View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Duisburg-Essen Publications Online

Universität Duisburg-Essen

Fakultät für Bildungswissenschaften

Lehrstuhl für Lehr-Lernpsychologie

Person Fit Analysis with Simulation-based Methods

Dissertation zur Erlangung des Grades Dr. phil.

vorgelegt von Christian Spoden

geboren am 27.01.1982 in Mülheim a.d. Ruhr

Erstgutachter: Prof. Dr. Dr. Detlev Leutner, Universität Duisburg-Essen

Zweitgutachter: Prof. Dr. Christian Tarnai, Universität der Bundeswehr München

Tag der mündlichen Prüfung: 16. Juli 2014

DANKSAGUNG (in German)

Mein großer Dank gilt Herrn Prof. Dr. Leutner für die gute kontinuierliche Betreuung der Arbeit und die Bereitstellung der nicht unerheblichen Ressourcen („computational power"), welche diese Arbeit erst ermöglicht haben.

Bedanken möchte ich mich zudem bei Herrn Prof. Dr. Tarnai für seine Bereitschaft, das Promotionsverfahren als Zweitgutachter zu unterstützen.

Einen besonderen Dank schulde ich Jens (Fleischer), der bei den zahlreichen Herausforderungen im Zusammenhang mit dieser Arbeit stets gleichermaßen ein offenes Ohr und guten Rat hatte.

Großer Dank gilt auch einer ganzen „Heerschar" von studentischen Hilfskräften, die mir in unterschiedlichen Phasen der Arbeit zahlreiche kleinere Aufgaben abgenommen haben.

Ich danke meinen Eltern, die mich während der Erstellung der Arbeit unterstützt haben. Ihnen sei die Arbeit gewidmet.

Der letzte und größte Dank gilt Jophia.

LIST OF CONTENT

LIST OF PREVIOUS PRESENTATIONS AND PUBLICATIONS OF PARTS OF THIS SCRIPT

Chapter 3:

Spoden, C., Fleischer, J. & Leutner, D. (2014). Applying the Rasch Sampler for person fit analysis under fixed nominal alpha level. *Journal of Applied Measurement, 15*, 276-291.

Spoden, C., Fleischer, J., Zischka, V., & Leutner, D. (2011, September). Hypothesentests bei Rasch Personen-Fit-Statistiken – Eine Alternative zum konventionellen Monte-Carlo-Verfahren [Hypothesis testing for Rasch person-fit ststistics – an alternative to conventional Monte Carlo methods]. Paper presented at 10th Meeting of the Fachgruppe für Methoden und Evaluation der deutschen Gesellschaft für Psychologie, Bamberg, Germany.

Spoden, C., Fleischer, J., & Leutner, D. (2011, Juli). Applying the Rasch Sampler to identify aberrant responding by person fit statistics under fixed α-level. Paper presented at the International Meeting of the Psychometric Society (IMPS), Hongkong.

Chapter 4:

Spoden, C., & Fleischer, J. (2012, September). Parametrische Personen-Fit-Statistiken mit robusten Fähigkeitsschätzern [Parametric person fit statistics with robust ability estimates]. Paper presented at 77th Meeting of the Arbeitsgruppe für Empirische Pädagogische Forschung (AEPF), Bielefeld, Germany.

Spoden, C., & Fleischer, J. (2012, April). Person fit analysis using robust latent trait estimates. Paper presented at the International Objective Measurement Workshop (IOMW), Vancouver, Canada.

LIST OF ABBREVIATIONS

2PLM	two parameter logistic item response model
3PLM	three parameter logistic item response model
4PLM	four parameter logistic item response model
AMT	Sine M-estimator in Mislevy and Bock (1982)
BS	bisquare weight (latent) ability estimates
CAT	computerized adaptive testing
CML	conditional maximum likelihood estimation
cor_{bis}	point biserial correlation
CTT	classical test theory
DMM	Mokken double monotonicity model
EAP	expected a posteriori (latent) ability estimates
EAP_{adj}	expected a posteriori estimator (latent) ability estimates adjusted
HU	Huber-type weighted (latent) ability estimates
IRF	item response function
IRT	item response theory
KFT	Kognitiver Fähigkeitstest
MAD	mean absolute difference
MAP	modal a posteriori (Bayes modal) (latent) ability estimates
MC	Monte Carlo
MCMC	Markov chain Monte Carlo estimation
MCMC(RS)	Markov chain Monte Carlo estimation
ML	maximum likelihood (latent) ability estimates
MML	marginal maximum likelihood estimation
MNSQ	mean square statistic

NOR	normalized versions of a person fit statistic
PISA	Programme for International Student Assessment
SE	standard error
SNIJ	Snijders’ correction for linear person fit statistics
TC_{HU}	tuning constant for the HU (latent) ability estimates
TT	true theta
WIM	estimation scheme by B. Wright and R. Mead in Mislevy and Bock (1982)
WL	weighted maximum likelihood (latent) ability estimates

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Aberrant responding in test or questionnaire data violating the principles of Item Response Theory is a prevalent phenomenon in psychological and educational sciences. By means of person fit statistics aberrant responding is identified that prevents the computation of inadequate ability estimates. Simulation-based methods for person fit analysis were investigated in simulation studies with regard to Type I error and statistical power to detect aberrancy. Real data analyses from psychological and educational sciences further illustrate the usefulness of person fit statistics based on the presented approaches.

In Study 1, a Markov chain Monte Carlo algorithm for sampling data matrices denoted as the *Rasch Sampler* is applied for simulating the null distribution of person fit statistics under the Rasch model. Results are compared to standardized statistics and illustrate the new approach (1) to correctly recover the nominal Type I error rates (while the standardized statistics deviate substantially) and (2) to offer predominantly similar or higher statistical power. Results from the application to Rasch-scalability problems of two subscales taken from Heller and Perleth’s (2000) multidimensional intelligence test (KFT) confirmed findings from the simulation studies.

In Study 2, the Type I error and power of person fit tests based on weighted maximum likelihood ability estimators and parametric bootstrap were evaluated. Results were compared to established methods for person fit analysis. Bootstrapping based on robust maximum likelihood estimators improves the statistical power but a satisfactory recovery of nominal Type I error rates requires strong downweighting of aberrant item responses. Bootstrapping based on the Warm’s (1989) estimator applied as scoring method to original and simulated data displayed promising results concerning Type I error recovery and statistical power. Results from the simulations were matched by findings from the analysis of four samples of students with disabilities participating in a state-wide administered large-scale assessment program to investigate whether assessment of competence is invalidated by test modifications for these students.

Both studies provide new insights on the benefits of simulation-based methods for the application of person fit tests to detect aberrant response behavior.

Keywords person fit, Item Response Theory, Monte Carlo simulation, weighted maximum likelihood scoring

ZUSAMMENFASSUNG (in German)

Abweichendes Antwortverhalten in Test- und Fragebogendaten gegenüber den Annahmen der Item-Response-Theorie stellt ein häufiges Phänomen in der Psychologie und den Bildungswissenschaften dar. Personen-Fit-Statistiken können herangezogen werden, um derartiges Antwortverhalten zu identifizieren und die Schätzung inadäquater Fähigkeitsausprägungen zu verhindern. Simulations-basierte Methoden zur Personen-Fit-Analyse werden mit Hilfe von Simulationsstudien in Bezug auf Typ-I-Fehler und statistische Power untersucht. Real-Daten aus der Psychologie und Bildungsforschung werden genutzt, um die Bedeutung der Ergebnisse beispielhaft zu untermauern.

In Studie 1 wird der *Rasch Sampler*, ein Markov-Chain-Monte-Carlo-Algorithmus zur Ziehung binärer Datenmatrizen, herangezogen, um die Verteilung von Personen-Fit-Statistiken für das Rasch-Modell zu simulieren. Die Ergebnisse werden mit standardisierten Personen-Fit-Statistiken verglichen und verdeutlichen (1) die Einhaltung des nominalen Typ-I-Fehlers (im Gegensatz zu deutlichen Abweichungen der standardisierten Statistiken) sowie (2) überwiegend vergleichbare oder höhere statistische Power im neuen Ansatz. Die Anwendung der Methode auf die Forschungsfrage nach der Rasch-Skalierbarkeit von zwei Subskalen von Heller und Perleth's (2000) multidimensionalem Intelligenztest (KFT) unterstreicht Ergebnisse der Simulationsstudien.

In der zweiten Studie werden Typ-I-Fehler und statistische Power verschiedener (parametrischer) Personen-Fit-Statistiken basierend auf gewichteten Maximum-Likelihood-Fähigkeitsschätzern untersucht und mit etablierten Ansätzen verglichen. Ein parametrischer Bootstrap basierend auf robusten Maximum-Likelihood-Schätzern erhöht die statistische Power, jedoch fällt die Einhaltung des nominalen Typ-I-Fehlers nur dann zufriedenstellend aus, wenn der Einfluss abweichender Item-Antworten bei der Berechnung des Schätzers durch Wahl einer geeigneten Gewichtung stark reduziert wird. Ein parametrischer Bootstrap basierend auf Warms (1989) Schätzer, angewendet auf Original- und simulierte Daten, verzeichnet vielversprechende Ergebnisse bezüglich der Einhaltung des Typ-I-Fehlers sowie der statistischen Power. Ergebnisse der Simulationen werden durch Ergebnisse einer Analyse von vier Stichproben von Förderschülern ergänzt, welche Erkenntnisse zur Invarianz zwischen konventioneller und angepasster Testadministration bei einem regionalen Large-Scale Assessment Programm erlauben.

