	.65
46	45	2.73	.10	0.07	.80	.66	.67	.76	.76
47	46	11.85	.00	5.10	.02	.43	.46	.63	.62
48	47	25.85	.00	15.38	.00	.61	.65	.80	.77
49	48	0.04	.84	0.26	.61	.44	.45	.55	.55
50	49	6.40	.01	2.09	.15	.84	.85	.91	.90

3.9 Conclusions

In the present study, a number of analogous tests for model violations to unidimensional item response models in a frequentist and Bayesian framework were compared. The LM tests procedure which is based on the MML framework is practical and useful tool for the evaluation of model fit. The LM statistics are useful because they are item oriented diagnostic tools, which give an indication of the source of model violations. Potentially, they offer the possibility of directed model relaxation to obtain sufficient model fit.

The main advantage of the Bayesian PPC procedure is that many model violations for all items can be assessed without complicated computations. They will gain in interest when they will be applied to more complex IRT models such as multidimensional models (Béguin & Glas, 2001) and multilevel IRT models (Fox & Glas, 2001, 2003).

The simulation studies showed that the LM test had good power characteristics and Type I error rates approximately equal to the nominal significance level. The PPCs have comparable power characteristics to the LM test except for short tests and small samples. The Type I error rates for the Bayesian procedure were conservative and well below than the nominal significance level. There was a clear tendency that both procedures were more efficient in flagging misfitting items in 2PNO model than in the 3PNO model.

Table 3.10. Outcomes of the LM tests for local independence for examination data combined with linking group data.

		1 PNO Model		2 PNO Model		X = 0		X = 1	
Item i	Item i-1	LM	Prob.	LM	Prob.	Obs.	Exp.	Obs.	Exp.
2	1	1.07	.30	0.34	.56	.85	.85	.88	.87
3	2	1.84	.17	3.31	.07	.56	.61	.67	.66
4	3	0.36	.55	0.22	.64	.32	.31	.41	.40
5	4	13.23	.00	16.41	.00	.86	.84	.90	.89
6	5	0.00	.99	0.05	.83	.19	.19	.28	.27
7	6	11.06	.00	7.62	.01	.91	.90	.95	.92
8	7	0.09	.77	0.54	.46	.92	.94	.96	.95
9	8	0.37	.54	0.29	.59	.67	.69	.76	.76
10	9	4.03	.04	0.82	.37	.87	.88	.94	.94
11	10	1.35	.25	3.66	.06	.88	.84	.91	.91
12	11	2.91	.09	0.01	.93	.86	.86	.94	.94
13	12	0.11	.74	0.00	.97	.34	.34	.51	.52
14	13	4.73	.03	2.93	.09	.80	.81	.89	.88
15	14	0.81	.37	0.15	.70	.92	.91	.96	.96
16	15	1.46	.23	0.18	.67	.40	.42	.60	.61
17	16	0.87	.35	2.36	.12	.77	.79	.85	.84
18	17	8.46	.00	3.76	.05	.92	.94	.97	.97
19	18	0.75	.39	5.47	.02	.11	.19	.21	.25
20	19	37.01	.00	63.56	.00	.33	.38	.35	.41
21	20	1.08	.30	5.43	.02	.53	.55	.57	.58

Table 3.10. Continued

22	21	0.83	.36	7.48	.01	.48	.51	.51	.54
23	22	1.10	.30	1.83	.18	.87	.88	.93	.92
24	23	0.13	.72	0.16	.69	.60	.61	.74	.73
25	24	6.45	.01	5.12	.02	.56	.52	.61	.61
26	25	0.00	.98	0.01	.94	.76	.76	.82	.83
27	26	1.30	.25	0.98	.32	.35	.37	.49	.50
28	27	3.40	.07	9.08	.00	.55	.58	.68	.68
29	28	0.01	.93	3.59	.06	.85	.84	.89	.90
30	29	1.62	.20	0.63	.43	.32	.34	.50	.48
31	30	0.23	.63	2.63	.10	.52	.53	.60	.60
32	31	7.50	.01	6.07	.01	.57	.54	.59	.60
33	32	0.22	.64	1.40	.24	.49	.50	.58	.57
34	33	0.69	.41	4.04	.04	.30	.32	.38	.37
35	34	1.30	.25	0.00	.99	.78	.78	.86	.84
36	35	8.11	.00	5.51	.02	.51	.55	.69	.69
37	36	1.29	.26	4.05	.04	.80	.82	.87	.86
38	37	0.15	.70	0.61	.43	.71	.73	.78	.78
39	38	8.30	.00	4.98	.03	.66	.70	.80	.78
40	39	6.82	.01	1.96	.16	.51	.48	.54	.56
41	40	0.55	.46	0.00	.99	.61	.61	.63	.66
42	41	1.48	.22	0.16	.69	.72	.72	.82	.81
43	42	0.72	.40	0.25	.62	.49	.50	.64	.64
44	43	16.86	.00	32.54	.00	.84	.79	.90	.86
45	44	1.45	.23	0.48	.49	.54	.53	.65	.65
46	45	12.09	.00	5.86	.02	.66	.69	.76	.78
47	46	17.50	.00	10.14	.00	.43	.48	.63	.63
48	47	13.82	.00	5.90	.02	.61	.64	.80	.76
49	48	0.09	.76	1.15	.28	.44	.46	.55	.56
50	49	0.93	.34	0.04	.84	.84	.83	.91	.89

4

Assessing Person Fit: A Comparative Study of Frequentist and Bayesian Frameworks

Abstract: Item response theory (IRT) models are used to model the likelihood that a test taker of a particular level of ability will answer a particular test item correctly. In reality, test takers may not always respond to test items in ways that are consistent with the IRT model. Some possible reasons for this “nonfitting” behavior include test anxiety, guessing, cheating in achievement tests, or faking responses in personality inventories. Several person fit statistics have been proposed to detect item score patterns that do not fit an IRT model. Traditionally, IRT models are evaluated in a frequency based framework, usually the maximum likelihood framework. However, nowadays, a Bayesian framework is emerging as an alternative. This paper compares the Type I error rate and power of a number of well known person fit statistics in both frameworks.

In a frequentist framework, Snijders (2001) presented a general framework for deriving the asymptotic null distribution for statistics which are linear in the item responses. This allows the standardization of linear person fit statistics with an estimated ability parameter. Further, a Lagrange multiplier tests for person fit is considered. To support a comparison with the Bayesian framework, all these tests are reformulated to the normal ogive representation of IRT models. Posterior predictive checks (PPCs) are a much used Bayesian model-checking tool. The person fit tests considered in the frequentist approach are redefined as PPCs. The results of simulation studies show that the power of the tests is greater in a frequentist framework and Type I error rate was conservative in the Bayesian framework.

This chapter has been submitted for publication as: M.N. Khalid and Cees A. W. Glas, Assessing Person Fit: A Comparative Study of Frequentist and Bayesian Frameworks.

4.1 Introduction

In the context of item response theory (IRT) modeling, several methods have been proposed to detect item score patterns that are not in agreement with the item score pattern expected based on a particular test model. These item score patterns should be detected because scores of such persons may not be adequate descriptions of their trait level. This area of research is commonly referred to as person fit research, and the majority of the research on person fit has concentrated on the development of statistics that can be used to identify nonfitting response vectors (van Krimpen-Stoop & Meijer, 1999; Meijer & Sijtsma, 2001). A response pattern is considered nonfitting or aberrant if it is found to be unlikely given the model according to a person fit statistics (PFS). These fit statistics focus on the appropriateness of the stochastic model on the level of the individual. For this reason they are commonly called person fit statistics. The causes for person misfit are many; all are factors extraneous to the measured ability that systematically and significantly affect performance and lead to inaccurate measurement of the ability. Person fit methods may help to identify invalid outcomes of a test caused by, for example, a lack of motivation to take the test seriously, concentration problems, an alignment error in marking the answer sheets, selective test preparation strategies, and faking on a personality test.

In the IRT context several person fit statistics have been proposed that can be used to detect individual item score patterns that do not fit the IRT model (Levine & Rubin, 1979; Wright & Stone, 1979; Tatsuoka, 1984; Smith, 1985, 1986; Klauer & Rettig, 1990; Drasgow, Levine, & McLaughlin, 1991; Glas & Dagohoy, 2007). See Meijer and Sijtsma (2001) for a review.

One of the difficulties in the assessment of person fit is the fact that the ability parameter of the examinee is unknown and needs to be estimated. The use of an estimated rather than the true value of the ability parameter has an effect on the distribution of the person fit statistic (Snijders, 2001). These estimates usually decrease the asymptotic variance of most statistics proposed in the literature. Therefore, their asymptotic distribution is usually unknown. To solve this problem, Snijders (2001) proposed a method for standardization of a specific class of person fit statistics for dichotomous items, such that their asymptotic distribution can be properly derived. Snijders (2001) applied the correction for the mean and variance of l_z and derived an asymptotic normal approximation for the conditional distributions of l_z when $\hat{\theta}$ is used in its calculation. Snijders performed a simulation study for relatively small tests consisting of 8 and 15 items and for large tests consisting of 50 and 100 items, fitting the two-parameter logistic model (the 2PL model), and estimating θ by maximum likelihood. The results showed that the correction was satisfactory at Type I error levels of $\alpha = .05$ and $\alpha = .10$ but that the empirical Type I error was smaller than the nominal Type I error for smaller values of α.

An alternative approach to account for the fact that the ability parameter is estimated are the Lagrange multiplier tests (LM tests) for person fit proposed by Glas and Dagohoy (2007). The purpose of the LM test is to compare two models, a model under the null-hypothesis and a more general model that is derived from the model under the null-hypothesis by adding parameters. Only the model under the null-hypothesis needs to be estimated. Simulation studies show that estimation of the item parameters has little effect on distribution so only the effect of estimation of θ is taken into account in the LM statistics.

Problems of sampling distributions are also avoided in a Bayesian framework in combination with Markov chain Monte Carlo (MCMC) computational methods. Exponential advances in current computing capabilities have made it possible to address the problem of the distribution of PFS using computationally intensive simulation based methods. An example are the posterior predictive checks (PPCs) for person fit the proposed by Glas and Meijer (2003). In this approach, samples from posterior distribution $p(\theta \mid X)$ are drawn, and these samples are used to generate replicate data X^{rep} that conform to the model. Using the data and replicate data, discrepancy measures $T(X, \theta)$ and $T(X^{rep}, \theta)$ are computed and the Bayesian equivalent of the p-value is approximated by determining the proportion of simulated values of $T(X^{rep}, \theta)$ at least as extreme as simulated values of $T(X, \theta)$. Compared to the traditional frequentist approach, the 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. Second, the person fit statistic may depend on unknown quantities such as person parameters. This uncertainty is explicitly taken into account. However, Bayarri and Berger (2000) have shown that using PPCs also comes with a problem. It often fails to detect model misfit and it had less than adequate behavior of posterior p-values (Berkhof, van Mechelen, & Gelman, 2002).

The purpose of this study is to describe and compare the person fit statistics that take in account Snijders correction, LM tests, and Bayesian procedures (PPCs) for evaluating goodness of fit in unidimensional dichotomous IRT models. These frameworks will be compared with respect to effects of sample size, test length, and the percentage of misfitting score patterns on Type I error rate and power. This chapter is organized as follows. First, the person fit statistics that are examined in this study are presented. The second section introduces the description of the LM statistic, Snijders' correction procedure and PPCs. The third section outlines the design of the simulation studies. Next, results from a study comparing empirical Type I error rates and power for the above frameworks are presented. Next an empirical data example is presented to show how the frameworks compare in practice. Finally, some conclusions are drawn, and some suggestions for further research are given.

4.2 Person Fit Statistics

Meijer and Sijtsma (2001) have presented a comprehensive overview of person fit statistics that are widely used in practice. A number of these statistics will be used below. A short description of IRT models is presented prior to the presentation of the fit statistics.

4.2.1 IRT Models

In the present study, we consider two- and three-parameter normal ogive models for dichotomously scored items (the 2PNO model and the 3PNO model). The normal ogive models were used because of computational advantages in the Bayesian framework (see, Albert, 1992) and to support direct comparability across the two frameworks. In the 3PNO model, the item is characterized by a difficulty parameter b_i, a discrimination parameter a_i and a guessing parameter c_i. Further, θ_n is the latent ability parameter of respondent n. The probability of correctly answering an item (denoted by $X_i = 1$) is given by

$$P_i \;=\; P(X_i = 1 \mid \theta) \;=\; c_i + (1 - c_i)\Phi(a_i(\theta - b_i)) \;, \tag{4.1}$$

where $\Phi(.)$ is

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-t^2/2\right)dt \;.$$

If the guessing parameter c_i is constrained to zero the model reduces to the 2PNO model and if also the discrimination parameter a_i is constrained to one the model reduces to the 1PNO model.

4.2.2 W Statistic

The W statistic (Wright & Stone, 1979) is defined by

$$W = \frac{\sum_{i=1}^{k} [X_i - P_i]^2}{\sum_{i=1}^{k} P_i[1 - P_i)]} \;, \tag{4.2}$$

where the difference between item score X_i and the expected item score P_i is weighted by the variance of the item score.

4.2.3 UB Statistic

A related statistic was proposed by Smith (1985, 1986), in which set of items is divided into S non-overlapping subsets denoted A_S (s = 1,…,S). Because the test length is the sample size of the statistic and test length is relatively short, usually $S = 2$. Then the unweighted between-sets fit statistic UB is defined as

$$UB = \frac{1}{S-1} \sum_{s=1}^{S} \frac{\left[\sum_{i \in A_S} (X_i - P_i)\right]^2}{\sum_{i \in A_S} P_i(1 - P_i)} \;. \tag{4.3}$$

The UB Statistics is a weighted W statistic computed at the subset level.

4.2.4 ξ_1 Statistic

Two complementary statistics, ξ_1 and ξ_2, were proposed by Tatsuoka (1984). The ξ_1 statistic is the standardization with a mean of zero and unit variance of

$$\xi_1 = \sum_{i=1}^{k}[P_i - X_i](n_i - \overline{n}) \ , \tag{4.4}$$

where n_i denotes the number of correct answers to item *i*, and $\overline{n}$ denotes the mean number of correctly answered items in the test. The index will be positive when easy items are incorrectly answered and difficult items are correctly answered, and it will also be positive if the number of correctly answered items deviates from the overall mean score of the respondents. If a response pattern is misfitting in both senses, the magnitude of the index will be largely positive.

4.2.5 ξ_2 Statistic

The ξ_2 statistic is a standardization of

$$\xi_2 = \sum_{i=1}^{k}[P_i - X_i](P_i - R/k) \ , \tag{4.5}$$

where *R* is the person's number-correct score on the test. The index will be positive if the response pattern is misfitting in the sense that easy items are incorrectly answered and difficulty items are correctly answered; the overall response tendencies of the total sample of persons are not important here.

4.2.6 Log-likelihood Statistic

Another well-known person fit statistic is the log-likelihood statistic which was first purposed by Levine and Rubin (1979). It is defined as

$$\ell = \sum_{i=1}^{k} \{X_i \log P_i + (1 - X_i) \log[1 - P_i]\} . \tag{4.6}$$

It was further developed by Drasgow, Levine, and Williams (1985) and Drasgow, Levine, and McLaughlin (1991).

4.2.7 l_z Statistic

Drasgow et al. (1985) proposed a standardized version l_z of ℓ, which is asymptotically standard normally distributed; l_z is defined as

$$l_z = \frac{l - E(l)}{(\mathrm{var}(l))^{1/2}} , \tag{4.7}$$

where $E(l)$ and $\mathrm{var}(l)$ denote the expectation and the variance of l, respectively.

4.3 Snijders' Correction Procedure

Deciding whether a response pattern is aberrant boils down to evaluating whether the outcome of a PFS for a particular response pattern is extreme in the conditional null distribution of the PFS given *θ*. There are two important criteria for evaluating the performance of a PFS (Drasgow, Levine, & McLaughlin, 1987). The first is standardization. A well standardized PFS has consistent conditional null distribution, which makes it possible to compare all values of the PFS to one cut score for all values of *θ*. The other criterion is power, which is a function of many different factors including test length, scoring method, ability level, and type and severity of aberrance.

A very well known likelihood-based PFS for assessing overall goodness of fit to two- and three-parameter logistic IRT models is the log-likelihood of a person's observed item scores (Levine & Rubin, 1979). Drasgow, Levine, and Williams (1985) developed a standardized version of ℓ, l_z, which is assumed to have a standard normal distribution across different levels of θ. However, the standardization that Drasgow et al. had assumed is inappropriate. Various researchers have found that even when true ability is used to calculate l_z, the conditional null distributions of l_z are neither standard normal nor consistent across different θ values. Instead, the distributions are negatively skewed, and often leptokurtic (Meijer & Sijtsma, 2001; Molenaar & Hoijtink, 1990; Nering, 1995). When the true θ is replaced by its estimate $\hat{\theta}$ in the calculation of l_z, the problems were exacerbated; variable errors at different levels of the estimated ability cause the variance of l_z to be very inconsistent across the θ levels. Therefore, empirical Type I error rates are lower than nominal levels, and person misfit is underestimated. In addition, the statistic has been shown to have relatively low power in detecting misfitting response patterns (van Krimpen-Stoop & Meijer, 1999).

To deal with this problem, Snijders (2001) introduced a correction for the mean and variance for a class of statistics which are linear in the item responses and derived an asymptotic normal approximation for the conditional distributions when $\hat{\theta}$ is used in its calculation. All statistics introduced above, except UB belong to this class. As an example, we introduce the approximation for the l_z statistic. For the exact derivation and for the expressions for other statistics refer to Appendix A.

The centred version of the fit statistic ℓ as given in (4.6) is written as

$$w_i(\theta) = \sum_{i=1}^{n} (X_i - \widehat{P}_i)^2 \omega_i(\theta) , \tag{4.8}$$

where weight is $\omega_i(\theta) = \log \frac{\widehat{P}_i}{(1-\widehat{P}_i)}$ and θ is the MLE of the ability parameter. To obtain a better approximation of the normal distribution, the weight $\omega_i(\theta)$ is modified as

$$w_i(\theta)' = \omega_i(\theta) - c_n(\theta) r_i(\theta) , \tag{4.9}$$

where $r_i(\theta) = \dfrac{P_i^/}{P_i(1-P_i)}$, $c_i(\theta) = \dfrac{\sum_{i=1}^{n} P_i^/ \omega_i(\theta)}{\sum_{i=1}^{n} P_i^/ r_i(\theta)}$ and $P_i^/$ is the first order derivative of P_i with respect to θ. The expected value and variance of the modified PFS are

$$E(w_i(\theta)) \approx -c_i(\theta) r_0(\theta) \;, \tag{4.10}$$

$$\mathrm{Va\,r}(w_i(\theta)) \approx n\tau_i^{\,2}(\theta) \;, \tag{4.11}$$

where $\tau_i^{\,2}(\theta) = \dfrac{1}{n}\sum_{i=1}^{n} \; w^2{}_i(\theta)^/ P_i(1-P_i)$ and $r_0(\theta) =$ 0 for maximum likelihood estimates.

4.4 An LM Test for Constancy of Theta

Item-oriented LM tests for IRT models have been proposed by Glas (1998, 1999), Glas and Suárez-Falcón (2003), and Jansen and Glas (2005). 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 modeling (Sörbom, 1989). The purpose of the LM test is to compare two models, a model under the null-hypothesis and a more general model that is derived from the model under the null-hypothesis by adding parameters. The important advantage of the LM test over likelihood ratio tests and Wald tests is that only the model under the null-hypothesis needs to be estimated. Recently, LM tests for person fit have been proposed by Glas and Dagohoy (2007). These authors use the logistic representation of IRT models. In the present application we use the normal ogive representation. The model under the null-hypothesis is the 3PNO model. Under the alternative hypothesis, it is assumed that there is a subset of items, say $A_,$ where the ability parameter is shifted, that is, the probability of a correct response is given by

$$P_i \;=\; P(X_i = 1 \mid \theta, i \in A) \;=\; c_i + (1 - c_i)\Phi(a_i(\theta + \delta - b_i)\,. \quad (4.12)$$

To test the constancy of theta across the response pattern, we define the hypotheses

$$\mathrm{H_o} \;:\; \delta = 0$$

and

$$\mathrm{H_1} \;:\; \delta \neq 0.$$

The test statistic is given by

$$\mathrm{LM} \;=\; \frac{\left(\dfrac{\partial \ell}{\partial \delta}\right)^2}{-\dfrac{\partial^2 \ell}{\partial \delta^2} + \left(\dfrac{\partial^2 \ell}{\partial \delta\, \partial \theta}\right)^2 \left(\dfrac{\partial^2 \ell}{\partial \theta^2}\right)^{-1}}\,, \qquad (4.13)$$

where ℓ stands for the log-likelihood function as defined in (4.6). Taking first and second order derivatives with respect to some item parameter results in the well-known general expressions

$$\ell' = \sum_i \frac{(X_i - P_i)P_i'}{P_i(1 - P_i)}$$

and

$$\ell'' = \sum_i \frac{(P_i')^2}{P_i(1 - P_i)}\;.$$

Further, for the 3PNO model we have $P_i' = (1 - c_i)\phi_i$, with $\phi_i = \left(1/\sqrt{2\pi}\right)\exp\left[-\left(a_i\left(\theta - b_i\right)\right)^2/2\right]$, and expression 4.13 becomes

$$\mathrm{LM} = \frac{\left(\displaystyle\sum_{i \in A} \frac{(X_i - P_i)a_i(1 - c_i)\phi_i}{P_i(1 - P_i)}\right)^2}{-\displaystyle\sum_{i \in A} \frac{a_i^2(1 - c_i)^2\phi_i^2}{P_i(1 - P_i)} + \left(\displaystyle\sum_{i \in A} \frac{a_i^2(1 - c_i)^2\phi_i^2}{P_i(1 - P_i)}\right)^2 \left(\displaystyle\sum_{i=1}^{K} \frac{a_i^2(1 - c_i)^2\phi_i^2}{P_i(1 - P_i)}\right)^{-1}}\;. \qquad (4.14)$$

The statistic has an asymptotic χ^2 distribution with one degree of freedom. If the absolute values of these derivatives are large, the fixed parameters are bound to change once they are set free. As a result, the test is significant, that is, the special model is rejected. If the absolute values of these derivatives are small, the fixed parameters will probably show little change should they be set free. So the test is not significant, that is, the special model is not rejected.

In the framework of the logistic representation of IRT, Glas and Dagohoy (2007) point out that the LM statistic for the constancy of theta can be viewed as a UB statistic corrected for the estimation of the ability parameter. In the normal ogive representation this relation between UB and LM is less obvious due to the presence of the factors $P_i' = (1 - c_i)a_i\phi_i$. In the logistic framework we have $P_i' = P_i a_i (1 - P_i)$.

4.5 Posterior Predictive Checks

Bayesian statistics has received considerable attention in statistics over the past decade. There has been a recent surge in the use of Bayesian estimation in IRT. Albert (1992), Patz and Junker (1999a, 1999b), Bradlow, Wainer, and Wang (1999), Janssen, Tuerlinckx, Meulders, and De Boeck (2000), Béguin and Glas (2001), Fox and Glas (2001, 2003), Bolt, Cohen, and Wollack (2002), Sinharay, Johnson, and Williamson (2003), and Wollack, Cohen, and Wells (2003) are only some of the recent examples of application of Bayesian estimation and MCMC algorithms (e.g., Gelman, Carlin, Stern, & Rubin, 2004) to fit complicated psychometric models. There are numerous instances in which classical methods fail, but the Bayesian approach offers a feasible method for assessing model fit. However, little attention has been given to assessing fit of IRT models from a Bayesian perspective.

In Bayesian statistics, a model can be checked in at least three ways: (1) examining sensitivity of inferences to reasonable changes in the prior distribution and the likelihood; (2) checking that the posterior inferences are reasonable, given the substantive context of the model; and (3) checking that the model fits the data. We address the third of these concerns using the posterior predictive distribution for a discrepancy, an extension of classical test statistics to allow dependence on unknown (nuisance) parameters. Posterior predictive assessment was introduced by Guttman (1967), applied by Rubin (1981), and given a formal Bayesian definition by Rubin (1984).

The posterior predictive model checks (PPCs) is a popular Bayesian model checking tool because of its simplicity. The method primarily consists of comparing the observed data with replicated data (those predicted by the model) using a number of test statistics. The important applications of the PPCs method in educational measurement and IRT context include Rubin and Stern (1994), Hoijtink and Molenaar (1997), Scheines, Hoijtink and Boomsma, 1999, Albert and Gosh (2000), Janssen, et al. (2000), van Onna (2003), Glas and Meijer (2003), and Fox and Glas (2001, 2003). The idea of PPCs is to generate simulated values from the posterior predictive distributions of replicated data and to compare these samples to the observed data. If the replicated data and the observed data differ systematically, it is an indication of a potential model misfit. The framework of PPCs can be defined as follows.

Letting X be the observed data and ϕ be the vector of all the parameters in the model, we then define $p(X|\phi)$ as the likelihood and $p(\phi)$ as the prior distribution on the parameters. The posterior predictive checking method suggests checking a model by examining whether the observed data appear extreme with respect to the posterior predictive distribution of replicated data, which is obtained by

$$p(X^{rep} \mid X) = \int p(X^{rep} \mid \phi) p(\phi \mid X) d\phi \,. \tag{4.15}$$

Usually, discrepancy measures or a test quantities $T(X,\phi)$ are defined, and the posterior distribution of $T(X,\phi)$ is compared to the posterior predictive distribution of $T(X^{rep},\phi)$, with substantial difference between them indicate model fit. The posterior predictive checking method allows a reasonable summary of such comparisons with the posterior predictive p- value (PPP-value):

$$\Pr(T(X^{rep},\phi) \geq T(X,\phi) \mid X) = \int_{T(X^{rep},\phi) \geq T(X,\phi)} p(X^{rep} \mid \phi) p(\phi \mid X) dX^{rep} d\phi \tag{4.16}$$

PPP-values that are close to 0 or 1 are indicative of model misfits.

Because of the difficulty in dealing with equations 4.15 or 4.16 analytically for all but simple problems, Rubin (1984) suggested simulating replicate data sets using MCMC from the posterior predictive distribution in practical applications of the PPCs method. The standard methods of MCMC estimation include the Gibbs sampler and the more general Metropolis-Hastings (MH) algorithm. The Gibbs sampler is a special case of the MH algorithm in which parameters are sampled from their full conditional distributions (rather than a relatively arbitrary proposal distribution). The algorithm generates a Markov chain of iterates that constitute a random walk over the posterior distribution, the results of which are equivalent to integrating the density over the data to obtain a sample from the posterior distribution for the parameters.

One generates *N* draws $\phi^1, \phi^2, \phi^3, ..., \phi^N$ from the posterior distribution $P(\phi \mid X)$ of ϕ, and draws $X^{rep,n}$ from the likelihood distribution $P(X \mid \phi^n)$, *n = 1, 2, . . . , N*. The process results in *N* replicated data sets. Equation 4.16 implies that the proportion of the *N* replications for which $T\,(X^{rep,n})$ exceeds *T*(*X*) provides an estimate of the PPP-value. Extreme PPP-values (close to 0, or 1, or both, depending on the nature of the test statistic) indicate model misfit. The details of MCMC framework for 2PNO model and 3PNO model can be found in Glas and Meijer (2003) and Béguin and Glas (2001).