Die Ergebnisse der beiden vorliegenden Studien erbringen neue Erkenntnisse bezüglich der Vorteile simulations-basierter Methoden bei der Anwendung von Person-Fit-Statistiken.

Schlüsselwörter Personen-Fit, Item-Response-Theorie, Monte-Carlo-Simulation, gewichtete Maximum-Likelihood-Fähigkeitsschätzer

1 PERSON FIT ANALYSIS IN ITEM RESPONSE THEORY

Do teachers differ in their instructional ability? Do religious people and atheists differ in educational level and intelligence? Are reading comprehension and listening comprehension two different skills or do they collapse to receptive competence? And was Obama the most liberal senator to become President of the United States? Measurement models are used in various research areas of social sciences to find an individual's position on some type of unobservable (latent) dimension, for example when psychologists are interested in personality domains (e.g., Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001), educational researchers analyze educational skills and competence (e.g., Hartig & Höhler, 2008), or when political analysts seek to evaluate positions like liberalness or conservativeness of a legislator (e.g., Bafumi, Gelman, Park, & Kaplan, 2005). These positions (in the following denoted as latent abilities) have to be uncovered from the analysis of the individual's behavior, in many settings the response behavior under a given stimulus, like a question, a task or a decision. In psychological research test theories formulate a theoretical and statistical framework for this process and define the relationship of the individual's behavior and the latent ability (e.g., Hambleton, Swaminathan, & Rogers, 1991; Embretson & Reise, 2000; Partchev, 2004). Though – according to Rost (1999) – most of all (scientifically grounded) testing instruments were developed according to the principles of Classical Test Theory (CTT; Allen & Yen, 2002), large-scale assessment programs in educational settings like *National Assessment of Educational Progress* (Pellegrino, 1999; Raju, Pellegrino, & Bertental, 2000) or *Programme for International Student Assessment* (PISA; OECD, 2010) strongly rely on Item Response Theory (IRT; e.g., Hambleton, Swaminathan, & Rogers, 1991; Embretson & Reise, 2000; Partchev, 2004) and have promoted interest in these methods. IRT gives a clear mathematical description for latent ability estimation but as various sources of aberrancy occur in social sciences data, it is a matter of discussion whether each type of response behavior is accurately represented by the ability estimates and how inaccurately represented behaviors might be identified (e.g., Meijer, 1996).

1.1 A short introduction to item response theory

Following Hambleton, Swaminathan and Rogers (1991), IRT models rely on two related assumptions, unidimensionality and local (stochastic) independence.

Unidimensionality of the item sample or item homogeneity is assumed when the item responses are determined by one single (latent) ability. This is hard to find purely due to several “disturbances” in real data analysis like test anxiety, cheating or response styles (Meijer, 1996; see below). Therefore, unidimensionality is generally assumed when a single trait was found to dominate the responses of the individuals to a particular item set (essential unidimensionality; e.g., Stout, 1987).

Local (stochastic) independence of the item responses means that the probabilities for L item responses $x_{\mathrm{i}} = x_1, x_2, \dots, x_L$ under an ability θ are defined by

$$P(x_1, x_2, \dots, x_L|\theta) = \prod_{i=1}^{L} P(x_i|\theta). \qquad [1.1]$$

As a consequence from local (stochastic) independence, the likelihood of the model can be computed by building the product over the item responses given θ (see below). This is the basis for estimation methods like maximum likelihood, which will be described in more detail in the next section. Unidimensionality and local independence are related to the point that the second property is always obtained when the first holds while the reverse is not generally true (Hambleton, Swaminathan, & Rogers, 1991, Chapter 2).

Inferences on the latent ability are valid under these two and a third assumption that the model equation holds which defines the item characteristic (item response function; IRF). For the Rasch model (Rasch, 1960), which is often interpreted as a special type of IRT model, the model equation is given by

$$P\left(x_{ip} = 1|\theta_{\mathrm{p}}, b_i\right) = \frac{\exp(\theta_{\mathrm{p}} - b_i)}{1 + \exp(\theta_{\mathrm{p}} - b_i)} \qquad [1.2]$$

where θ_{p} is the latent ability for the individual p and b_i is the item difficulty for item i. Rasch-type models are characterized by a property (e.g., Baker & Kim, 2004, Chapter 5; Molenaar, 1995) called specific objectivity which addresses comparisons between two items or two individuals. Given that specific objectivity holds, comparisons between two items or between two individuals no more depend on distributional assumptions of the analyzed population or item sample and their metric is only determined by the matrix of item responses (Baker & Kim, 2004, Chapter 5). Specific objectivity is a consequence of the data matrix marginals serving as sufficient statistics for the latent ability and the item difficulties (i.e., the individuals' raw score contains all information about the latent ability, the number of correct responses on a particular item contains all information about the item difficulty). This issue is outlined in detail in Chapter 3.

From the point of item response modeling, the Rasch model is a rather restrictive model from a larger class of psychometric models (for a discussion see Andrich, 2004). The first relaxation of the Rasch model arises when a discrimination parameter a is introduced which converts equation 1.2 into

$$P\left(x_{ip} = 1|\theta_{\mathrm{p}}, a_i, b_i,\right) = \frac{\exp\left[a_i(\theta_{\mathrm{p}} - b_i)\right]}{1 + \exp\left[a_i(\theta_{\mathrm{p}} - b_i)\right]}. \qquad [1.3]$$

This model is known as two-parameter logistic latent trait model (2PLM). Even more relaxed models arise when the lower (c_i) and the upper asymptote (d_i) of the IRFs are estimated, which leads to the more flexible three (3PLM; Birnbaum, 1968) and four parameter IRT models (4PLM). The four parameter IRT model (Barton & Lord, 1981) is defined by

$$P\left(x_{ip} = 1|\theta_{\mathrm{p}}, a_i, b_i, c_i, d_i\right) = c_i + (d_i - c_i)\frac{\exp\left[a_i(\theta_{\mathrm{p}} - b_i)\right]}{1 + \exp\left[a_i(\theta_{\mathrm{p}} - b_i)\right]}. \qquad [1.4]$$

Note that substituting the logit by the probit link function leads to a second group of IRT models (e.g., Embretson & Reise, 2000, Chapter 4), which is mainly constituted by similar properties.

When the typical assumptions underlying standard IRT models (see above) are not met, the parameters of any of these common IRT models can be decomposed into additional model parameters. Two examples for the decomposition of the latent ability are presented here: For example, the assumption of unidimensional abilities is questionable for many constructs in psychology. Therefore, a multidimensional IRT model can be formulated (for the 2PLM; e.g., Reckase, 2009) according to

$$P\left(x_{ip} = 1|\theta_{\mathrm{ph}}, a_{ih}, b_i\right) = \frac{\exp\left(\sum_h a_{ih}\theta_{\mathrm{ph}} - b_i\right)}{1 + \exp\left(\sum_h a_{ih}\theta_{\mathrm{ph}} - b_i\right)} \qquad [1.5]$$

with a decomposition of the ability vector in h elements θ_1,…, θ_{H} and with a_{ih} as the weight (loading) of the particular ability. Another useful extension in educational settings is the decomposition of ability dispersion into individual and cluster level ability variance (where clusters may be formed by different classes, different schools, different measurement occasions etc.) by

$$\theta_{\mathrm{pg}} = \bar{\theta}_{..} + (\bar{\theta}_{.\mathrm{g}} - \bar{\theta}_{..}) + (\theta_{\mathrm{pg}} - \bar{\theta}_{.\mathrm{g}}) \qquad [1.6]$$

where $\bar{\theta}_{..}$ is the ability grand or overall mean, $\bar{\theta}_{.\mathrm{g}}$ is the group mean for cluster g, $(\bar{\theta}_{.\mathrm{g}} - \bar{\theta}_{..})$ is the variation of group means (the between-level variation) and $(\theta_{\mathrm{pg}} - \bar{\theta}_{.\mathrm{g}})$ is the variation of the respondents' individual ability from her or his particular cluster mean (the within-level variation; see e.g., Höhler, Hartig, & Goldhammer, 2010). Within- and between-level variation may also be regressed on covariates. This approach has become known as multilevel IRT (e.g., Fox & Glas, 2001; Kamata, 2001). As a main advantage to conventional approaches based on multilevel analysis of raw scores in CTT, multilevel IRT incorporates a useful method to handle uncertainty in the dependent variable by combining the IRT measurement model with a multilevel analysis (Fox & Glas, 2001). Several other forms of decompositions of IRT model parameters including item (property) covariates and person covariates are possible (e.g., De Boeck & Wilson, 2004).

Apart from these parametric IRT models, Mokken (1971) has proposed another psychometric approach relying on order restrictions and relaxed assumptions, which is currently known as nonparametric IRT (Sijtsma & Molenaar, 2002). The Double Monotonicity Model (DMM) is the more relevant of the two nonparametric IRT models. Unidimensionality, local independence and the property of monotonicity are assumed under the DMM (Sijtsma & Molenaar, 2002). The IRFs are also assumed to be non-intersecting but may have different forms which are not characterized by a limited number of parameters as it is the case in parametric IRT.

To demonstrate differences between typical IRT models, Figure 1.1 shows IRFs under parametric IRT and the DMM.

1.1.1 Estimation of item response functions in parametric IRT

In parametric item response theory, the item response functions are derived from a relatively small number of parameters, most often estimated by variants of maximum likelihood methods (e.g., Baker & Kim, 2004). For the Rasch model it is based on solving the likelihood equation

$$L(\theta, b|x) = \prod_{p=1}^{N}\prod_{i=1}^{L} P(x_{pi}|\theta_{\mathrm{p}}, b_i) = \prod_{p=1}^{N}\prod_{i=1}^{L} \frac{\exp[x_{pi}(\theta_{\mathrm{p}} - b_i)]}{1 + \exp(\theta_{\mathrm{p}} - b_i)} \qquad [1.10]$$

by iterative algorithms like Newton Raphson or expectation maximization (e.g., Tanner, 1994). The main difficulty arises from the existence of unknown quantities of both abilities and item parameters. Therefore, modern maximum likelihood methods divide IRT parameter estimation in two steps, an item parameter estimation step where θ is removed from the likelihood equation and a subsequent ability estimation step which is operated with the estimated item parameters treated as known (Baker & Kim, 2004).

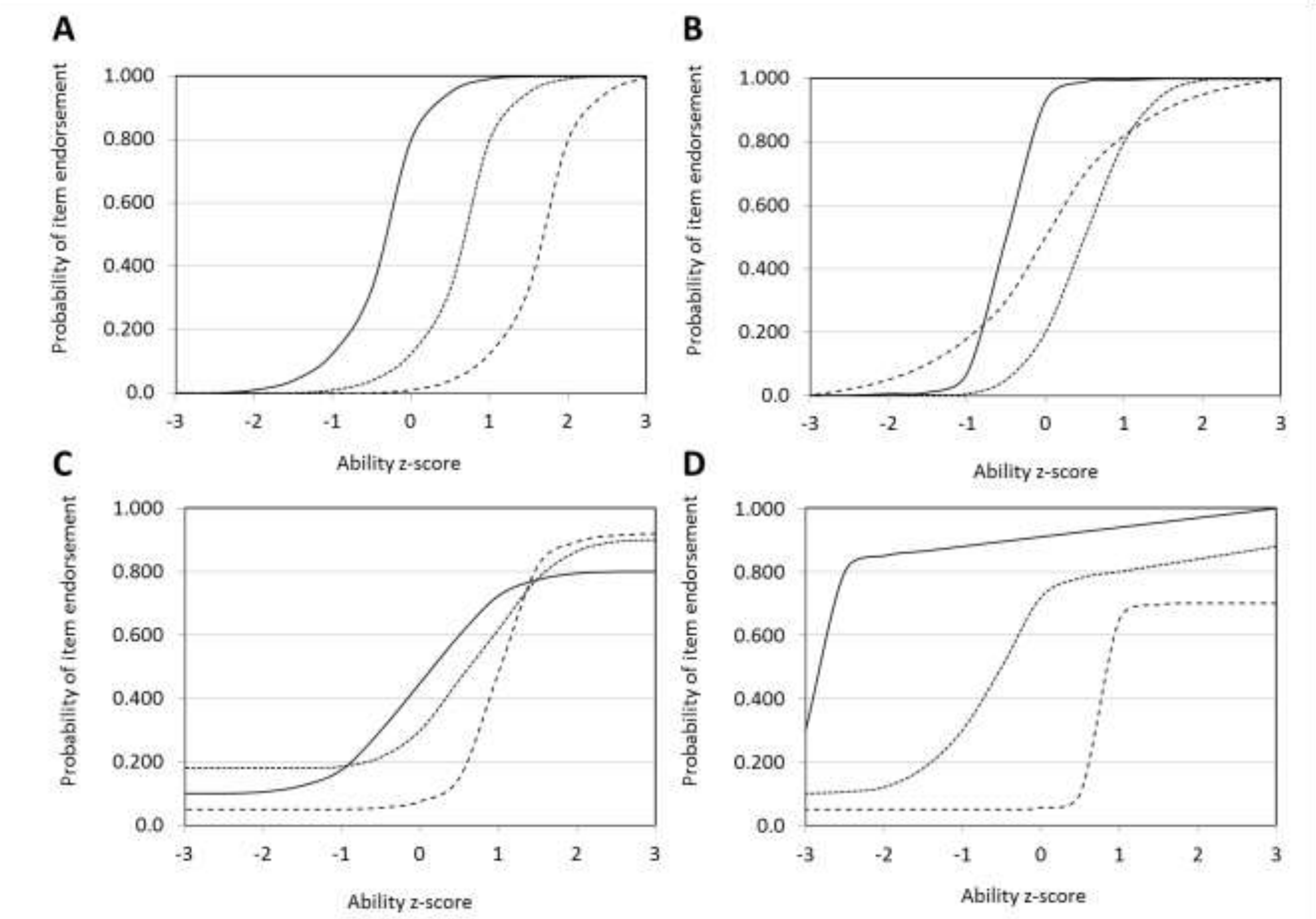

***Figure 1.1*. Item response functions under the Rasch model (A), the 2PLM (B), the 4PLM (C) and Mokkens DMM (D).**

Several variants of maximum likelihood methods have been developed for item parameter estimation but two of these are most prevalent in software packages (for an overview on the technical details of each estimation method see, e.g., Baker & Kim, 2004). Marginal maximum likelihood (MML) estimation is the more flexible approach but also more voluminous to outline. Complete MML algorithms are, for example, given by Baker and Kim (2004, Chapter 6), Harwell, Baker and Zwarts (1988) or Thissen (1982). As a basic idea of MML, the latent abilities θ are numerically integrated out to estimate the item parameters which involves assumptions of a latent marginal population distribution and principles of Bayes statistics. A numerical solution is usually found by applying an EM-algorithm (e.g., Baker & Kim, 2004, Chapter 6; Tanner, 1994, Chapter 4). In contrast to the MML method, conditional maximum likelihood estimation (CML; e.g., Baker & Kim, 2004, Chapter 5; Embretson & Reise, 2000, Chapter 8; Mair & Hatzinger, 2007) offers an option for item parameter estimation of Rasch-type models. Due to its relative simplicity compared to MML, the principle of CML is described here in more detail to illustrate how the likelihood with two unknown quantities can be solved. CML estimation for the Rasch model is provided as follows (e.g., Baker & Kim, 2004, Chapter 5; Embretson & Reise, 2000, Chapter 8; Mair & Hatzinger, 2007): By substituting the ability by $\mathrm{\xi_p} = \exp(\mathrm{\theta_p})$ and given the Rasch “easiness“ $\mathrm{\varepsilon_i} = \exp(-\,b_i)$, the Rasch model is now defined by

$$P\left(x_{ip} = 1 | \mathrm{\xi_p}, \mathrm{\varepsilon_i}\right) = \frac{\mathrm{\xi_p \varepsilon_i}}{1+\mathrm{\xi_p \varepsilon_i}}. \qquad [1.11]$$

The probability of a particular response vector $\boldsymbol{x}_p$ is then given by multiplying the (predicted) probabilities over the L items by

$$P\left(\boldsymbol{x}_p | \mathrm{\xi_p}, \boldsymbol{\varepsilon}\right) = \prod_{i=1}^{L} \frac{(\mathrm{\xi_p \varepsilon_i})^{x_i}}{1+\mathrm{\xi_p \varepsilon_i}} = \frac{\mathrm{\xi_p}^{r_p} \prod_{i=1}^{L} \mathrm{\varepsilon_i}^{x_{ip}}}{\prod_{i=1}^{L}(1+\mathrm{\xi_p \varepsilon_i})}, \qquad [1.12]$$

and the probability for a given raw score *r* under any of the response vectors satisfying $\sum_{i=1}^{L} x_i = r_p$ is now

$$P(r_p|\xi_\mathrm{p}, \boldsymbol{\varepsilon}) = \sum_{x|r_p} \prod_{i=1}^{L} \frac{(\xi_\mathrm{p}\varepsilon_\mathrm{i})^{x_i}}{1+\xi_\mathrm{p}\varepsilon_\mathrm{i}} = \frac{\xi_\mathrm{p}{}^{r_p} \sum_{\boldsymbol{X}|r_p} \prod_{i=1}^{L} \varepsilon_\mathrm{i}{}^{x_{ip}}}{\prod_{i=1}^{L}(1+\xi_\mathrm{p}\varepsilon_\mathrm{i})} \qquad [1.13]$$

where $\Sigma_{x|r_p}$ describes the sum across all these response vectors with raw score *r* and where

$$\gamma_\mathrm{r}(\varepsilon_\mathrm{i}) = \sum_{\boldsymbol{X}|r_p} \prod_{i=1}^{L} \varepsilon_\mathrm{i}{}^{x_{pi}} \qquad [1.14]$$

is efficiently estimated by several algorithms (e.g., Gustafsson, 1980; Liou, 1994). The basic symmetric function γ_r gives the combinatoric solution to reaching a score of *r* by products of ε_i. In case of a complete design, the symmetric functions are given by