4.6 Simulation Design

The simulation study consisted of two parts. In the first part, the Type I error rate as a function of test length and sample size was investigated. In the second part, the detection rates of the different statistics for different model violations, test lengths, and sample sizes were investigated. In all reported simulation studies, the statistics *l*, *W*, *UB*, ξ_1, ξ_2 and LM were used as defined above. The simulation studies were performed under 2PNO and 3PNO models. The data were generated using fixed item parameters; the ability parameters were drawn from a standard normal distribution. The guessing parameters c_i were fixed to .20 for all items. Item difficulty and discrimination parameters were chosen as follows:

- For a test length K = 30, three values of the discrimination parameter, a_i, 0.5, 1.0, and 1.5, were crossed with 10 item difficulties $b_i = -2.00 + 0.40(i - 1)$, $i = 1, \ldots, 10$.

- For a test length K = 60, three values of discrimination parameters, a_i, 0.5, 1.0, and 1.5, were crossed with 20 item difficulties $b_i = -2.00 + 0.20(i - 1)$, $i = 1, \ldots, 20$.

Data matrices with two samples sizes were used: N = 400 and N = 1,000. For MML the item parameters were fixed at their true values. For MCMC, the item parameters were re-estimated for each data set. The non-informative priors and true values of the parameters were used as starting values for the MCMC procedure. The procedure had a run length of 4,000 iterations with a burn-in period of 1,000 iterations. That is, the first 1,000 iterations were discarded. In the remaining 3,000 iterations, $T(X^{rep}, \xi)$ and $T(X, \xi)$ were computed for every iteration. So the posterior predictive checks were based on 3,000 draws. For the statistics that use a partitioning of the items into subtests, the items were ordered according to their item difficulty and then two subtests of equal size were formed, one with the difficult and one with the easy items. Finally, for every condition, 100 replications were made, and the proportion of replications with a *p* value less than .05 was determined under both frameworks. So for N = 400, a number of 40,000 replications of each statistics were made and for N = 1000, a number of 100,000 replications were made. The frequentist versions of the statistics were computed with Snijders' correction, except for the UB statistic where this correction is not available.

4.7 Results

4.7.1 Type I Error Rate

The results for the Bayesian and frequentist frameworks are shown in Table 4.1. The test characteristics that were manipulated (shown in the tables) include the sample size denoted by N, test length denoted by K, and the person fit statistics denoted by PFS. It can be seen that in general, the significance probabilities converge to their nominal value of .05 as a function of sample size and test length, and the nominal significance probability was best approximated by the combination of a test length

K= 60 and a sample size N = 400 or N = 1,000 in both frequentist and the Bayesian frameworks. Note that for K = 30, the significance probabilities were slightly smaller than the significance probabilities under the frequentist framework. In general, significance probabilities are conservative in the Bayesian approach. There are no clear effects for specific person fit statistics, except that the *UB* seemed to be quite conservative under both frame works. Finally, the Snijders' correction that was applied to derive the null distribution for PFS works quite well and error rates approaches to the nominal significance level. Further, also the significance probabilities of the LM test were very close to the nominal significance level.

4.7.2 Detection Rates for Guessing

To evaluate the detection rate of guessing, a number of simulation studies were carried out. These studies generally had the same setup as the studies of the Type I error rate reported above. Item parameters were as above, unless reported otherwise. The data were generated in such a way that guessing occurred for 10% of the simulees, so data matrices with N = 400 simulees had 40 aberrant simulees, and data matrices with N = 1,000 simulees had 100 aberrant simulees. For these aberrant simulees, guessing was imposed in three conditions, where 1/6, 1/3, or 1/2 of the test scores were corrupted by random responding. So for the test with K = 30 items, the number of corrupted items was either 5, 10, or 15, and for the test with K = 60 items, the number of corrupted items was either 10, 20, or 30. Guessing was always imposed on the items with the lowest item difficulty. This was done because guessing on the easiest items has the most detrimental effect on the estimation of θ (Meijer & Nering, 1997), and thus, detection of these item score patterns is important. The probability of a correct response to these items by aberrant simulees was .20.

Table 4.1. Type I error rates under the Bayesian and Frequentist frameworks**.**

		Bayesian				Frequentist	
		K=30		K=60		K=30	K=60
Model	PFS	N=400	N=1000	N=400	N=1000	N=1000	N=1000
2PNO	LM					0.048	0.052
	UB	0.028	0.035	0.037	0.012	0.011	0.013
	Like	0.039	0.031	0.038	0.033	0.037	0.066
	W	0.041	0.039	0.044	0.031	0.046	0.061
	Zeta1	0.042	0.041	0.047	0.040	0.045	0.048
	Zeta2	0.078	0.062	0.065	0.062	0.051	0.062
3PNO							
	LM					0.045	0.051
	UB	0.023	0.025	0.027	0.012	0.001	0.010
	Like	0.029	0.031	0.032	0.033	0.036	0.062
	W	0.031	0.032	0.034	0.031	0.038	0.053
	Zeta1	0.047	0.043	0.041	0.040	0.043	0.049
	Zeta2	0.075	0.061	0.065	0.065	0.046	0.051

Bayesian methods used posterior estimates of items parameters. Frequentist methods used true values of item parameter estimates. Hence there is no column for N=400 in frequentist case, because the sample size for the estimation of item parameters plays no role share.

Test statistics were computed in the same way as in the Type I error rates. The both procedures were run using the data of all simulees, both the aberrant and nonaberrant ones. 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. As in the previous study, for the computation of statistics based on a partitioning of the test, two subtests of equal size were formed: a difficult and an easy one. As a result, the corrupted items were in the easiest test, although the partitioning did not completely conform to the pattern of corrupted and uncorrupted items. So in this sense, the partition was not optimal.

The proportions of “hits”, that is, the proportion of correctly identified aberrant simulees, are shown in Table 4.2. The proportions of “false alarms”, that is, the proportion of normal simulees incorrectly identified as aberrant, are not shown because they were analogous to the Type I error rates under both frameworks. The main overall trend for all tests was that the detection rate increased for 2PNO model than 3PNO model. The detection rates were higher for identification of guessing simulees in frequentist framework than Bayesian.

The optimal condition for the detection of guessing was a large sample size and a large test length. Therefore, the results of the condition with N = 1,000 simulees and K = 60 items is discussed first. The main overall trend for all tests was that the detection rate decreased as the number of affected items increased. This can be explained by the bias in the ability parameters of simulees with nonfitting response vectors. It can be seen that the ability estimates for the misfitting simulees was grossly inflated because of misfit items for p = 1/2 and p = 1/3 than for p = 1/6 accordingly. It can also be concluded that the presence of 10% misfitting simulees in the calibration sample affected the ability estimates for the fitting simulees to some degree. As the number of affected items increased, the ability estimates become biased, and because the fit statistics were computed conditionally on θ, the detection rate decreased. Inspection of the results in the condition with N = 400 simulees and K = 60 shows that the detection rate was little affected by the smaller calibration sample.

For a test length of K = 30 items, the detection rate was slightly less than for K = 60 items. This was as expected, because the statistics were computed on an individual level, and on this level, the test length was the number of observations on which the test was based. Finally, the false alarm rates were well controlled with N = 400, 1000 and K = 30, 60 under both frame works. However, false alarm rates were slightly less than the nominal level with K = 30 in frequentist frame work. But on other hand overall they are conservative in the Bayesian frame work.

Table 4.2. Detection rates for guessing simulees under the Bayesian and Frequentist frameworks.

		K=30						K=60					
		Bayesian			Frequentist			Bayesian			Frequentist		
Mode	PFS	1/6	1/3	1/2	1/6	1/3	1/2	1/6	1/3	1/2	1/6	1/3	1/2
2PNO	LM				0.911	0.865	0.665				0.839	0.789	0.656
	UB	0.762	0.519	0.112	0.854	0.748	0.366	0.829	0.756	0.641	0.820	0.755	0.657
	Like	0.729	0.527	0.163	0.977	0.955	0.903	0.747	0.742	0.632	0.995	0.986	0.963
	W	0.725	0.524	0.131	0.973	0.948	0.496	0.789	0.761	0.635	0.989	0.981	0.816
	Zeta1	0.744	0.683	0.412	0.921	0.849	0.621	0.802	0.796	0.698	0.960	0.887	0.627
	Zeta2	0.803	0.718	0.608	0.972	0.967	0.857	0.889	0.897	0.855	0.991	0.984	0.904
3PNO													
	LM				0.751	0.734	0.646				0.569	0.450	0.347
	UB	0.542	0.238	0.167	0.667	0.620	0.392	0.594	0.351	0.247	0.519	0.445	0.335
	Like	0.516	0.232	0.191	0.881	0.871	0.787	0.556	0.322	0.251	0.880	0.693	0.558
	W	0.472	0.251	0.212	0.821	0.786	0.432	0.552	0.332	0.262	0.753	0.696	0.555
	Zeta1	0.363	0.269	0.198	0.660	0.585	0.520	0.370	0.232	0.317	0.408	0.271	0.238
	Zeta2	0.556	0.319	0.259	0.865	0.835	0.777	0.608	0.373	0.295	0.840	0.689	0.549

The detection rates for N=400 were analogous.

4.7.3 Detection Rates for Item Disclosure

In high-stakes testing, persons may be tempted to obtain knowledge about the type of test questions or even about the correct answers to the items in the test. In computerized adaptive testing, this is one of the major threats to the validity of test scores. But also in standardized paper-and-pencil tests this is a realistic problem. Item preknowledge on a few items will only have a minor effect on the number-correct score (Meijer & Nering, 1997). Also, item preknowledge of the correct answers on the easiest items in the test will only slightly improve the number-correct score. This suggests that in particular item preknowledge on the items of median difficulty and on the most difficult items may have an effect on the total score. Thus, the effect of item preknowledge will be important in particular for persons with a low ability level who answer many difficult items correctly.

The setup of the simulation study to the detection rate of the tests for item disclosure was analogous to the study to the detection rate for guessing. So data were generated for sample sizes of N = 400 and N = 1,000 simulees and test lengths of K = 30 and K = 60 items; item disclosure was prominent for 10% of the simulees; and for these simulees, 1/6, 1/3, or 1/2 of the difficult items in the test were corrupted. The probability of a correct response to these items was chosen to be .80. Test statistics were computed in the same way as in the guessing study.

The proportions of hits are shown in Table 4.3. The false alarms were analogous to Type I error rates so they are not presented. Now the items in the second part of the test, that is, the difficult items, were affected by the model violations. It can be concluded that the effects of test length and proportion of affected items are also found here. Furthermore, the absence of an effect of calibration sample size was replicated.

The optimal condition for the detection of item disclosure was a large sample size and a large test length. The results showed that in general under 2PNO model the detection rates were higher under both approaches. Over all there was tendency that the detection rates were higher for the LM test and fit statistics for which Snijders' correction take into account. There was also tendency that proportion of hits inflated

Table 4.3. Detection rates for item disclosure under the Bayesian and Frequentist frameworks.

		K=30						K=60					
		Bayesian			Frequentist			Bayesian			Frequentist		
	PFS	1/6	1/3	1/2	1/6	1/3	1/2	1/6	1/3	1/2	1/6	1/3	1/2
2PNO	LM				0.528	0.686	0.756				0.567	0.727	0.798
	UB	0.209	0.372	0.421	0.251	0.429	0.597	0.612	0.657	0.631	0.646	0.652	0.689
	Like	0.651	0.478	0.511	0.651	0.861	0.901	0.611	0.638	0.673	0.927	0.975	0.988
	W	0.372	0.519	0.522	0.673	0.780	0.897	0.617	0.649	0.619	0.916	0.922	0.926
	Zeta1	0.453	0.541	0.601	0.628	0.747	0.761	0.577	0.744	0.752	0.667	0.779	0.795
	Zeta2	0.474	0.541	0.601	0.737	0.892	0.918	0.769	0.846	0.857	0.899	0.945	0.960
3PNO													
	LM				0.338	0.641	0.692				0.352	0.691	0.714
	UB	0.113	0.294	0.352	0.211	0.389	0.527	0.241	0.423	0.594	0.241	0.419	0.597
	Like	0.131	0.319	0.394	0.513	0.771	0.831	0.258	0.536	0.512	0.681	0.871	0.921
	W	0.151	0.379	0.382	0.543	0.716	0.737	0.313	0.574	0.511	0.653	0.806	0.797
	Zeta1	0.249	0.546	0.578	0.513	0.532	0.691	0.496	0.748	0.756	0.608	0.737	0.741
	Zeta2	0.321	0.629	0.678	0.673	0.752	0.824	0.564	0.838	0.861	0.707	0.872	0.898

The detection rates for N=400 were analogous in conditions with increase of misfit items. The detection rates of *Zeta2* were relatively higher and for *UB* were relatively low among other tests. In general it's revealed that detection rates were higher for guessing than item disclosure. The guessing violation is more severe than item disclosure. In item disclosure the item parameters were not changed, so for the aberrant simulees the probabilities of correct responses were uniformly shifted for the affected part of the test. In guessing it implies that the original items parameters lose their meaning, that is, all items are equally difficult.

4.8 An Empirical Example

The main purpose of this example is to show the type of outcome to expect in a well-fitting data set and to assess the agreement between the two estimation frameworks. The example pertains to data of the central examination in Secondary Education (MAVO-D level) of language proficiency in German in the Netherlands. This centralized examination is a high-stakes test. The data were collected in 1995. The examination consisted of 50 items with 2 response categories of each. The total sample used here consisted of 2021 students. The item parameters were estimated by MML and MCMC assuming standard normal distributions for the θ-parameters. For frequentist framework, given the MML estimates of the item parameters, the θ- parameters were estimated by maximum likelihood (ML), and the fit statistics were computed. In the Bayesian MCMC framework, all parameters were estimated concurrently with non-informative priors for the item parameters (see Albert, 1992).

Table 4.4 gives the results of detection rate in percentages for person fit statistics under frequentist and the Bayesian framework at 5% and 20% level of significance. The column labeled 'PFS' denotes the fit statistics that were evaluated; the next four columns show the number of persons that were identified as aberrant by each PFS under the respective frameworks. The statistics computed in the frequentist framework had higher detection rates than the PPCs computed in the Bayesian framework. The UB statistic had the lowest detection rate, probably because Snijders' correction could not be applied.