$\gamma_0 = 1,$

$\gamma_1 = \varepsilon_1 + \varepsilon_2 + \cdots + \varepsilon_\mathrm{L},$

$\gamma_2 = \varepsilon_1\varepsilon_2 + \varepsilon_1\varepsilon_3 + \cdots + \varepsilon_{\mathrm{L}-1}\varepsilon_\mathrm{L},$

…

$\gamma_\mathrm{L} = \varepsilon_1\varepsilon_2 \ldots \varepsilon_\mathrm{L}.$

The probability of an item response vector ***x*** under a known raw score *r* and **ε** is defined by the ratio of Equation 1.12 and Equation 1.13 (e.g., Embretson & Reise, 2000, pp. 222-225; Mair & Hatzinger, 2007). This ratio simplifies as the (scaled) abilities ξ and the common denominator vanish to the following formula:

$$P(\boldsymbol{x_p}|\boldsymbol{r_p}, \boldsymbol{\varepsilon}) = \frac{\prod_{i=1}^{L} \varepsilon_\mathrm{i}{}^{x_{pi}}}{\gamma_\mathrm{r}(\varepsilon_\mathrm{i})} = \frac{\prod_{i=1}^{L} \varepsilon_\mathrm{i}{}^{x_{pi}}}{\sum_{\boldsymbol{X}|r_p} \prod_{i=1}^{L} \varepsilon_\mathrm{i}{}^{x_{pi}}} \qquad [1.15]$$

By computing the product over all persons in the sample ***X*****,** it follows:

$$P(\boldsymbol{X}|\boldsymbol{r},\boldsymbol{\varepsilon}) = \prod_{p=1}^{N} \frac{\prod_{i=1}^{L} \varepsilon_{\mathrm{i}}^{x_{pi}}}{\gamma_{\mathrm{r}}(\varepsilon_{\mathrm{i}})} = \prod_{p=1}^{N} \frac{\prod_{i=1}^{L} \varepsilon_{\mathrm{i}}^{x_{pi}}}{\sum_{\boldsymbol{X}|r_p} \prod_{i=1}^{L} \varepsilon_{\mathrm{i}}^{x_{pi}}}, \qquad [1.16]$$

or when returning to the item difficulties b_i:

$$P(\boldsymbol{X}|\boldsymbol{r},\boldsymbol{b}) = \prod_{p=1}^{N} \frac{\exp(-\sum_{i=1}^{L} x_{pi}\, b_i)}{\sum_{\boldsymbol{X}|r} \exp(-\sum_{i=1}^{L} x_{pi}\, b_i)}. \qquad [1.17]$$

This is the conditional likelihood for the sample (e.g., Embretson & Reise, 2000, Chapter 8; Mair & Hatzinger, 2007). Note that this likelihood needs to be solved iteratively, e.g., by the Newton-Raphson algorithm which involves the first (partial) derivative of the basic symmetric functions. According to Pfanzagl (1994), CML has preferable statistical properties like consistency, asymptotic unbiasedness, asymptotic efficiency and asymptotic normal distribution. As a main drawback of CML, this method is only used for Rasch-type IRT models (but see Verhelst & Glas, 1995). Under valid distributional assumptions, also MML is assumed to have similar properties as CML (e.g., de Leeuw & Verhelst, 1986; Pfanzagl, 1994).

Other popular approaches include joint maximum likelihood estimation (Wright & Panchepakesan, 1969) and nonparametric approaches based on splines (Woods & Thissen, 2006) or Ramsay curve smoothing (Ramsay, 1991; Woods, 2006, 2008b). Additionally, alternative methods from Bayesian statistics based on the Markov chain Monte Carlo approach (MCMC) have grown rapidly in psychometric research in the last two decades (Fox, 2010; Jackman, 2009, Chapter 9). MCMC is an umbrella term for several iterative sampling algorithms that are based on the construction of a Markov chain (e.g., Meyn & Tweedie, 2009). Sampling schemes for MCMC estimation of Bayesian IRT models are given by Albert (1992) or Patz and Junker (1999) beside others; Fox (2010) has given an overview on Bayesian methods for IRT. There is also a growing number of research papers relying on MCMC that offer new options for IRT model estimation in a Bayesian nonparametric framework (e.g., Duncan & MacEachern, 2008; Karabatsos & Walker, 2009; Miyazaki & Hoshino, 2009; San Martín, Jara, Rolin, & Mouchart, 2011). While both item parameters and abilities are generated in joint maximum likelihood estimation and in Bayesian approaches, remember the latent abilities to be estimated separately when the more common MML or CML methods are applied.

1.1.2 Latent ability estimation

In this script, the term "ability" is applied for many different characteristics ranging from clinical symptom scores over political positions to cognitive or personality measures. In the same way as the item parameters, the latent abilities are computed iteratively under most scoring methods (Baker & Kim, 2004, Chapter 3 and 7). Both maximum likelihood and Bayesian methods are applied to determine the ability estimates. Common methods are (Baker & Kim, 2004, Chapter 3 and 7; Hoijtink & Boomsma, 1995):

- (Unweighted) maximum likelihood estimation (ML; e.g., Baker & Kim, 2004, Chapter 3),
- Weighted maximum likelihood estimation (WL) by Warm (1989),
- Robustly weighted maximum likelihood estimation by bisquare weighting (BS; Mislevy & Bock, 1982) or Huber-type weighting (HU; Schuster & Yuan, 2011),
- Expected a posteriori estimation (EAP) and modal a posteriori estimation (also empirical Bayes estimates; MAP).

Methods 1-3 belong to the group of maximum likelihood scoring methods while EAP and MAP are Bayesian estimators. A description of some of these methods is given in Chapter 4. It is not intended to give a detailed discussion on statistical properties of scoring methods, but previous analyses found, for example, that the WL has small bias while EAP / MAP are inwards and ML outwards biased (in finite samples; Hoijtink & Boomsma, 1995). With response disturbances being present in the data, also BS and HU are less biased than the ML with the BS displaying even lower bias than the HU (Schuster & Yuan, 2011). Due to using prior information, the Bayesian estimators EAP / MAP have smaller variance than WL and in particular ML (Hoijtink & Boomsma, 1995). The HU was found to have a smaller sampling variance than the BS (Schuster & Yuan, 2011). For more details see Hoijtink and Boomsma (1995) or Schuster and Yuan (2011).

Considering the various sources of aberrancy in social sciences data in general and educational data in particular, the question arises whether these scoring methods generally display the underlying ability in an accurate way.

1.1.3 Inaccuracy of ability estimates

Results from different fields of psychological research raise some doubt on the assumption that latent ability estimates generated by the methods described above are generally accurate indicators of a respondents' true ability: Cognitive and affective factors like fatigue, test anxiety or inattention may impair the performance of examinees in a testing situation (e.g., Haladyna, 2004, Chapter 10; Meijer, 1996). To stress only one single example, imagine a competent examinee suffering from test anxiety while working on an exam. Answering the first items the examinee will be unable to concentrate due to cognitive interference as a main feature of test anxiety (Tobias, 1992). After some time the examinee has adapted to the situation, feels more confident and his responses on the following items may therefore reflect his true competence. But when the complete response vector is analyzed to estimate the ability, a spuriously low score might be observed for the anxious examinee and the true ability might be underestimated.

Several such types of aberrant response behaviors have been collected by Meijer (1996) and Haladyna (2004, Chapter 10). Table 1.1 gives hypothetical vectors for these types of responding and shows the point biserial correlation cor$_{bis}$ with the fictitious percentage of correct responses to these items. The cor$_{bis}$ is positive for model-conform vectors while it shows negative values for vectors arising from aberrancy. Hence, cor$_{bis}$ is an indicator of aberrancy of the respective vector and belongs to the class of person fit statistics to be discussed below.

Though the underlying mechanisms of aberrancy are not always clear, there is some evidence that aberrant responses may go along with several problems including biased ability estimates (see Chapter 1.2.1). Consequently, much effort has been exerted to develop and improve tools for the detection of response aberrancy. This field of research is referred to as person fit analysis (Meijer & Sijtsma, 2001).

1.2 Person fit

In item response theory the adequacy of a measurement model can be analyzed by means of statistical indicators and tests (e.g., Embretson & Reise, 2000, Chapter 9; Hambleton, Swaminathan, & Rogers, 1991, Chapter 4). The usage of item and person fit statistics has spread in typical IRT software packages. Item fit analysis allows identifying items which display low conformity with the IRFs (e.g., Reise, 1990); this may, for example, be a sign of systematic item misinterpretation by the respondents. Person fit statistics can be used to evaluate the consistency of each individual response vector with the applied model and have proven to be a useful psychometric tool considering the serious consequences related to individual aberrancy (Meijer, 1996; Meijer & Sijtsma, 2001).