Table 4.4. Detection rate of fit statistics in percentages.

PSF	Frequentist 5%	Bayesian 5%	Frequentist 20%	Bayesian 20%
Lz	4.8	2.7	18.8	3.0
W	4.2	1.8	16.8	2.9
Zeta 1	9.6	2.7	16.8	6.0
Zeta 2	12.6	2.5	17.1	7.5
UB	0.9	0.3	8.2	1.4
LM	6.5	-	23.9	-

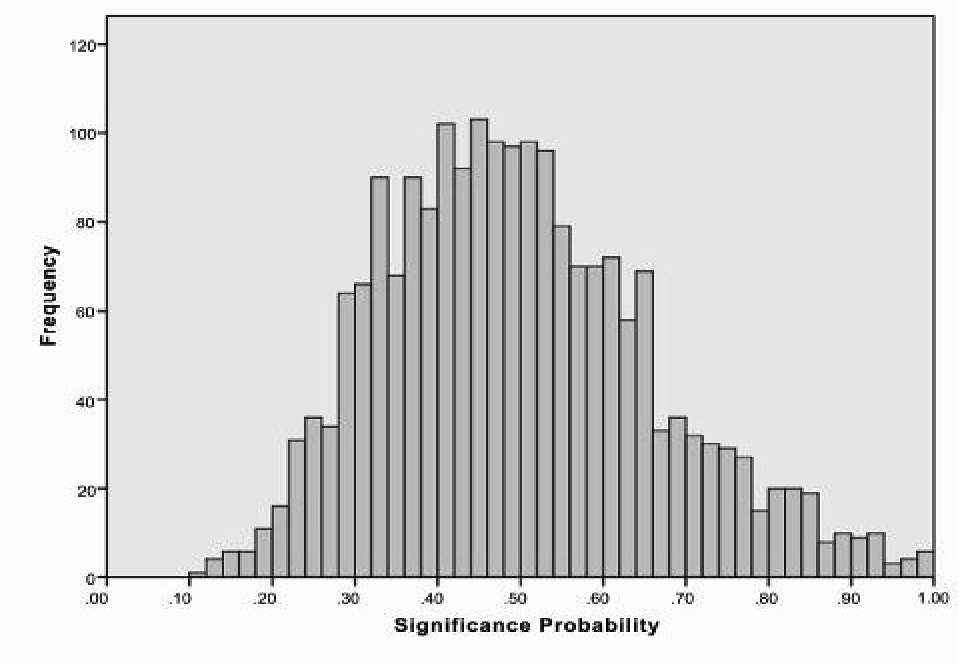

Figure 4.1. Distribution of significance probabilities of UB in the Bayesian framework.

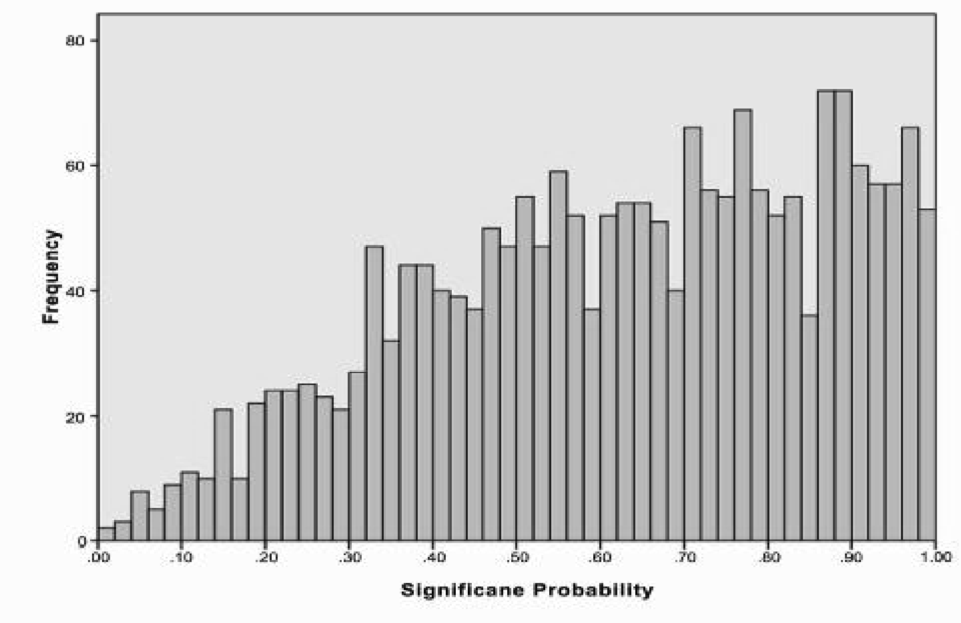

Figure 4.2. Distribution of significance probabilities of UB in the frequentist framework.

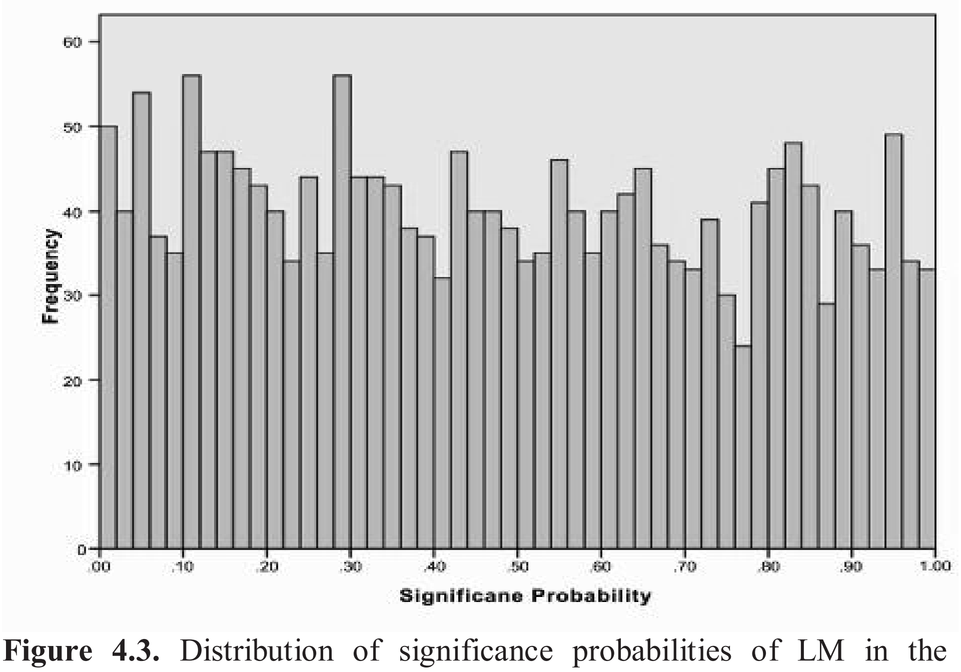

Figure 4.3. Distribution of significance probabilities of LM in the frequentist framework.

Figures 4.1, 4.2 and 4.3 display the distributions of significance probabilities of the UB test in the Bayesian and frequentist framework, and distribution of significance probabilities of the LM test. Above it was argued that the LM statistic could be seen as a corrected UB statistic. Comparison of the three figures shows that the number of significant outcomes is much lower for the two UB tests. The distribution of the Bayesian version of UB has more kurtosis than the other two. This is in line with the usual findings on the behavior of posterior p-values (Bayarri & Berger, 2000; Berkhof, van Mechelen, & Gelman, 2002). Further, compared to the distribution of LM, the frequentist version of UB is clearly skewed to the right. As a result the number of significant outcomes is low. Finally, the outcomes of the LM test are a very good approximation to a uniform distribution. The number of values of LM significant at 10% was 225 (out of 2021 students). Further, Pearson's X^2-test on the frequencies in the 10 deciles yielded a value of 2.35 with 9 degrees of freedom and a significance probability of 0.98. If the model holds, the outcomes should be (approximately) uniform. The "qualification approximate" is used because the significance probabilities are not direct observations but (functions of) estimates, but the influence of this is negligible. So the result can be seen as good support for model fit.

4.9 Discussion and Conclusions

Aberrant response behavior in psychological and educational testing may result in inadequate measurement. Therefore, it is important to detect misfitting item scores. To classify an item score pattern as nonfitting, the researcher can simply take the top 1% or top 5% of aberrant cases. As an alternative, the researcher can use a theoretical sampling distribution or simulate reference data based on the estimated item parameters in the sample. Person fit statistics can be used as descriptive statistics. In this study, however, person fit statistics were used to test the hypothesis that an item score pattern is not in agreement with the underlying test model. With some exceptions (Snijders, 2001; Glas & Meijer, 2003; Glas & Dagohoy, 2007), most detection methods do not take into account the uncertainty about the item and person parameters of the IRT model. In this study, we used the Bayesian based method of PPCs and the frequentist based methods of LM tests and statistics with Snijders' corrections to take into account this uncertainty. Although Bayesian methods are statistically superior to other simulation methods (such as the parametric and non-parametric bootstrap), a drawback is that they are relatively complex and computational intensive and have conservative significance probabilities (Sinharay & Stern, 2003). In this article, the various approached were compared.

From the results, it can be concluded that for a test 60 items the Type I error is well under control at a nominal .05 for most statistics studied in both approaches. In general, the Type I error is slightly conservative, that is lower, in the Bayesian procedure. Similar results were found in an earlier study of Glas and Meijer (2003). The nominal significance level for the LM tests and frequentist PFS with Snijders' correction were quite close to the 5% nominal significance level. This is in accordance with findings in earlier studies (e.g., van Krimpen-Stoop & Meijer, 1999). Detection rates differed for different statistics and different types of model violations simulated. In general, it can be concluded that the detection rates for guessing were higher than detection rates for item disclosure. The fit statistics have higher power in the 2PNO model than in the 3PNO model. This could be the fact that more item parameters need to be estimated for the 3PNO model, which leads to a loss of power. The frequentist procedures had a higher power than the Bayesian procedures. Aggregated over all conditions, the ξ_2-test had the highest power, while (uncorrected) *UB*-test had the lowest detection rates. The detection rates decreased when the number of items affected by guessing increased.

Glas and Dagohoy (2007) have shown that the Lagrange multiplier statistic can take both the effects of estimation of the item parameters and the estimation of the person parameters into account. However, the estimation of the item parameters usually has little impact. The Lagrange multiplier statistic has an asymptotic χ^2-distribution. Finally, a remark can be made about the relation of the LM test to other tests of person fit. An essential feature of the test is that a model violation is translated into an explicit alternative model by introducing extra parameters that represent the model violation. The test then amounts to the evaluation whether the additional parameters are equal to zero. This distinguishes the LM approach from the use of more general tests such as the test based on the likelihood statistic l_z by Drasgow et al. (1985) and the test based on the Pearson-type *W*-statistic by Wright and Stone (1979). These tests have an unspecified general alternative, so they have a more global nature. The LM approach allows for targeting specific model violations and can also be used to target a number of other model violations rather than the one studied here in detail.

Appendix

4.A The Null Distributions of PFS

The person fit statistics $W_n(\theta)$ that is linear in the item responses following Snijders defined as

$$W_n(\theta) \ = \ (X_i - P_i)\omega_i(\theta) \tag{4.17}$$

where $\omega_i(\theta)$ is the weight function. To obtain the approximation of the normal distribution mean defined as

$$E\,(W_n(\theta)) \approx -c_n(\theta)r_0(\theta) \tag{4.18}$$

where $c_i(\theta) = \dfrac{\sum_{i=1}^{n} P_i^{/}\omega_i(\theta)}{\sum_{i=1}^{n} P_i^{/} r_i(\theta)}$ and $r_i(\theta) = \dfrac{P_i^{/}}{P_i(1-P_i)}$. The variance can be defined as

$$\operatorname{var}(W_n(\theta)) \approx n\tau_n^{\,2}(\theta) \tag{4.19}$$

The $\tau_n^{\,2}(\theta)$ and $w_i(\theta)^{/}$ defined as

$$\tau_n^{\,2}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ w_i^{\,2}(\theta)^{/} P_i(1-P_i) \tag{4.20}$$

$$w_i(\theta)^{/} = \omega_i(\theta) - c_n(\theta)r_i(\theta) \tag{4.21}$$

where $P_i^{/}$ is the first order derivate of P_i with respect to θ. For the ML estimator of the ability $r_0(\theta) = 0$ and

$$r_i(\theta) = a_i \qquad \text{(2PLM)} \tag{4.22}$$

$$r_i(\theta) = \frac{a_i\phi(a_i\theta - b_i)}{P_i(1-P_i)} \qquad \text{(2PNO)} \qquad (4.23)$$

$$r_i(\theta) = \frac{\dfrac{a_i(1-P_i)(P_i-c_i)}{(1-c_i)}}{P_i(1-P_i)} \qquad \text{(3PLM)} \qquad (4.24)$$

$$r_i(\theta) = \frac{(1-c_i)a_i\phi(a_i\theta - b_i)}{P_i(1-P_i)} \qquad \text{(3PNO)} \qquad (4.25)$$

$$P_i^/ = a_iP_i(1-P_i) \qquad \text{(2PLM)} \qquad (4.26)$$

$$P_i^/ = a_i\Phi(a_i\theta - b_i) \qquad \text{(2PNO)} \qquad (4.27)$$

$$P_i^/ = \frac{a_i(1-P_i)(P_i-c_i)}{(1-c_i)} \qquad \text{(3PLM)} \qquad (4.28)$$

$$P_i^/ = (1-c_i)a_i\Phi(a_i\theta-b_i) \qquad \text{(3PNO)} \qquad (4.29)$$

The above defined expressions are common for all PFS and the below ones are PFS specific.

l_z Statistic

The weight function for l_z is

$$\omega_i(\theta) = (a_i\theta - b_i) \qquad \text{(2PLM)} \qquad (4.30)$$

$$\omega_i(\theta) = \log[\frac{\phi(a_i\theta - b_i)}{1-\phi(a_i\theta - b_i)}] \qquad \text{(2PNO)} \qquad (4.31)$$

W Statistic

Weight function for W is

$$\omega_i(\theta) = \frac{1-2P_i}{\sum_{i=1}^{n} P_i[1-P_i]} \tag{4.32}$$

ξ_1 Statistic

Weight function for ξ_1 is

$$\omega_i(\theta) = (n_i - \bar{n}) \tag{4.33}$$

ξ_2 Statistic

The weight function for ξ_2 is

$$\omega_i(\theta) = (P_i - R/k) \tag{4.34}$$

The weight functions for W, ξ_1 and ξ_2 statistics are common for both logistic and ogive models.