1.2.1 Consequences related to individual misfit

Assessments often involve major implications for the examinee. Inaccurate ability estimates may cause unfair decisions when, for example, unqualified individuals are being awarded a degree, qualified individuals are denied a degree or being excluded from academic or professional programs (e.g., Schmitt et al., 1999). Disadvantages may also arise in more informal educational assessments when curriculum or instructional processes are adjusted by teachers based on feedback about the competence of students (Leutner, Fleischer, Spoden, & Wirth, 2007). The effects of person misfit have been studied intensively in psychometric research. In summary, results indicate that effects of low person fit of an individual's response vector may include inaccurate latent abilities being estimated for misfitting individuals (Meijer & Nering, 1997), a decrease in correct diagnostic classification (mastery) decisions (Hendrawan, Glas, & Meijer, 2005) and a decrease in the validity of the test instrument (Meijer, 1997, 1998; Schmitt, Cortina, & Whitney, 1993; Schmitt et al., 1999).

Table 1.1

Fictious vectors of Rasch-conform and Rasch-aberrant response behavior

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	cor$_{bis}$
Rasch vectors	1	1	1	0	1	0	0	0	0	0.735
	1	1	1	0	1	0	1	0	0	0.572
	1	0	1	1	0	1	0	1	0	0.245
	1	1	1	1	0	1	1	0	0	0.602
test anxiety	0	0	0	1	1	1	1	1	1	-0.775
uninformed guessing	0	0	0	1	0	0	1	0	0	-0.097
cheating	1	1	0	0	0	0	1	1	1	-0.163
inattention	1	0	1	1	1	1	1	0	0	0.344
item difficulty	.90	.80	.70	.60	.50	.40	.30	.20	.10	

Notes. One is a correct item response, zero is an incorrect item response.

Inaccurate latent ability estimation: Meijer and Nering (1997) analyzed the effects of misfitting response vectors on latent abilities estimated by three scoring methods (ML, EAP, BS). Results illustrated that latent ability estimates were indeed biased for misfitting response vectors though results differed systematically for the different scoring methods, different types of model violations and different levels of true ability (e.g., extreme levels of true θ were generally much more biased than medium levels). By applying the robust BS method, bias was not only reduced for extreme abilities compared to the scoring methods ML or EAP, but also the statistical power of the applied person fit statistic (l_z) to detect model violations was found to be enhanced (Meijer & Nering, 1997).

Decrease of correct mastery decisions: Hendrawan, Glas and Meijer (2005) studied the interaction of person fit and mastery classification decisions. They found that for fitting response vectors and several scoring methods, the classification rates were very much comparable to the mastery decisions in groups formed according to the true status. For misfitting response vectors, however, serious impact on the classification decision was found and especially in one of the simulated conditions, random guessing, classifications were found to be random decisions. The authors concluded "... that classification decisions cannot be justified if the model does not fit.“ (p. 43) and "…that person-fit statistics can indeed be used for identifying a subsample in which the model fits and mastery testing is appropriate.“ (p. 43).

Effects on validity: Effects of person misfit on test validity were studied in both real and simulated data. Analyzing four data sets from personnel selection, Schmitt, Cortina and Whitney (1993) found an interaction effect with person fit in the analysis of criterion-related validity estimated as a regression of supervisory performance ratings on performance test scores in a sample of mechanics. Results indicated that applying a person fit statistic to identify questionable response vectors might help to improve the validity of test results. In a second study, Schmitt et al. (1999) re-investigated person fit as a moderator in the regression of students' grade-point average on cognitive and personality test scores (as a measure of the criterion-related validity) in a sample of undergraduate university students. The validity was substantially higher for model-conform respondents compared to misfitting respondents. The authors concluded that „…person fit can have a substantial practical impact on the validity of tests for subgroups identified by level of fit." (p. 49). Meijer (1998) analyzed item responses by candidates for a job application in computer occupations (e.g., systems analysts, programmers etc.). Compared to respondents with conform vectors, he found persons with misfitting responses to be less predictable by results from an intelligence tests with regard to behavioral assessment ratings collected over four years. Removing misfitting respondents from the data set increased the correlations between predictor and criterion. While most studies relied on real data analysis, Meijer (1997) modeled the effect of misfit on test-score validity in a simulation study with a systematic (experimental) variation of several factors. In the presence of misfitting respondents, validity was indeed deteriorated – but this effect was small in magnitude and was only found when a strong correlation between the predictor and criterion scores (.3 or .4) existed and a high percentage of misfit (.15 or higher) was present. Removing misfitting item-score vectors from a predictor influenced results on criterion-related validity very modestly.

In summary, results from person fit analysis indicate that low person fit may indeed be related to inaccurately estimated abilities and may impair the validity of test results under some conditions. The magnitude of these problems certainly depends on the testing context and on the statistical power of the particular person fit statistic applied to identify misfit.

1.2.2 Nonparametric, parametric and other types of person fit statistics

The number of proposed person fit statistics has grown rapidly in the last decades. For readers interested in the details of the several statistics, Meijer and Sijtsma (2001) present a useful overview including information on similarities between the statistics, and Karabatsos (2003) provides a general comparison of correct fit / misfit classifications under 36 statistics for the Rasch model. Differences between the two main classes of person fit statistics, nonparametric and parametric statistics, are outlined below.

Group-based nonparametric statistics: Group-based statistics do not include any (predominantly iteratively estimated) IRT parameters and are therefore also denoted as nonparametric statistics. These statistics are based on the computation of correlations (or covariances) between the individual response vectors of two respondents or between an individual response vector and some aggregated quantity of the sample, like, e.g., percentages of correct responses per item. The estimates of the point-biserial correlation between the individual responses and the item difficulties presented in Table 1.1 of the previous chapter, and statistic *U3* presented in Chapter 3 are examples for group-based nonparametric statistics. Karabatsos (2003) found that four out of the best five performing person fit statistics for the Rasch model were group-based statistics. Several other studies based on the two- and three-parameter model found that group-based nonparametric statistics perform about as well as parametric statistics (Emons, 2008, 2009; Meijer, 1994). A major problem with nonparametric statistics is the absence or inadequacy (Emons, Meijer, & Sijtsma, 2002) of standardizations which complicates the definition of cut values to decide on whether a response vectors does or does not fit according to the IRT model parameters. This issue is addressed in Chapter 3 based on the Rasch model.

Parametric IRT-based statistics: Parametric statistics either rely on IRT parameters or on the (estimated) item response probability (from now on denoted by $P_i(\theta)$) computed based on these parameters. As Snijders (2001) has shown, parametric person fit statistics are often based on the form

$$V(\theta) = \sum_{i=1}^{L} [x_i - P_i(\theta)]\, v_i(\theta)\,. \tag{1.18}$$

where $v_i(\theta)$ is a weight specific for the given statistic. The statistics depends on $P_i(\theta)$ which needs to be computed before $V(\theta)$ can be determined; $P_i(\theta)$ depends on the estimated item parameters and the estimated ability. In general, person fit statistics implemented in conventional software packages most often belong to the group of parametric person fit statistics (see Table 1.2). The evaluation of parametric statistics is not clear without ambiguity: While Karabatsos (2003) found parametric statistics to underperform compared to group-based nonparametric statistics based on the Rasch model (see above), Meijer (2003) found them to be more efficient in a real data example than nonparametric statistics.

Table 1.2

Typical IRT software packages with implemented person fit statistics

software	class	statistic	scoring method
ConQuest 3.0 (Adams, Wu, & Wilson, 2010)	parametric	*infit / outfit*	weighted likelihood estimation
eRm (R-package; Mair, Hatzinger, & Maier, 2012)	parametric	*infit / outfit*	maximum likelihood (with spline interpolation for non-observed and 0/full responses)
ltm (R-package; Rizopoulos, 2006)	parametric	l_z (Drasgow, Levine, & Williams, 1985)	empirical Bayes estimates
RSP (Glas & Ellis, 1993)	parametric	*M* (Molenaar & Hijtink, 1990)	-
Winmira (von Davier, 1997)	parametric	extension of *M* (von Davier & Molenaar, 2003)	-
Winsteps (Linacre, 2012)	parametric	*infit / outfit*	(adjusted) joint maximum likelihood

Note. Ability estimates may be substituted by applying the basic symmetric functions of the Rasch model in parametric statistics; this is implemented in RSP and Winmira.

Much of the research on parametric person fit statistics has focused on the problem of biased latent ability estimation (especially for short tests) and its impact on the statistics (see Chapter 3 and 4). As each of these statistics depend in some form on the latent ability while true abilities are unknown, an ability estimate has to be computed before it can be "plugged in" to compute the statistic. Research has shown that due to biased ability estimates and improper standardization, "standardized" parametric statistics might not reproduce the nominal α-level in a correct way (e.g., Li & Olejnik, 1997; Nering, 1995; Reise, 1995; Reise & Due, 1991; van Krimpen-Stoop & Meijer, 1999). Several approaches were developed to compensate this effect quite successfully (de la Torre & Deng, 2008; Dimitrov & Smith, 2006; Molenaar & Hoijtink, 1990; Snijders, 2001). This issue is addressed in Chapter 3 based on the Rasch model and in Chapter 4 with focus on IRT models in general.