5

A Comparison of Top-down and Bottom-up approaches in the Identification of Differential Item Functioning using Confirmatory Factor Analysis

Abstract: Measurement invariance (MI) can be thought of as invariance of measures of the same attribute under different conditions. Confirmatory factor analytic (CFA) procedures can be used to provide evidence of measurement invariance. When conducting DIF studies using the CFA, there is variation in the way nested models are constructed. In this paper, Constrained baseline (top-down) and Free baseline (bottom-up) approaches for detection of test structure differences across groups under several conditions were examined. More specifically, sample size, test length, underlying response models, effect size, percentage of DIF items, and kind of DIF were evaluated. Power and Type I error were examined to evaluate the accuracy of detecting a lack of measurement invariance for both approaches. Implications of the results are discussed and recommendations for best practice are provided.

This chapter has been submitted for publication as: M.N. Khalid, Bernard P. Veldkamp and Cees A. W. Glas, A Comparison of Top-down and Bottom-up approaches in the Identification of Differential Item Functioning using Confirmatory Factor Analysis.

5.1 Introduction

Measurement can be defined as the systematic assignment of numbers on variables to represent characteristics of persons, objects, or events. In the behavioral sciences, measurement processes typically are aimed at describing characteristics of individuals or groups that are of some substantive interest to the researcher. But in order to compare individuals or groups based on their scores, the scores have to reflect true differences in these characteristics. Therefore, there is currently a great deal of interest in the assessment of measurement invariance, in the field of psychometrics. Measurement invariance, also referred to as measurement equivalence (ME; Vandenberg, 2002) in the literature, can be thought of as invariance of measures of the same attribute under different conditions (Horn & McArdle, 1992). It is considered to be a prerequisite for meaningful group comparisons (e.g., Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Vandenberg & Lance, 2000).

A test or a subscale is said to have measurement invariance across groups or populations if persons with identical scores on the underlying/latent construct have the same expected raw score or true score at the item level, the subscale total score level, or both (Drasgow & Kanfer, 1985). When measurement invariance is present, the relationship between the latent variable and the observed variable remains invariant across populations. In this case, the observed mean difference may be viewed as reflecting only the true difference between the populations. However, in selection contexts, practitioners may be concerned that differences in test scores across groups are caused by the instrument rather than by differences in proficiency (Drasgow, 1987; Stark, Chernyshenko, & Drasgow, 2004). To solve these issues a methodology is needed that can distinguish a lack of measurement invariance (i.e., differential item functioning; DIF) from impact.

Currently, there are two popular methods for establishing measurement invariance. One method is based on structural equation modeling (or, more specifically, confirmatory factor analysis (CFA)), and the other is based on item response theory (IRT). These approaches are often viewed as distinctly different alternatives. Each alternative has its own terminology and approach for examining the relationships among items and scales, and have evolved in relative isolation. Mc-Donald (1999) and

Raju et al., (2002) provided comprehensive reviews of the methodological similarities and differences among CFA and IRT.

When conducting DIF studies using the CFA and IRT, there is a variation in the way models are constructed. Both bottom-up and top-down approaches are being applied. In the top-down approach, a baseline model is formed by constraining the parameters for all items to be equal across groups and a series of augmented models is formed by freeing the parameters for the studied item(s), one at a time, and examining the changes in G^2 (e.g., Thissen, 1991; Bolt, 2002). In the first step of a top-down approach, no DIF is assumed for all items. Next, models are estimated that allow for one additional DIF item at a time. When the fit of a subsequent model does not improve significantly, the approach stops. The approach is also referred to as the Constrained baseline approach. This term will be used through out this paper. When this approach is used, it is not necessary to specify a referent item for identifying the metric, because, in each comparison, all items except the studied item are constrained. However, it is necessary to anchor the metric by choosing a reference group whose latent mean is set to zero for parameter estimation.

Constrained baseline (top-down) approaches, like for example the traditional likelihood ratio (LR) tests, are being applied for several reasons. First, it has been shown that Constrained baseline approaches work fairly well in simulation studies where the number of DIF items is not too large relative to the number of items examined (see Bolt, 2002; Cohen, Kim, & Wollack, 1996; Kim & Cohen, 1998). Besides, it stems from a general belief that it is better to establish a common metric by using more than one item, as is often done in CFA. Using just one referent could reduce the accuracy of the implicit linking that is required for parameter estimation, thus leading to inaccurate DIF detection. Linking based on one referent could also considerably be problematic if the referent's discrimination and location parameters differ appreciably across groups. Consequently, additional steps might be needed, in practice, to ensure that the choice of the referent is appropriate. Generally speaking, IRT researchers prefer to implement a Constrained baseline approach for detecting DIF.

In the Free baseline (bottom-up) approach, a baseline model is typically one in which all parameters except the referent are free to vary and an item is studied by additionally constraining its parameters to be equal across groups. In the first model, DIF is assumed for all items except the referent. In each subsequent step, DIF is excluded for one additional item. The approach stops when a subsequent model does not show a significantly better fit.

The Free baseline approach is preferred due to following argument. According to statistical theory (Maydeu-Olivares & Cai, 2006) for the difference between a baseline model and constrained models to follow a central chi-square distribution under the null hypothesis, the baseline model has to fit the data. If the baseline model contains a number of DIF items, then it might not fit adequately and DIF detection could be adversely affected. Thus, from a statistical standpoint, the approach of comparing nested models in Free baseline approach is theoretically appropriate, whereas the traditional Constrained baseline approach is not. In most CFA based methods for DIF detection, a Free baseline approach is implemented.

To summarize, between CFA-based and the IRT-based approaches for DIF detection, there are common themes but also discrepancies concerning the details of implementation. In both cases, the comparison models usually involve fixing or freeing a studied item or a subset of items. In general CFA based methods implement a Free baseline approach where as most IRT based DIF detection methods implement a Constrained baseline approach. There have been some notable exceptions, though. For example, Reise et al. (1993) used the Free baseline approach to conduct an IRT DIF study of mood ratings collected in Minnesota and China, and compared the results with those using CFA. Besides, Chan (2000) used the Constrained baseline approach for CFA analysis of cognitive styles across gender and occupational groups in Singapore.

These studies are interesting and important because they illustrate alternative approaches for group comparisons. However, they did not address the question of accuracy concerning DIF detection. Moreover, given the variation in methods that has been discussed in the literature, it is premature to advocate strongly a particular sequence of steps, because there have been no direct comparisons of the Free and Constrained baseline strategies in either domain, let alone across domains, in the context of simulation studies, where the truth about DIF items is known.

Given the current variation in the way models are constructed, the purpose of this research is to make a comprehensive comparative simulation study and to explore the factors that have an impact on their performance. The present study will provide a comparison of Free baseline and Constrained baseline approaches for DIF detection to shed some light on unexplored issues. Both approaches will be implemented in CFA based method called Mean and Covariance Structure (MACS; Sörbom, 1974). This method was selected because the basic process for identifying DIF items in MACS is similar to the IRT based LR method (see Thissen, Steinberg, & Wainer, 1993). It involves the comparison of a baseline model with a series of augmented or constrained models. In addition, it presumably capable of detecting DIF due to differences in item discrimination (loadings) and location (item means or intercepts). The rest of chapter organized as follows. First, a concise overview of CFA frame work and CFA based DIF approach will be described. Second, the outline of simulation study is described. Next, the results of study are tabulated. Then both approaches are applied in the context of an empirical example. Finally, a discussion and recommendations for further study are provided.

5.2 CFA Based DIF Procedure

The following section provides an overview of the confirmatory factor analysis (CFA) and differential item functioning (DIF) for examining measurement invariance. Only the single underlying factor CFA model will be presented.

5.2.1 Confirmatory Factor Analysis

Following Jöreskog and Sörbom (1996), the linear confirmatory factor analysis (CFA) model may be represented as

$$x = \lambda_x \xi + \delta, \tag{5.1}$$

where *x* represents a vector of observed or measured variables, ξ is a vector of latent variables or underlying factors, λ_x is a regression coefficient or factor loading matrix that relates factors to observed variables, and δ is a vector of measurement errors/residuals in *x*. Equation 5.1 is commonly referred to as the measurement model for the exogenous variables in SEM. Typically, the vector *x* represents items that serve as the indicator variables (i.e., the observed variables generated by their underlying latent constructs); different items serve as indicator variables for different latent constructs (ξ) in a CFA. As a consequence, the regression paths or lambdas linking the items to their underlying latent constructs are of primary interest. If the factor loadings are equal across groups and measurement errors are assumed to have a mean of zero, then the expected value of x will be equal across groups. This implies that persons of equal ability (i.e., same factor score) will have the same expected raw score on the item (Raju, Laffitte, & Byrne, 2002).

5.2.2 DIF Procedure

In CFA literature, the term measurement invariance is broadly used, in the sense that invariance can exist to different degrees or in different forms (see e.g., Vandenberg & Lance, 2000). The weakest form of invariance is known as configural invariance (Horn & McArdle, 1992), which implies that the same number of factors and similar patterns of loadings exist across groups. The next two forms of invariance are known as metric invariance and scalar invariance. Metric invariance implies equality of factor loadings and scalar invariance implies equality of intercepts for the regressions of items on the latent variables they represent (Steenkamp & Baumgartner, 1998). One can also examine invariance of factor correlations or covariances, uniquenesses, and latent means, but generally there is no clear prescription about the order or need for performing these tests. Obviously, configural invariance must be established before examining metric or scalar invariance. However, metric invariance is cited as a prerequisite for meaningful examination of scalar invariance in many expositions (e.g., Vandenberg & Lance, 2000).
In CFA, testing for equivalence of loadings and intercepts across groups is relatively straight forward. First, one specifies a baseline model, where at least those items are fixed that are needed for identification. The most common way to set the metric is to select a reference item (i.e., a referent) whose loading is set equal to 1 in both groups. In addition, to

examine intercepts in a MACS analysis, one must also constrain the intercepts for the referent to be equal across groups and the latent mean for one group to be zero (for alternative approaches to identification, see Byrne, 1998; Jöreskog & Sörbom, 1996; Reise et al., 1993). The next step is to specify a series of compact (constrained) models where, in each case, the respective loadings and/or intercepts for one or more items are constrained to be equal across groups. The baseline and constrained models are then estimated in succession to obtain a chi-square goodness-of-fit statistic for each. Because each constrained model is nested within the baseline model, one can then compare the change in chi-square with respect to the baseline with a critical value having degrees of freedom equal to the difference in degrees of freedom for the respective models. For each comparison, a statistically significant result is viewed as evidence that the hypothesis of equivalence of the constrained parameters is untenable.

5.3 Study Design

The current study seeks to explore and compare the efficacy of Constrained baseline and Free baseline models with detecting DIF using a using unidimensional and multidimensional scale under a variety of conditions. The following variables were manipulated in the study. First, the sample size was varied. Earlier studies have found the effects of sample size (Glas, 1999; Meade & Lautenschlager, 2004). Sample sizes of 400 (small) and 1000 (large) were chosen as they frequently occurred in the educational and psychological measurement. Earlier studies found that increase in number of items also have an effect on power and Type I error rates of DIF detection methods (Reise, 1990; Finch, 2005; Glas & Dagohoy, 2007). Therefore the test length varied from 10 (small), 20 (average), and 40 (long test).

Besides, several response models were applied. The 2-parameter logistic model (2PL), the 3-parameter logistic model (3PL), the Graded response model (GRM), and a Multidimensional model were chosen as they are the most commonly applied IRT models, and estimation procedures for these models are well defined. Multidimensional models were also chosen because they have not been studied before with CFA procedures. The discrimination parameters were drawn from a lognormal distribution, and difficulty parameters were drawn from a standard normal distribution. For the 3PL, guessing parameters were fixed at 0.20. For GRM, Sample sizes of N = 400, and N = 1000 were crossed with test lengths of K = 10, K = 20, and K = 40. For the test length K = 10, the item parameters were equal to $\beta_i = -2.00 + 0.20(i - 1), i = 1, \ldots, 10$. For the test lengths K = 20 and K = 40 these values were repeated two and three times, respectively.

The ability distributions of both the reference and the focal group follow a standard normal distribution $N(0, 1)$. For the multidimensional (between & within) models, the mean was set equal to zero and the correlation between dimensions was 0.40 and 0.80 respectively. Both uniform and non-uniform DIF were simulated. The amount of DIF varied. Effect sizes were set at 0.5 or 1.0. Finally, the percentage of DIF items varied, either 10% or 20% of the items were simulated to have DIF.

The procedure that implements this study design can be described as follows.

1) First, a reference group is designated, whose latent mean is set to zero. The latent mean of the other group is free to vary.

2) Second, a baseline model is specified. For the Free baseline approach, item 1 is chosen to be the referent item. Only the parameters of this item (marker item) are constrained across groups. For the Constrained baseline approach, the parameters for all items are constrained across groups. Constraining items in MACS implies that the loadings are constrained to be 1 and the intercepts are constrained to be equal across groups. For a multidimensional model, the first item on each dimension is selected as a referent in the Free baseline approach.

3) Third, to detect DIF items, a series of constrained models is formed, where one item at a time has its discrimination (loading) and location (intercept) parameters simultaneously constrained to be equal/freed across groups.

4) Finally, DIF items are identified by comparing the respective changes in chi-square using a critical value. The critical chi-squares used for MACS was 5.99 with 2 degrees of freedom for

> dichotomous, polytomous, and (between) multidimensional conditions because the number of parameters that constrained/free for nested models were equal. The critical chi-squares 7.81 with 3 degrees of freedom were used for within multidimensional models due to loading of each indicator on each dimension.