Other types of person fit statistics: Beside these two main groups of person fit statistics, several variants exist. For example, item-group statistics are a special case of statistics determined in item subsamples (e.g., Drasgow, Levine & McLaughlin, 1991; Smith, 1986). Uniformly most powerful person fit tests defined by Klauer (1991, 1995) represent another subclass of person fit statistics designed to test individual conformity to the estimated (Rasch) model against an alternative generalization of this model. Also cumulative sum statistics (e.g., Tendeiro & Meijer, 2012) and local misfit analysis with the person response function (e.g., Conijn, Emons, van Assen, & Sijtsma, 2011; Emons, Sijtsma, & Meijer, 2004; Reise, 2000; Sijtsma & Meijer, 2001; Woods, 2008a) have aroused growing interest in the last years. Furthermore, there is a substantial number of studies that analyzed person fit in more complex models. This includes person fit analysis for polytomous items (e.g., Conijn, Emons, & Sijtsma, 2014; Drasgow, Levine, & Williams, 1985), continuous responses (e.g., Ferrando, 2010), mixture IRT models (von Davier & Molenaar, 2003), latent class models (Emons, Glas, Meijer, & Sijtsma, 2003), multiscale or multidimensional data (e.g., Conijn, Emons, &

Sijtsma, 2014; Drasgow, Levine, & McLaughlin, 1991), computer adaptive testing (e.g., McLeod & Lewis, 1999; Nering, 1997; van Krimpen-Stoop & Meijer, 1999) and cognitive diagnosis models (Cui & Leighton, 2009; Liu, Douglas, & Henson, 2009). Additionally, differential person functioning is a concept related to person fit. The term "differential person functioning" has been suggested by Johanson and Alsmadi (2002) analogously to the well-established principle of differential item functioning (e.g., Osterlind & Everson, 2009) and is related to the person response function, but not intended as a measure of fit (Johanson & Alsmadi, 2002). Answer copying indices like those presented by, for example, Sotaridona, van der Linden, and Meijer (2006) or van der Linden and Sotaridona (2004) are conceptually different from person fit tests but likewise constructed to detect a prevalent type of aberrant response behavior in achievement testing and are partly embedded in IRT (Wollack, 1997; Wollack & Cohen, 1998).

The comparison of several person fit statistics has not only demonstrated substantial differences in the statistical power between the different statistics (e.g., Karabatsos, 2003) but has also uncovered differences between the same statistics depending on test characteristics (like test length, percentage of misfit etc.; see Chapter 2), the underlying estimation procedure (e.g., de la Torre & Deng, 2008; Meijer & Nering, 1997; Reise, 1995) and the underlying method to determine cut values (e.g., by evaluating *p*-values or significance probabilities; de la Torre & Deng, 2008; van Krimpen-Stoop & Meijer, 1999). Given the potential consequences of inaccurately estimated abilities or impairment of the test validity outlined in detail in this chapter, the usage of the most powerful variant of each statistic to identify misfit is desirable. This is one of the main research aims in person fit research in general and in this script in particular: How can person fit analysis be enhanced with regard to statistical power while preventing inflation of Type I error rates at the same time?

1.3 References

Adams, R., Wu, M., & Wilson, M. (2012). Conquest 3.0. [computer program]. Melbourne: Australian Council for Educational Research.

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

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*, 7–16.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis, 13*, 171–187.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel-Dekker.

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item-response model*. Princeton, NJ: Educational Testing Services.

Birnbaum, A. (1968). Some latent trait models and their uses in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

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

Conijn, J. M., Emons W. H. M., & Sijtsma, K. (2014). Statistic l_z-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement, 38*, 122–136.

Conijn, J. M., Emons W. H. M., van Assen, M. A. L. M., & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research, 46*, 365–388.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person-fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 429–449.

De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159–177.

de Leeuw, J., & Verhelst, N. (1986). Maximum likelihood estimation in generalized Rasch models. *Journal of Educational Statistics, 11*, 183–196.

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person fit statistics. *Journal of Applied Measurement, 7*, 170–183.

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.

Duncan, K., & MacEachern, S. (2008). Nonparametric Bayesian modeling for item response. *Statistical Modelling, 8*, 41–66.

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

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*, 224–247.

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from vectors of summed polytomous item. *Applied Psychological Measurement, 33*, 599–619.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement, 26*, 88–108.

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement*, *27*, 459–478.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39*, 1–35.

Ferrando, P. J. (2010). Some statistics for assessing person-fit based on continuous-response models. *Applied Psychological Measurement, 34*, 219–237.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Fox, J.-P., & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*, 269–286.

Glas, C.A.W. & Ellis, J.L. (1993). *RSP, Rasch scaling program, computer program and user's manual*. Groningen: ProGAMMA.

Gustafsson, J. E. (1980). A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous items. *Educational and Psychological Measurement, 40*, 377–385.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Hartig, J., & Höhler, J. (2008). Representation of competencies in multidimensional IRT models with within-item and between-item multidimensionality. *Journal of Psychology, 216*, 89–101.

Harwell, M. R., Baker, F. B., & Zwarts, M. (1988). Item parameter estimation via marginal maximum likelihood and EM algorithm: A didactic. *Journal of Educational and Behavioral Statistics, 13*, 243–271.

Hendrawan, I., Glas, C. A.W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement, 29*, 26–44.

Höhler, J., Hartig, J. & Goldhammer, F. (2010). Modeling the multidimensional structure of students' foreign language competence within and between classrooms. *Psychological Test and Assessment Modeling, 52*, 323–340.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models. Foundations, recent developments and applications* (pp. 53–68). New York, NY: Springer.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester: Wiley.

Johanson, G., & Alsmadi, A. (2002). Differential person functioning. *Educational and Psychological Measurement, 62*, 435–443.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement, 38*, 79–93.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298.

Karabatsos, G., & Walker, S. (2009). Coherent psychometric modelling with Bayesian nonparametrics. *British Journal of Mathematical and Statistical Psychology, 62*, 1–20.

Klauer, K. C. (1991). An exact and optimal standardized person test for assessing consistency with the Rasch model. *Psychometrika, 56*, 213–228.

Klauer, K. C. (1995). The assessment of person fit. In G. H. Fischer &. I. W. Molenaar (Eds.), *Rasch models, foundations, recent developments, and applications* (pp. 97–110). New York, NY: Springer-Verlag.

Leutner, D., Fleischer, J., Spoden, C., & Wirth, J. (2007). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik [State-wide standardized assessments of learning between educational monitoring and individual diagnostics]. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8*, 149–167.

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

Linacre, J. M. (2012). *Winsteps® Rasch measurement computer program user's guide*. Beaverton, OR: Winsteps.com.

Liou, M. (1994). More on the computation of higher-order derivatives of the elementary symmetric functions in the Rasch model. *Applied Psychological Measurement, 18,* 53–62.

Liu, Y., Douglas, J. A., & Henson, R. A. (2009). Testing person fit in cognitive diagnosis. *Applied Psychological Measurement, 33*, 579–598.

Mair, P., & Hatzinger, R. (2007). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26–43.

Mair, P., Hatzinger, R., & Maier, M. J. (2012). *Package 'eRm'. Reference Manual*. (Ver. 0.15-1). Retrieved September 2012 from http://cran.r-project.org/web/packages/eRm/eRm.pdf

McLeod, L. D., & Lewis, C. (1999). Detecting item memorization in the CAT environment. *Applied Psychological Measurement, 23*, 147–160.

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

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3–8.

Meijer, R. R. (1997). Person-fit and criterion-related validity: An extension of the Schmitt, Cortina and Whitney study. *Applied Psychological Measurement, 21*, 99–113.

Meijer, R. R. (1998). Consistency of test behavior and individual difference in precision of prediction. *Journal of Occupational and Organizational Psychology, 71*, 147–160.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using IRT based person-fit statistics. *Psychological Methods, 8*, 72–87.

Meijer, R. R., & Nering, M. L. (1997). Ability estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321–336.

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

Meyn, S. P., & Tweedie, R. L. (2009). *Markov chains and stochastic stability* (2nd ed.). Cambridge: Cambridge University Press.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*, 725–737.

Miyazaki, K., & Hoshino, T. (2009). A Bayesian semiparametric item response model with Dirichlet process priors. *Psychometrika, 74,* 375–393.

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

Molenaar, I. W. (1995). Some background for item response theory and the Rasch model. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 3–14). New York, NY: Springer Verlag.

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

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

Nering, M. L. (1997). The distribution of indexes of person fit within the computerized adaptive testing environment. *Applied Psychological Measurement, 21*, 115–127.

OECD (2010). PISA 2009 results: What students know and can do – Student performance in reading, mathematics and science (Volume I). Retrieved from http://dx.doi.org/10.1787/9789264091450-en

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning*. Thousand Oaks, CA: Sage Publishing.

Partchev, I. (2004). *A visual guide to item response theory*. Retrieved from http://metheval.uni-jena.de/irt/VisualIRT.pdf

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

Pellegrino, J. W. (Ed.). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress*. Washington, DC: National Academy Press.