MACS analyses were conducted using the LISREL 8 computer program (Jöreskog and Sörbom, 1996). For each replication, baseline and constrained or augmented models, constructed as described previously, were run in succession; the chi-square goodness-of-fit statistics were extracted from the LISREL output files using a FORTRAN program. Chi-square difference statistics for the nested model comparisons were evaluated with respect to critical *p*-values. More specifically, when the observed chi-square difference was greater than the corresponding critical chi-square value, the item was flagged as having DIF.

5.4 Results

Tables 5.1-5.10 present power and Type I error rate for the simulation study. Power represents the proportion of DIF items correctly identified as having DIF across the 100 replications in each condition. Type I error represents the proportion of times an item having no DIF was incorrectly flagged. Because of the large number of simulated conditions, we have presented the interpretation of results in according to the generating IRT models.

5.4.1 2PL

Tables 5.1 and 5.2 show the results for the 2PL model based on uniform and non-uniform DIF conditions. The first column labeled K denotes test length; the second column labeled δ denotes effect size and third column labeled N denotes sample size. The next columns show the power and Type I error rate as a function of test length, effect size and sample size over 100 replications at the 5% level of significance. Labels 10% and 20% show the presence of DIF items as percentage of test length.

Table 5.1. The Power under the 2PL model using Constrained baseline and Free baseline approach.

			Cons.		Free		Cons.		Free	
			Number of items with uniform DIF				Number of items with non-uniform DIF			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.79	0.67	0.80	0.69	0.27	0.21	0.31	0.27
		1000	0.99	0.97	0.99	0.98	0.62	0.44	0.64	0.43
	1.00	400	1.00	1.00	1.00	1.00	0.64	0.43	0.65	0.42
		1000	1.00	1.00	1.00	1.00	0.94	0.84	0.95	0.86
20	0.5	400	0.83	0.72	0.87	0.73	0.27	0.26	0.29	0.28
		1000	1.00	1.00	1.00	1.00	0.70	0.65	0.79	0.69
	1.00	400	1.00	1.00	1.00	1.00	0.74	0.67	0.77	0.67
		1000	1.00	1.00	1.00	1.00	1.00	0.98	1.00	0.98
40	0.5	400	0.77	0.74	0.82	0.85	0.37	0.32	0.40	0.34
		1000	0.98	0.99	0.99	0.98	0.77	0.69	0.80	0.68
	1.00	400	1.00	1.00	1.00	1.00	0.88	0.78	0.90	0.79
		1000	1.00	1.00	1.00	1.00	1.00	0.99	1.00	1.00

Table 5.2. The Type I error rate under the 2PL model using Constrained baseline and Free baseline approach.

			Cons.		Free		Cons.		Free	
			Number of items with uniform DIF				Number of items with non-uniform DIF			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.14	0.14	0.07	0.08	0.10	0.14	0.08	0.08
		1000	0.15	0.23	0.07	0.08	0.12	0.16	0.09	0.10
	1.00	400	0.19	0.28	0.08	0.09	0.13	0.15	0.07	0.09
		1000	0.21	0.43	0.07	0.08	0.14	0.18	0.08	0.10
20	0.5	400	0.09	0.09	0.06	0.07	0.08	0.10	0.08	0.09
		1000	0.09	0.11	0.08	0.08	0.11	0.13	0.08	0.09
	1.00	400	0.10	0.12	0.07	0.08	0.10	0.12	0.07	0.09
		1000	0.11	0.16	0.07	0.08	0.14	0.15	0.07	0.08
40	0.5	400	0.07	0.06	0.06	0.05	0.06	0.06	0.05	0.06
		1000	0.08	0.09	0.06	0.06	0.09	0.06	0.06	0.06
	1.00	400	0.08	0.09	0.06	0.07	0.07	0.08	0.06	0.07
		1000	0.08	0.09	0.06	0.07	0.07	0.10	0.06	0.07

For uniform DIF, the detection rates in most instances were close or equal to 1.0 and much smaller for similar conditions in non-uniform DIF. Note that detection rates have main effects of test length, effect size and sample size. The increase in test length and sample size inflates the power because of the more reliable estimates of parameters. The large effect size makes a substantial difference in response probabilities among the groups that inflates the power. Overall, the power to detect DIF was better when item means (thresholds) were different for the two groups (uniform DIF) rather than just the loadings (non-uniform DIF). This might be explained as follows. Increasing the thresholds made the DIF items more difficult for focal group members, essentially shifting their item response functions toward the right (indicating higher trait levels). Thus, focal group members of equal standing on the latent attribute had a lower probability of endorsement than reference group members, regardless of trait level. The increase in the thresholds can be viewed as a nuisance dimension. On the other hand, reducing the loadings made the DIF items less discriminating for focal group members. When item discrimination varies across groups, the item response functions that relate probability of endorsement to trait level cross, meaning that the reference group is favored at some trait levels and the focal group at others. But the underlying model is still followed and due to crossing of favourability the net effect is moderate. The detection rates were comparable among both Free baseline and Constrained baseline approaches and slightly in favour of the Free baseline approach, for instance when test length was 10.

Type I error rates were inflated for Constrained baseline conditions irrespective of the kind of DIF. For example, when uniform DIF was simulated, the Constrained baseline strategy yielded Type I error rates ranging from .14 to .43, for small tests. These error rates were slightly lower when non-uniform DIF was simulated. The error rates have main effects of percentage of DIF items, effect size, and test length. Overall, it appears that the presence of items having large DIF in the baseline model substantially increased the Type I error but did not effect the power. Note also that when items having small DIF were present in the Constrained baseline models, error rates were still inflated but to a lesser degree.

In contrast to the highly inflated Type I error rates in the DIF conditions for the Constrained baseline approach, the Free baseline strategy showed excellent results regardless of sample size, percentage of DIF items and effect size. For example, when effect size is large, Type I error rates for the Free baseline conditions were between .06 and .08 in both uniform and non-uniform DIF conditions. These results are similar to the earlier findings of Stark, Chernyshenko, and Drasgow (2006). Therefore, having a baseline model that fits the data, (i.e., did not include DIF items) was critical for accurate detection. This point elaborates as follows. When the least restrictive model holds, relative fit assessment can be safely performed for much larger models than those whose absolute fit can be tested with the χ^2 and G^2 statistics. However, when the least restrictive model being compared is misspecified, statistical inferences based on G^2 (dif) can be misleading. This is because in this case a chi-square distribution is no longer the appropriate large sample reference distribution for this statistic (Maydeu-Olivares & Cai, 2006). Recently, Yuan and Bentler (2004) showed that when a likelihood ratio statistic is used to compare two nested models but the least restrictive model is misspecified inflated Type I errors are obtained. Their results thus concur with those presented here.

5.4.2 3PL

Tables 5.3 and 5.4 show the results for the power and Type 1 error rates when data was simulated using the 3PL. The detection of DIF decreased dramatically in most instances as compared to similar conditions observed under the 2PL. The decrease in the power is more evident when the DIF was simulated using non-uniform rather uniform DIF. To some extent detection rates have main effects of test length, sample size and effect size; specifically for non-uniform DIF. Detection rates were higher for uniform DIF. In general, it can be seen that the detection rates were comparable among both approaches and slightly in favour of the Free baseline approach for both uniform and non-uniform DIF.

Table 5.3. The Power under the 3PL model using Constrained baseline and Free baseline approach.

			Cons.		Free		Cons.		Free	
			Number of items with uniform DIF				Number of items with non-uniform DIF			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.67	0.40	0.70	0.56	0.11	0.10	0.11	0.17
		1000	0.93	0.83	0.96	0.86	0.17	0.15	0.18	0.16
	1.00	400	1.00	0.97	1.00	0.99	0.19	0.17	0.29	0.20
		1000	1.00	0.99	1.00	1.00	0.41	0.38	0.42	0.40
20	0.5	400	0.63	0.51	0.64	0.54	0.19	0.17	0.21	0.19
		1000	0.96	0.95	0.98	0.96	0.23	0.21	0.23	0.21
	1.00	400	1.00	0.99	1.00	1.00	0.26	0.17	0.28	0.27
		1000	1.00	1.00	1.00	1.00	0.56	0.51	0.60	0.54
40	0.5	400	0.60	0.55	0.64	0.58	0.20	0.18	0.25	0.24
		1000	0.94	0.92	0.97	0.91	0.28	0.23	0.38	0.33
	1.00	400	1.00	0.99	1.00	1.00	0.30	0.28	0.32	0.31
		1000	1.00	1.00	1.00	1.00	0.46	0.47	0.59	0.56

Table 5.4. The Type I error rate under the 3PL model using Constrained baseline and Free baseline approach.

			Cons.		Free		Cons.		Free	
			Number of items with uniform DIF				Number of items with non-uniform DIF			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.11	0.10	0.07	0.08	0.07	0.08	0.04	0.05
		1000	0.12	0.18	0.08	0.09	0.06	0.06	0.07	0.07
	1.00	400	0.13	0.25	0.09	0.11	0.07	0.08	0.07	0.07
		1000	0.14	0.28	0.10	0.14	0.07	0.08	0.08	0.08
20	0.5	400	0.07	0.08	0.04	0.05	0.04	0.06	0.05	0.05
		1000	0.10	0.10	0.07	0.08	0.06	0.05	0.06	0.05
	1.00	400	0.10	0.11	0.07	0.08	0.05	0.06	0.06	0.06
		1000	0.11	0.14	0.09	0.09	0.06	0.06	0.06	0.07
40	0.5	400	0.06	0.07	0.04	0.05	0.04	0.04	0.05	0.05
		1000	0.07	0.08	0.06	0.07	0.05	0.05	0.04	0.05
	1.00	400	0.08	0.09	0.06	0.06	0.06	0.04	0.04	0.05
		1000	0.08	0.09	0.07	0.07	0.05	0.05	0.05	0.05

Type I error rates were inflated for the Constrained baseline approach in case of uniform DIF conditions. When uniform DIF was simulated, the Constrained baseline strategy yielded Type I error rates ranging from .14 to .28. The error rates were close to nominal significance level when non-uniform DIF was simulated. Overall, it appears that the when the underlying model was 3PL, the performance of both approaches appears to decrease. A possible explanation might be the presence of a guessing parameter, since in CFA there is no parameter that accommodates for the guessing behaviour.

5.4.3 GRM

Tables 5.5 and 5.6 show the results when the data was simulated using the Graded Response Model (GRM). It is evident that increasing the number of response categories from two to five improved the accuracy of DIF detection. Detection rates were slightly higher in case of uniform DIF. The increase in detection rates when the underlying model was the GRM instead of the 2PL may be explained as follows. When the number of response options increases, violations of normality and continuity becomes less of an issue and the power of approach improves. The power of both the constrained based approach and the Free baseline approach was comparable, but the Free baseline approach showed much smaller Type I error rates. Note that the error rates remained fairly low for the Free baseline conditions but that they are slightly higher than those observed in the dichotomous conditions.

Table 5.5. The Power under the GRM model using Constrained baseline and Free baseline approach.

			Cons.		Free		Cons.		Free	
			Number of items with uniform DIF				Number of items with non-uniform DIF			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.87	0.85	0.88	0.86	0.67	0.61	0.67	0.62
		1000	0.99	0.98	1.00	0.99	0.93	0.84	0.94	0.85
	1.00	400	1.00	1.00	1.00	1.00	0.84	0.80	0.85	0.82
		1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	0.5	400	0.93	0.87	0.95	0.86	0.70	0.68	0.71	0.69
		1000	1.00	1.00	1.00	1.00	0.96	0.94	0.97	0.95
	1.00	400	1.00	1.00	1.00	1.00	0.88	0.86	0.88	0.87
		1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
40	0.5	400	0.97	0.94	0.98	0.95	0.77	0.72	0.78	0.72
		1000	0.98	0.99	0.99	0.98	0.98	0.96	0.99	0.97
	1.00	400	1.00	1.00	1.00	1.00	0.98	0.96	0.99	0.97
		1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 5.6. The Type I error rate under the GRM model using Constrained baseline and Free baseline approach.

			Cons.		Free		Cons.		Free	
			Number of items with uniform DIF				Number of items with non-uniform DIF			
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.14	0.15	0.07	0.08	0.11	0.13	0.08	0.08
		1000	0.16	0.26	0.07	0.08	0.13	0.15	0.07	0.08
	1.00	400	0.19	0.29	0.07	0.08	0.14	0.17	0.08	0.09
		1000	0.24	0.42	0.07	0.08	0.16	0.20	0.08	0.10
20	0.5	400	0.08	0.09	0.06	0.07	0.09	0.11	0.06	0.07
		1000	0.10	0.11	0.06	0.07	0.09	0.13	0.07	0.08
	1.00	400	0.11	0.14	0.07	0.08	0.10	0.12	0.07	0.08
		1000	0.13	0.17	0.07	0.08	0.16	0.18	0.06	0.07
40	0.5	400	0.06	0.06	0.06	0.05	0.06	0.07	0.05	0.06
		1000	0.05	0.07	0.05	0.06	0.09	0.06	0.05	0.06
	1.00	400	0.07	0.08	0.06	0.07	0.07	0.08	0.06	0.07
		1000	0.08	0.09	0.05	0.07	0.08	0.10	0.06	0.07

5.4.4 Multidimensional (between) Models

Tables 5.7 and 5.8 show the results when data were simulated using the multidimensional (between) model. The power in most instances was close or equal to 1. Effects of effect size and sample size were found. Overall, the power to detect DIF was comparable irrespective of correlation between the dimensions. Type I error rates were inflated for Constrained baseline conditions and remain close to nominal significance level for the Free baseline approach. The detection rates were comparable among both Free baseline and Constrained baseline approaches and slightly in favour of the Free baseline approach for instance when effect size and sample size were small.