Pfanzagl, J. (1994). On item parameter estimation in certain latent trait models. In G. Fischer, & D. Laming (Eds.), *Contributions to mathematical psychology, psychometrics, and methodology* (pp. 249–263). New York, NY: Springer.

Raju, N., Pellegrino, J. W., & Bertental, M. W. (Eds.). (2000). *Grading the nation's report card: Research from the evaluation of NAEP*. Washington, DC: National Academy Press.

Ramsay, J. O. (1991). Kernel smoothing approaches to nonparametric item characteristic curve estimation. *Psychometrika, 56*, 611–630.

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

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data dit in IRT. *Applied Psychological Maesurement, 14*, 127–137.

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

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*, 543–570.

Reise, S. P., & Due, A. M. (1991). Test characteristics and their influence on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217–226.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Rost, J. (1999). Was ist aus dem Rasch-Modell geworden? [Where has the Rasch model gone?]. *Psychologische Rundschau, 50*, 140–156.

San Martín, E., Jara, A., Rolin, J.-M., & Mouchart, M. (2011). On the bayesian nonparametric generalization of IRT-type models. *Psychometrika, 76*, 385–409.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23,* 41–53.

Schmitt, N. S., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17*, 143–150.

Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics, 36*, 720–735.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika*, *66*, 191–208.

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

Smith, R. M. (1986). Person fit in the Rasch model. *Educational and Psychological Measurement, 46*, 359–370.

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

Sotaridona, L., van der Linden, W. J., & Meijer, R. R. (2006). Detecting answer copying using the Kappa statistic. *Applied Psychological Measurement, 30*, 412–431.

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

Tanner, M. (1994). *Tools for statistical inference*. New York, NY: Springer-Verlag.

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement, 36*, 420–442.

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

Tobias, S. (1992). The impact of test anxiety on cognition in school learning. In K. A. Hagtvet, & T. Baker Johnson (Eds.), *Advances in test anxiety research. Vol. 7* (pp. 18–31). Amsterdam/Lisse: Swets & Zeitlinger.

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

Verhelst, N. D., & Glas, C. A. W. (1995). The one parameter logistic model. In G. H. Fischer, & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 215–238). New York, NY: Springer Verlag.

van der Linden, W. J., & Sotaridona, L. S. (2004). A statistical test for detecting answer copying on multiple-choice tests. *Journal of Educational Measurement, 41*, 361–377.

von Davier, M. (1997). WINMIRA - program description and recent enhancements. *Methods of Psychological Research – Online, 2 (2)*, 25–28.

von Davier, M., & Molenaar, I. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68,* 213–228.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Wollack, J.A. (1997). A nominal response model approach to detect answer copying. *Applied Psychological Measurement, 21*, 307–320.

Wollack, J. A., & Cohen, A. S. (1998). Detection of answer copying with unknown item and trait parameters. *Applied Psychological Measurement, 22*, 144–152.

Woods, C. M. (2006). Ramsay-curve item response theory (RC-IRT) to detect and correct for nonnormal latent variables. *Psychological Methods, 11*, 253–270.

Woods, C. M. (2008a). Monte-Carlo evaluation of two-level logistic regression for assessing person fit. *Multivariate Behavioral Research, 43*, 50–76.

Woods, C. M. (2008b). Ramsay-curve item response theory for the three-parameter logistic item response model. *Applied Psychological Measurement, 32,* 447–465.

Woods, C. M., & Thissen, D. (2006). Item response theory with estimation of the latent populational distribution using spline-based densities. *Psychometrika, 71*, 281–301.

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

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

2 RESEARCH AIM, METHODOLOGY AND OVERVIEW OF STUDIES

In the previous chapter, person fit analysis has been introduced as a psychometric tool to identify respondents for which the applied IRT model may fail to give accurate information on the latent ability. Summarizing several results on misfitting response vectors, a decrease in accuracy of the latent ability estimation for those vectors (Meijer & Nering, 1997), a decrease in the number of correct diagnostic classification (mastery) decisions (Hendrawan, Glas, & Meijer, 2005), and a decrease in the validity of the test instrument (e.g., Schmitt, Chan, Sacco, McFarland, & Jennings, 1999) has to be expected if the IRT data set is not adjusted (e.g., by downweighting or excluding misfitting response vectors). Additionally, consequences resulting from inaccurate ability estimation for misfitting response vectors have been summed up. Given these consequences of inaccurate representation of abilities, the detection of those individuals with inappropriate item responses by means of person fit statistics is a reasonable way to improve the assessment process.

2.1 Challenges for person fit analysis and research aim

Difficulties in the interpretation of the results from person fit statistics are often related to problems with the standardization of these statistics. Standardized or normalized person fit statistics are usually preferred to unstandardized statistics as these allow diagnostic decisions to be made under a given Type I error rate (false alarm rate) and allow comparisons of fit across different latent abilities even when the original statistics are not independent from the ability level (e.g., Tatsuoka, 1984). But as indicated in Chapter 1, statistics standardized by normalization formulas are often inadequate for nonparametric (Emons, Meijer, & Sijtsma, 2002) as well as for parametric person fit statistics (e.g., Li & Olejnik, 1997; Molenaar & Hoijtink, 1990; Nering, 1995; van Krimpen-Stoop & Meijer, 1999). Reise (1995) found the power of the standardized statistic l_z to be influenced by the estimation method: Results were best for the true ability levels, but person fit analysis based on the robust scoring method BS (Mislevy & Bock, 1982) outperformed person fit analysis based on the ML and EAP scoring method (see Chapter 1). Differences between the scoring methods were also related to the characteristics of the test information function, the ability level, and the amount of misfit (percentage of misfitting responses). Results also uncovered an underestimated variance of the statistic for extreme abilities. Meijer and Nering (1997) replicated these results and found the advantage of BS to be related to lower bias of robust scoring methods in the presence of aberrancy (e.g., Meijer & Nering, 1997; see also Schuster & Yuan, 2011). Similar results were also found by Nering (1995) under the 2PLM and 3PLM for statistic l_z. Li and Olejnik (1997) analyzed the distributions of five standardized parametric person fit statistics for the Rasch model and found significant deviations from the standard normal distribution with each of these five statistics. Snijders (2001; see Chapter 4) proposed a method which corrects the normalization when true abilities are replaced by ability estimates. Béland, Magis, Raîche, and Talbot (2010) as well as Magis, Raîche, and Béland (2012) illustrated Snijders' method to improve the approximation of standardized mean square statistics (infit and outfit statistics). Molenaar and Hoijtink (1990), Tarnai and Rost (1990), as well as Dimitrov and Smith (2006) found the statistical power of parametric person fit statistics to be enhanced if the ability estimate was eliminated by making use of the basic symmetric functions of the Rasch model (see Chapter 1; see also Baker & Kim, 2004).

Another option to correct inadequate standardizations and to facilitate the interpretation of person fit statistics by reporting *p*-values (significance probabilities) to practitioners is given by computer-intensive Monte Carlo (MC) simulation methods (de la Torre & Deng, 2008; Rizopoulos, 2013; Seo & Weiss, 2013; van Krimpen-Stoop & Meijer, 1999). A reference null distribution is generated by simulating and analyzing new response vectors under the given IRT parameters (details described in in Chapters 3 and 4). Simulation-based methods have been established as the method of choice to facilitate the interpretation of person fit statistics by significance probabilities; recently published research articles on person fit analysis completely relied on MC simulation to obtain critical values for the statistics (Conijn, Emons, & Sijtsma, 2014; Seo & Weiss, 2013). Referring to standardized person fit statistics, de la Torre and Deng (2008, p. 176) argued that "…one can no longer justify the use of coarse approximation in performing PFA [person fit analysis]." and that „…the use of computer-intensive methods…should be given further attention because of the viability and promises of such approaches."

The general research aim of the following studies was to enhance the statistical power of person fit analysis (while at the same time controlling for Type I error level inflation) by simulation-based methods. In Study 1, an alternative method for person fit analysis in the Rasch model is proposed and evaluated based on simulating new data matrices with given marginals. In Study 2, parametric bootstrapping for person fit analysis with different underlying scoring methods is presented. Recent developments in psychometric methods as well as their user-friendly implementation in open-source software like R (R Core Development, 2013) facilitate the availability and the analysis of the methods applied in the following simulation studies. The methodological approach to evaluate the usefulness of the methods is stochastic simulation.

2.2 Evaluating Type I error and statistical power of person fit statistics

Given the underlying stochastic nature of IRT models, it is impossible to determine which response vectors classified as misfitting in real data were truly subject to aberrant response behavior. Type I error and statistical power of person fit statistics are therefore investigated in simulated data (e.g., Rupp, 2013). For this purpose, data is generated under a specified psychometric model and the Type I error can be found by analyzing the proportion of respondents classified as misfitting given an a priori defined cut value. For normalized statistics this is the z-score for the most extreme 5 %, z ≈ -1.65 (or the respective z-score for other nominal Type I error rates; e.g., Reise, 1995). For non-standardized statistics this cut score may be found by rule-of-thumb values defined in previous analyses (e.g., Karabatsos, 2003), by outlier analysis (Zijlstra, van der Ark, & Sijtsma, 2007) or – and of importance for the studies presented here – by simulation of the null distribution (de la Torre & Deng, 2008; Rizopoulos, 2013; van Krimpen-Stoop & Meijer, 1999). To investigate the statistical power of a person fit statistic, several types of model violations are imputed into model-conform response data. Again, classifications using the values of a person fit statistic and the respective cut score are compared to the known true classification (e.g., Reise, 1995, pp. 220-221). As an alternative, receiver operating curve analyses (e.g., Karabatsos, 2003) can be applied to analyze the accuracy of fit / misfit classifications and to optimize statistical power and Type I error computation.