Table 5.7. The Power under the multidimensional (between) model with correlation 0.40 and 0.80 using Constrained baseline and Free baseline approach.

			Correlation 0.40				Correlation 0.80			
			Cons.		Free		Cons.		Free	
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.71	0.74	0.73	0.75	0.72	0.77	0.74	0.76
		1000	0.99	0.97	0.99	0.98	1.00	1.00	1.00	0.99
	1.00	400	0.99	1.00	0.99	0.98	1.00	1.00	1.00	0.99
		1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
20	0.5	400	0.76	0.80	0.79	0.80	0.76	0.80	0.77	0.78
		1000	0.99	0.98	0.99	0.97	0.99	1.00	0.94	0.89
	1.00	400	1.00	1.00	1.00	1.00	1.00	1.00	0.97	0.95
		1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
40	0.5	400	0.79	0.76	0.78	0.77	0.75	0.73	0.76	0.72
		1000	0.99	0.99	0.98	0.99	1.00	1.00	0.88	0.99
	1.00	400	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99
		1000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 5.8. The Type I error rate under the multidimensional (between) model with correlation 0.40 and 0.80 using Constrained baseline and Free baseline approach.

			Correlation 0.40				Correlation 0.80			
			Cons.		Free		Cons.		Free	
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.07	0.11	0.05	0.05	0.06	0.11	0.04	0.06
		1000	0.09	0.16	0.05	0.05	0.09	0.17	0.06	0.07
	1.00	400	0.14	0.24	0.05	0.07	0.13	0.22	0.03	0.05
		1000	0.25	0.50	0.05	0.06	0.23	0.46	0.05	0.04
20	0.5	400	0.05	0.06	0.05	0.05	0.05	0.06	0.04	0.05
		1000	0.05	0.07	0.04	0.05	0.05	0.05	0.05	0.04
	1.00	400	0.06	0.07	0.05	0.05	0.06	0.07	0.05	0.04
		1000	0.08	0.13	0.04	0.05	0.08	0.12	0.04	0.05
40	0.5	400	0.04	0.05	0.04	0.04	0.04	0.07	0.04	0.04
		1000	0.05	0.07	0.04	0.04	0.05	0.06	0.04	0.04
	1.00	400	0.06	0.07	0.05	0.04	0.06	0.08	0.05	0.05
		1000	0.08	0.12	0.05	0.05	0.08	0.11	0.04	0.05

5.4.5 Multidimensional (within) Models

The results are shown in tables 5.9 and 5.10 when data were simulated using the multidimensional (within) model. Overall, the power to detect DIF and Type I error rates were lower than those observed for the between multidimensional models. The decrease in power can be explained as follows. There are more parameters that have to be estimate for the within multidimensional models. That results in loss of precision. Type I error rates were inflated for Constrained baseline approach. Detection rates were comparable for both approaches and slightly in favour of Free baseline approach.

Table 5.9. The Power under the multidimensional (within) model with correlation 0.40 and 0.80 using Constrained baseline and Free baseline approach.

			Correlation 0.40				Correlation 0.80			
			Cons.		Free		Cons.		Free	
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.61	0.59	0.62	0.63	0.63	0.62	0.65	0.64
		1000	0.82	0.77	0.80	0.79	0.83	0.79	0.82	0.80
	1.00	400	0.89	0.82	0.90	0.85	0.89	0.83	0.90	0.87
		1000	0.98	0.93	0.95	0.93	0.98	0.93	0.95	0.93
20	0.5	400	0.63	0.60	0.71	0.67	0.65	0.62	0.71	0.67
		1000	0.72	0.75	0.72	0.76	0.72	0.77	0.73	0.76
	1.00	400	0.93	0.95	0.97	0.95	0.93	0.95	0.97	0.95
		1000	1.00	1.00	0.98	0.98	1.00	1.00	0.99	0.98
40	0.5	400	0.62	0.54	0.62	0.56	0.65	0.58	0.64	0.57
		1000	0.95	0.91	0.95	0.90	0.97	0.92	0.95	0.92
	1.00	400	0.97	0.96	0.99	0.96	0.98	0.96	0.99	0.97
		1000	1.00	1.00	1.00	0.99	1.00	1.00	1.00	0.99

Table 5.10. The Type I error rate under the multidimensional (within) model with correlation 0.40 and 0.80 using Constrained baseline and Free baseline approach.

			Correlation 0.40				Correlation 0.80			
			Cons.		Free		Cons.		Free	
K	δ	N	10%	20%	10%	20%	10%	20%	10%	20%
10	0.5	400	0.07	0.11	0.05	0.05	0.06	0.07	0.05	0.06
		1000	0.09	0.16	0.05	0.05	0.10	0.14	0.05	0.06
	1.00	400	0.14	0.24	0.05	0.07	0.15	0.19	0.04	0.08
		1000	0.25	0.50	0.05	0.06	0.16	0.32	0.06	0.07
20	0.5	400	0.05	0.06	0.05	0.05	0.04	0.06	0.05	0.05
		1000	0.05	0.07	0.04	0.05	0.05	0.07	0.06	0.08
	1.00	400	0.06	0.07	0.05	0.05	0.05	0.08	0.06	0.05
		1000	0.08	0.13	0.04	0.05	0.06	0.11	0.06	0.06
40	0.5	400	0.04	0.05	0.04	0.04	0.04	0.06	0.05	0.05
		1000	0.05	0.07	0.04	0.04	0.05	0.07	0.06	0.08
	1.00	400	0.06	0.07	0.05	0.04	0.05	0.08	0.06	0.05
		1000	0.08	0.12	0.05	0.05	0.06	0.09	0.06	0.06

5.5 An Empirical Example

Since most differences in performance between the Free baseline and the Constrained baseline approach in the simulation study were found for small test lengths, both methods were applied to detect DIF in a short length test. The example pertains to the number series scale of a Dutch intelligence test. This subscale consisted of seven dichotomous items. The scale was administered in a job selection context. The respondents were divided into two groups on the basis of gender. The sample consisted of 263 males and 410 females.

The evaluation of DIF was done using MACS. Table 5.11 gives the results for both approaches. The column labeled ‘Prob’ gives the probability value at 5% level of significance; the column labeled ‘Constrained baseline and Free baseline’ denotes how the baseline and nested models were formed. The statistic has a chi-square distribution with two degree of freedom. Item 1 was used as reference item for which the loading was fixed to one and intercept equal to zero; hence no significance probabilities were shown for item 1. Mean of the latent variable was zero in the reference group while it was estimated for the focal group. The analysis has shown that item 4 was identified as a DIF item by both approaches. The general conclusion is that for the studied data set, both approaches were efficient for identification of non-fitting items.

Table 5.11. Evaluation of DIF in both approaches.

	Cons. baseline	Free baseline
Item	Prob	Prob
1	-	-
2	0.118	0.782
3	0.117	0.531
4	0.021	0.011
5	0.110	0.645
6	0.109	0.752
7	0.122	0.679

5.6 Discussion and Conclusions

Measurement invariance has long been a central concern in organizational, cross-cultural, and educational research. However, researchers in these fields seemed to have developed different preferences for the type of methodology used to examine this issue. For many years, cross-cultural and industrial organizational psychologists relied primarily on comparisons of group means and correlations. Eventually these approaches were supplemented or replaced by more complex CFA based methods, such as simultaneous factor analysis of several populations and MACS. However, in educational measurement, psychometricians and applied researchers have generally preferred IRT methods for examining measurement equivalence, because these methods were designed specifically to distinguish DIF from impact, an issue that lies at the heart of the polemic about standardized testing.

In this study, we have compared the accuracy of both approaches to detecting DIF while varying sample size, the number of response categories, amounts and types of DIF, and the number of items. Rather than adhering to conventions for data analysis, many of which have not been examined systematically using simulation studies, we attempted to illuminate the correspondence between the two approaches. These strategies, are the Free baseline and Constrained baseline, draw on the strength of the respective methodologies. A fully Free baseline model (with the exception of a referent), which, provided it fits the data, is statistically appropriate as the basis for subsequent nested model comparisons where one item at a time is constrained to be equal across groups (Maydeu-Olivares & Cai, 2006). From IRT, it draws on the ideas of simultaneous comparisons of item parameters (discrimination–loadings and locations–intercepts) and strict *p*-values for flagging DIF items. Here, we examined the Power and Type I error rate of these strategies for DIF detection methods and compared the results with those of an alternative approach involving fully constrained-baseline and augmented models. In the present study, CFA-based mean and covariance structures method (MACS; Sörbom, 1974) was explored using Free baseline and Constrained baseline models. This method was selected because the basic process for identifying DIF items is similar to the IRT based LR method (see Thissen, Steinberg, & Wainer, 1993). It involves the comparison of a baseline model with a series of augmented or constrained models. In addition, it presumably capable of detecting DIF due to differences in item discrimination (loadings) and location (item means or intercepts).

Our results indicated that, with respect to the power of the Free baseline and Constrained baseline approaches, they are similar in their DIF detection efficacy. The Constrained baseline approach exhibited a higher Type I error rate, especially when DIF was simulated on item thresholds. Especially for small tests, Type I error rate of the Constrained baseline approach was much higher than those of the Free baseline approaches. In contrast, the statistically correct Free baseline strategy worked well in all conditions. Power to detect DIF was high, while Type I errors were near the nominal level (.05). Moreover, main effects like sample size, effect size, number of items were evident for instance when DIF was simulated on item loadings. Both strategies showed lower detection rates, when DIF was simulated for 3PL. A possible explanation might be the failure to accommodate guessing behavior. As was expected, both approaches performed better with polytomous data than with dichotomous data. In fact, with polytomous data involving five response options, they showed higher power and only slightly higher error rates. For multidimensional models, the error rates were slightly more controlled due to the constraining of more than one referent item. Another finding of this study is that the between multidimensional models have higher accuracy of DIF detection compared to the within multidimensional models. A possible explanation might be that more parameters have to be estimated which results in loss of precision.

Based on the extensive simulation study, one might conclude in favour of the Free baseline approach. To evaluate the differences in practice both approaches for compared for a small test length, since largest differences between the methods found for small test in the simulation study. For this specific intelligence test, no difference in the performance was found.

Although this study involved variations of experimental conditions, it still has several limitations and implications for future research. One important question is how to find an unbiased referent to implement the free-baseline strategy in practice. This seems to be of critical importance, because having bias in the linking subtest severely inflated the Type I error rates (see the results for our DIF conditions with the constrained-baseline strategy). Strategies to identify potentially DIF free referent items have been discussed in several articles (e.g., Candell & Drasgow, 1988; Cheung & Rensvold, 1999; Stark et al., 2006; Thissen et al., 1988). Another topic of interest can be the number of more than one referent items and their impact on the performance in terms of power and Type I error rates.

Summary

Item response theory (IRT) is a collection of statistical models that used for development, evaluation, and scoring of instruments. IRT models describe, in probabilistic terms, the relationship between a person's response to a test item and his or her standing on the construct being measured by the test. These measured constructs include any latent (i.e., unobservable) variable, such as depression, achievement, or attitude, that requires multiple test items to estimate a person's level on the construct. The properties of these models offer many well-known advantages in testing applications. However, the extent to which these properties are attained is dependent on the degree to which the IRT model itself is appropriate. This thesis is concerned with evaluation of model fit from two perspectives: the items and the respondents. In the first case, for every item, residuals and item fit statistics are computed to assess whether the item violates the model. In the second case, residuals and person fit statistics are computed for every person to assess whether the responses to the items follow the model.

In Chapter 2, a method for testing model fit due to differential item functioning (DIF) was proposed in the framework of marginal maximum likelihood (MML) estimation. The fit of the model is evaluated using the Lagrange multiplier tests. The tests are based on residuals, that is, differences between observed and expected mean scores, that support an appraisal of the seriousness of the model violation. In practice, more than one DIF item may be present, and therefore, step-wise procedures are used where DIF items are identified one at a time. In this process of test purification, items with DIF are identified using statistical tests and DIF is modeled using group-specific item parameters. The two major issues addressed in this study were the following. The first problem addressed is that the dependency of these statistics might cause problems in the presence of relatively large number DIF items. However, simulation studies are presented that show that the power and Type I error rate of a step wise procedure where DIF items are identified one at a time are good. The second problem pertains to the importance of DIF, i.e., the effect size, and related problem of defining a stopping rule for the searching procedure. Simulations are presented that show that the importance of DIF and the stopping rule can be based on the estimate of the difference between the means of the ability distributions of the

studied groups of respondents. The searching procedure is stopped when the change in this effect size becomes negligible.

In Chapter 3, statistics were developed and evaluated that are sensitive to specific unidimensional IRT model assumptions. An essential feature of these statistics is that they are targeted at item fit and based on information that is aggregated over persons. Examples of assumptions that evaluated using fit statistics were subpopulation invariance (DIF), the form of the item response function, and local stochastic independence. To date the most of the goodness-of-fit tests that have been proposed are often poorly rooted in statistical theory. To address this issue, LM statistics that proposed by Glas (1998, 1999) were used. LM statistics were computed using the estimates of a null-model (the IRT model of interest) and gauge the effects of adding parameters that represent violations of the null-model. The obvious advantage of LM test is that many model violations can be evaluated using estimates of the null-model only. However, MML estimation method is complicated (in terms of computation) for multilevel and multidimensional psychometric models (Fox & Glas, 2001; Béguin & Glas, 2001) due to complex dependency structures of models and require the evaluation of multiple integrals needed to solve the estimation equations for parameters. An alternative to the MML framework with LM tests is the Bayesian approach with posterior predictive checks (PPCs) for the assessment of model violations in unidimensioal IRT models. The results of the both frameworks for fit statistics were compared using a number of combinations of test length, sample size, effect size and percentage of misfit items. The overall conclusion was that the LM statistics which were based on frequentistic framework have an edge over the PPCs.

In Chapter 4, person fit measures were proposed to assess the consistency of an individuals' response pattern with an IRT measurement model. To classify an item score pattern as a misfit, a distribution of the statistic under the null hypothesis of fitting response behavior, the null distribution, is needed. In practice, the ability parameter is unknown and has to be estimated. This results in a conservative classification of score patterns as misfitting (Nering, 1995, 1997; Molenaar & Hoijtink, 1990). In this study we have compared the frequentist and Bayesian frameworks that into account effect of estimation of θ. In the first case, Snijders (2001) correction procedure was employed to derive an asymptotic sampling distribution for a family of person fit statistics that are linear in item responses. An alternative to Snijders procedure, LM statistics which are based on residuals, that is, differences between observed and expected mean scores were also evaluated. In the second case, posterior predictive checks (Glas & Meijer, 2003) which are computed in a Markov chain Monte Carlo (MCMC) framework were employed. The two important model violations, Guessing and Item disclosure, were studied using a number of fit statistics in both frameworks. The simulation studies for the Type I error rate and power were presented. In general, the detection rates for aberrant examinees were higher for LM and Snijders' procedure while Type I error rates were conservative for PPCs.

In all these previous chapters, the evaluation of model fit have been done in the frequentist estimation method using MML and Bayesian estimation method using MCMC. As an alternative to IRT models, a linear method based on confirmatory factor analysis (CFA) is considered in Chapter 5 for the assessment of measurement equivalence (DIF). Recently, a few studies (Raju, Laffitte, & Byrne, 2000; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006) have offered a review and a comparison of results from CFA and IRT. Currently there is a variation in the way nested models are constructed that involve fixing or freeing a studied item or a subset of items. In Bottom-up, a baseline model is typically one in which all parameters except the referent are free to vary and an item is studied by additionally constraining its parameters to be equal across groups. In Top-down, a baseline model is formed by constraining the parameters for all items to be equal across groups and a series of augmented models is formed by freeing the parameters for the studied item(s), one at a time, and examining the changes in G^2 (e.g., Thissen, 1991; Bolt, 2002). A comprehensive comparative simulation study was conducted using the 2PL, 3PL, GRM and a multidimensional underlying models with combinations of various test length, sample size, effect size and percentage of misfit items. The results of the study show that detection rates were comparable among all underlying models and approaches except for 3PL. The Type I error rates were inflated for the Top-down approach while close to nominal significance levels for the Bottom-up approach.

References

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

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

Albert, J.H., & Ghosh, M. (2000). Item response modeling. In D. Dey, S. Ghosh & Mallick (eds.), *Generalized Linear Models: A Bayesian Perspective* (pp. 173-193). Marcel-Dekker, New York.

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. *Psychometrika, 38,* 123–140.

Andersen, E.B. (1980). *Discrete statistical models with social science applications*. Amsterdam, North Holland.

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

Bayarri, M, J., & Berger, J.O. (2000). P-values for composite null models. *Journal of the American Statistical Association, 95*, 1127–1142.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some fit analysis of multidimensional IRT models. *Psychometrika, 66,* 471-488.

Berkhof, J., Van Mechelen, I., & Gelman, A. (2002). *Posterior predictive checking using antisymmetric discrepancy functions.* technical report.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinees ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397-424). Reading MA: Addison-Wesley.

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

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

Bock, R. D., & Zimowski, M. F. (1997). Multiple Group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 433-448). New York: Springer Verlag.

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15,* 113–141.

Bolt, D. M., Cohen, A. S., & Wollack, J. A. (2002). Item parameter estimation under conditions of test speededness: Application of a mixture Rasch model with ordinal constraints. *Journal of Educational Measurement, 39*, 331–348.

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

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic concepts, applications, and programming.* Mahwah, NJ: Erlbaum.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage.

Chan, D. (2000). Detection of differential item functioning on the Kirton Adaption–Innovation Inventory using multiple-group mean and covariance structure analyses. *Multivariate Behavioral Research, 35,* 169–199.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differential item functioning test items. *Educational Measurement: Issues and Practice, 17*, 31-44.

Cohen, A. S., Kim, S.-H., & Baker, F. B. (1993). Detection of differential item functioning in the graded response model. *Applied Psychological Measurement, 17,* 335–350.

Cohen, A. S., Kim, S.-H., & Wollack, J. A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement, 20,* 15–26.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72,* 19–29.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*, 662–680.

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

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

Efron, B. (1977). The efficiency of Cox's Likelihood function for censored data. *Journal of the American Statistical Association, 72***,** 557–565.

Embretson, S.E., & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Finch, H (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measurement*, *29*, 278-295.

Finch, W.H., & French, B.F. (2007). Detection of crossing differential item functioning: A comparison of four methods. *Educational and Psychological Measurement, 67,* 565-582.

Fischer, G.H. (1993). Notes on the Mantel-Haenszel procedure and another chi-squared test for the assessment of DIF. *Methodika, 7,* 88-100.

Fischer, G.H. (1995). Some neglected problems in IRT. *Psychometrika, 60,* 449-487.

Fox, J.P., & Glas, C.A.W. (2001). Bayesian estimation of a multi-level IRT model using Gibbs sampling. *Psychometrika, 66,* 271-288.

Fox, J. -P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*, 169–191.

French, B. F., & Maller, S. J. (2007). Iterative purification and effect size use with logistic regression for differential item functioning detection. *Educational and Psychological Measurement, 67*, 373-393.

Gelman A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.

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

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

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

Glas, C. A. W., & Falcón, J. C. S. (2003). A comparison of item-fit statistics for the three-parameter logistic model. *Applied Psychological Measurement, 27*, 87-106.

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.

Glas, C.A.W., & Verhelst, N.D. (1989). Extensions of the partial credit model. *Psychometrika, 54*, 635-659.

Glas, C.A.W., & Verhelst, N.D. (1995). Testing the Rasch model. In: G. H. Fischer & I.W. Molenaar (eds.). *Rasch models. Their foundations, recent developments and applications.* (pp.69-96). New York: Springer.

Glas, C.A.W., & Verhelst, N.D. (1995). Tests of fit for polytomous Rasch models. In G. H. Fischer & I.W. Molenaar (eds.). *Rasch models. Their foundation, recent developments and applications.* (pp.325-352). New York: Springer.

Guttman, I. (1967). The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society, Series B*, *29,* 83-100.

Hambleton, R.K., & Han, N. (2005). Assessing the fit of IRT models to educational and psychological test data: A five step plan and several graphical displays. In: Lenderking, W.R., Revicki, D. (Eds.), *Advances in Health Outcomes Research Methods, Measurement, Statistical Analysis, and Clinical Applications*, Degnon Associates, Washington.

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.

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*, 577–601.

Holland, P., & Rosenbaum, P. (1986). Conditional association and unidimensionality in monotone latent variable models. *Annals of Statistics, 14*, 1523–1543.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Holland & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18,* 117–144.

Hulin, C. L., Lissak, R. I., & Drasgow, R. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Jansen, M., & Glas, C. A. W. (2005). Checking the Assumptions of Rasch's Model for Speed Tests. *Psychometrika, 70*, 671-684.

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

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, *49*, 223-245.

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika*, *54*, 681-697.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27*, 307-327.

Kim, S.-H., & Cohen, A. S. (1998). Detection of differential item functioning under the graded response model with the likelihood ratio test. *Applied Psychological Measurement, 22,* 345–355.

Klauer, K.C., & Rettig, K. (1990). An approximately standardized person test for assessing consistency with a latent trait model. *British Journal of Mathematical and Statistical Psychology, 43,* 193–206.

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

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

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading M.A: Addison-Wesley.

Maller, S. J. (2001). Differential item functioning in the WISC-III: Item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement, 61*, 793-817.

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

Maydeu-Olivares, A., & Joe, H. (2005). Limited- and full-information estimation and goodness-of-fit testing in 2(n) contingency tables: A unified framework. *Journal of the American Statistical Association, 100,* 1009-1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71,* 713-732.

Maydeu-Olivares, A., & Cai, L. (2006). A cautionary note on using G^2(dif) to assess relative model fit in categorical data analysis. *Multivariate Behavioral Research, 41,* 55–64.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Erlbaum.

Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods, 7,* 361–388.

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

Meredith, W., & Millsap, R.E. (1992). On the misuse of manifest variables in the detection of measurement bias. *Psychometrika*, *57*, 289–311.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

McKinley, R., & Mills, C. (1985). A comparison of several goodness-of-fit statistics. *Applied Psychological Measurement, 19*, 49-57.

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika, 49*, 359–381.

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

Mislevy, R. J., & Bock, R. D. (1989). A hierarchical item-response model for educational testing. In R. D. Bock (Ed.), *Multilevel analysis for educational data* (pp. 57-74). San Diego, CA: Academic Press.

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

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

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* Berlin: De Gruyter.

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

Muthén, B. O. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213-238). Hillsdale, NJ: Lawrence Erlbaum.

Nering, M.L. (1995). The distribution of person fit statistics using true and estimated person parameters. *Applied Psychological Measurement, 19*, 121–129.

Navas-Ara, M. J. & Gómez-Benito, J. (2002). Effects of ability scale purification on identification of DIF. *European Journal of Psychological Assessment*, *18*, 9-15.

Oort, F. J. (1998). Simulation study of item bias detection with restricted factor analysis. *Structural Equation Modeling, 5*, 107-124.

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

Orlando, M., & Thissen, D. (2003) Further examination of the performance of S-X^2, an item fit index for dichotomous item response theory models. *Applied Psychological Measurement, 27*, 289-298.

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

Patz, R.J., & Junker, B. (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.

Penfield, R.D., & Camilli, G. (2007). Differential item functioning and item bias. In S. Sinharay & C.R. Rao (Eds.), *Handbook of Statistics, Volume 26: Psychometrics* (pp. 125-167). New York: Elsevier.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87,* 517–529.

Rao, C.R. (1948). Large sample tests of statistical hypothesis concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society, 44,* 50–57.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114,* 552–566.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Vol. 1. Equivalence in measurement* (pp. 21–50). Greenwich, CT: Information Age.

Rigdon, S. E., & Tsutakawa, R. K. (1983). Estimation in latent trait models. *Psychometrika, 48*, 567–574.

Roussos, L. A., & Stout, W. (2004). Differential item functioning analysis. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 107-116). Thousand Oaks, CA: Sage.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12,* 1151-1172.

Rubin, D. B., & Stern, H. S. (1994). Testing in latent class models using a posterior predictive check distribution. In A. von Eye & C.C. Clogg (Eds.), *Latent variables analysis. Applications for developmetnal research* (pp. 420-438). Thousand Oaks: Sage.

Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37-52.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.

Sinharay, S. (2005), Assessing fit of unidimensional item response theory models using a Bayesian approach. *Journal of Educational Measurement, 42,* 375-394.

Sinharay, S., Johnson, M. S., & Stern, H. S. (2006). Posterior predictive assessment of item response theory models. *Applied Psychological Measurement, 30*, 298-321.

Sinharay, S., Johnson, M. S., & Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Educational and Behavioral Statistics*, *28*, 295–313.

Sijtsma, K., & Meijer, R. R. (1992). A method for investigating the intersection of item response functions in Mokken's nonparametric IRT model. *Applied Psychological Measurement, 16*, 149–157.

Silvey, S. D. (1959). The Lagrangian multiplier test. *Annals of Mathematical Statistics, 30*, 389-407.

Smith, R.M. (1985). A comparison of Rasch person analysis and robust estimators. *Educational and Psychological Measurement*, *45*, 433–444.

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

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27,* 229–239.

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

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2004). Examining the effects of differential item/test functioning (DIF/DTF) on selection decisions: When are statistically significant effects practically important? *Journal of Applied Psychology, 89,* 497–508.

Stark, S., Chernyshenko, O., & Drasgow, F. (2006). Detecting Differential Item Functioning with CFA and IRT? Toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306.

Steenkamp, J. E., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25,* 78–90.

Stone, C.A., & Zhang, B. (2003). Assessing goodness of fit of item response theory models: A comparison of traditional and alternative procedures. *Journal of Educational Measurement, 40***,** 331–352.

Stout, W.F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, *52*, 589-617.

Stout, W.F. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika*, *55*, 293-325.

Swaminathan, H., & Rogers, H.J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement*, *27*, 361-370.

Tatsuoka, K.K. (1984). Caution indices based on item response theory. *Psychometrika, 49*, 95–110.

Tsutakawa, R.K., & Johnson J.C. (1990). The Effect of Uncertainty of Item Parameter Estimation on Ability Estimates. *Psychometrika, 55*, 371-390.

Tsutakawa, R.K., & Soltys M.J. (1988). Approximation for Bayesian Ability Estimation. *Journal of Educational Statistics, 13*, 117-130.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D. (1991). *MULTILOG user's guide (Version 6.0) [Computer manual]*. Mooresville, IN: Scientific Software.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Van Onna, M. J. H. (2003). General and specific misfit of nonparametric IRT-models. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J.J. Meulman (Eds.), *New developments in psychometrics: Proceedings of the International Meeting of the Psychometric Society* (IMPS 2001) (pp. 239-246). Tokyo: Springer Verlag.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods, 5*, 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Verhelst, N. D., & Glas, C. A. W. (1995). The generalized one parameter model: OPLM. *In Rasch Models: Their Foundations, Recent Developments and Applications* (Edited by G.H. Fischer and I. W. Molenaar). Springer, New York.

Wang, W.-C., & Su, Y.-H. (2004a). Effects of average signed area between two item characteristic curves and test purification procedures on the DIF detection via the Mantel-Haenszel method. *Applied Measurement in Education*, *17*, 113-144.

Wang, W.-C., & Su, Y.-H. (2004b). Factors Influencing the Mantel and generalized Mantel-Haenszel methods for the assessment of differential item functioning in polytomous items. *Applied Psychological Measurement, 28*, 450-480.

Wilson, M., & Masters, G. N. (1993). The partial credit model and null categories. *Psychometrika, 58*, 87-99.

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

Wollack, J. A., Cohen, A. S., & Wells, C. S. (2003). A method for maintaining scale stability in the presence of test speededness. *Journal of Educational Measurement*, 40, 307–330.

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

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.

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: Scientific Software International, Inc.