Relying on a suggestion by Levine and Rubin (1979) and Drasgow (1982), a common distinction is often made between spuriously low and spuriously high scores. Aberrancy will be labeled as spuriously low if incorrect responses are given to easy items by high ability respondents. Conversely, spuriously high scores will be found if correct responses are given to difficult items by low ability respondents. As an example for spuriously high scores, cheating on difficult items of a test has been simulated in several studies (de la Torre & Deng, 2008; Dimitrov & Smith, 2006; Emons, Sijtsma, & Meijer, 2004) by fixing the probability of a correct response on these items to $P_i = 0.90$ or $P_i = 1.00$ for low ability respondents.

Note that nominal Type I error rates are arbitrarily defined and that researchers may look differently at power and Type I error depending on the research question. Emons, Sijtsma, and Meijer (2004) argued that – as long as power still exists – conservative person fit test might be useful to identify the most serious model violations under a certain degree of confidence. According to Meijer (2003, p. 81), incorrectly flagging a respondent as aberrant only implies a more thorough investigation of response vectors and has usually no serious consequences while ignoring misfit might cause biased latent ability estimates and incorrect mastery classifications (Hendrawan et al., 2005; Meijer & Nering, 1997). Additionally, statistical power of person fit statistics was often found to be critically low for short to medium test lengths (e.g., Emons, Glas, Meijer, & Sijtsma, 2003, p. 476). Considering the serious consequences of inaccurate ability estimation for misfitting respondents and reports on low power for short to medium test lengths, conservative test statistics do not seem useful for person fit analysis.

Though Type I error recovery and statistical power are the most relevant criteria for selecting an appropriate person fit statistic, some minor criteria have an impact on the choice of a person fit statistics. In first place, availability of the statistics in typical software packages (see Chapter 1) may determine which person fit statistic is applied in practice. In the studies presented below focus was laid on the recovery of nominal Type I error rates and the statistical power influenced by several psychometric characteristics.

2.3 Psychometric characteristics influencing the statistical power of person fit statistics

Previous studies found the power of person fit statistics to be influenced by several characteristics (see the summary by Meijer & Sijtsma, 2001; see also St-Onge, Valois, Abdous, & Germain, 2009). Some of these characteristics are usually varied in simulation studies to investigate interactions of these design factors with Type I errors and statistical power. In the two simulation studies presented in the next chapters, decisions were made regarding test characteristics (test length, item discrimination, spread of item difficulties), ability levels and the model violations simulated to investigate statistical power (type of model violations, percentage of misfitting response vectors). A short summary of previous findings on these characteristics in the context of person fit research is given below to justify the choice of settings and design factors of these studies.

Test characteristics: As demonstrated in several studies (e.g., Cui & Leighton, 2009; Emons, Glas, Meijer, & Sijtsma, 2003; Emons, Sijtsma, & Meijer, 2004; Karabatsos, 2003; Li & Olejnik, 1997; Meijer, 1994, 1996; Nering & Meijer, 1998; Reise & Due, 1991; Rogers & Hattie, 1987; Rudner, 1983), the statistical power of person fit statistics grows with an increasing number of items. Emons et al. (2003, p. 476) argued that lack of detection accuracy of person fit statistics was mainly related to low item numbers in realistic settings. According to Meijer, Molenaar, and Sijtsma (1994) person fit analysis may also be useful for short tests with “sufficiently reliable” items (p. 111). They found the statistical power to be strongly influenced by the (mean) item discrimination when analyzing a nonparametric person fit statistic (*U3*). The authors argued that “…although it is not desirable to use short tests for person-fit analysis, the use of highly reliable (highly discriminating) items may yield a detection rate that is approximately the same as for longer tests with weakly discriminating items” (Meijer et al., 1994, p. 119). Cui and Leighton (2009) discussed that ”…the best detection rates were achieved when the [person fit statistic] HCI was applied to tests that consisted of a relatively large number of high discriminating items…” (p. 446). Emons eta al. (2003, p. 476) stated that weak (mean) discrimination was one of the main flaws in person fit research beside short test length. Beside others, Meijer, Molenaar, and Sijtsma (1994) also investigated higher statistical power to detect misfit for larger spreads of item difficulties. Reporting similar results, Reise and Due (1991) as well as Reise (1995) also showed that the spread of item difficulty relative to the ability determined the statistical power of person fit statistics.

Respondents' characteristics (ability levels, type and percentage of misfitting response vectors): Latent abilities of the respondents were usually generated by sampling from a standard normal distribution (see the summary by Rupp, 2013). Researchers interested in person fit for extreme ability levels often simulated response vectors in several intervals of the ability distribution with both medium and extreme levels included (e.g., de la Torre & Deng, 2008; Reise, 1995; Sijtsma & Meijer, 2001). This is an important variation given that standardized person fit statistics often do not recover Type I error levels in an accurate way, in particular for extreme abilities (e.g., Emons et al., 2002; Reise, 1995).

The several types of aberrancies that are simulated to analyze the statistical power of person fit statistics represent the many types of disturbances psychologists might think of when reflecting a typical testing situation. But there is no agreement in person fit research on which types of model violations should be simulated to study the power of person fit statistics (though some spuriously high and spuriously low scores are usually differentiated; see below; see also Rupp, 2013). Karabatsos (2003) found that the statistical power of person fit statistics differed under several types of misfit conditions, but there were few differences in the rank ordering of the statistics. In Karabatsos' (2003) study, smaller rates of misfitting respondents (5 %, 10 %, and 25 %) were equally likely to be detected, while misfit was hard to detect under a rate of 50 %. Others (e.g., Armstrong & Chi, 2009a, 2009b; Emons, 2009) found slight differences in statistical power, mostly a decrease in power with a growing percentage of aberrant response vectors.

These findings on item characteristics and model violations were considered to define settings and design factors in the simulation studies described in the following chapters.

2.4 Settings and design factors of the simulation studies

Simulation methods are computer intensive methods mainly applied to inspect the characteristics of statistical methods in relation to the “truth” (Burton, Altman, Royston, & Holder, 2006). They are designed to imitate reality and should either be based on real data or rely on what is typical for real data (Burton et al., 2006; see also Rupp, 2013). The design of person fit studies is certainly worth a thorough consideration which cannot be provided here; Rupp (2013) has given a critical comment on person fit research in the time period 2001-2010. Considering the effects presented above (Chapter 2.3), settings and design factors for the studies on person fit statistics in this script (Chapter 3 and 4) are justified as follows:

The number of items between 20 and 60 items represent low to moderate test lengths. For example, the item number of the state-wide administered large-scale assessments of competencies typically varied between 20 and 50 (Fleischer, personal communication). Rupp (2013) summarized most person fit studies to use item samples of lengths 20 to 60. The simulated item parameters and ability levels represent typical choices in previous studies (Rupp, 2013). Item difficulties with moderate ranges, for example [-2, 2] or [-2.75, 2.75], were selected. A Rasch model was underlying Study 1 (Chapter 3), two discrimination levels were analyzed in 2PLM data in Study 2 (Chapter 4). The underlying abilities of the respondents were generated following previous studies which relied on the standard normal distribution or relevant ability intervals. To investigate the statistical power in Study 1, two types of misfit scenarios for low achieving students were simulated following exactly Dimitrov and Smith (2006); the number of abilities affected by misfit (1-27th percentile of the standard normal distribution) is assumed to yield about maximum detection rates with few replications (see Chapter 2.3). In Simulation Study 2 (Chapter 4), ability intervals following previous studies by, for example, de la Torre and Deng (2008), Reise (1995) or Sijtsma and Meijer (2001), were generated to vary abilities from medium to extreme levels. In line with these studies, item response vectors representing both spuriously high and spuriously low scores as suggested by Levine and Rubin (1979) and Drasgow (1982) were simulated to investigate the statistical power. Two recognized person fit statistics were analyzed; statistic *U3* (van der Flier, 1980, 1982; e.g., see studies by Emons et al., 2002; Emons et al., 2004; Emons et al., 2005; Meijer et al., 1994) and statistic l_0 respectively l_z (Levine & Rubin, 1979; e.g., see studies by de la Torre & Deng, 2008; Nering, 1995; Reise, 1995; Seo & Weiss, 2013; van Krimpen-Stoop & Meijer, 1999). For example, Li and Olejnik (1997) found the standardized statistic l_z to be “as good or better than the alternatives considered” in their study (p. 228). When comparing two or more methods, the moderately independent simulation design outlined by Burton et al. (2006, pp. 4281/4282) was followed; thus, the same data sets were used to compare the methods. These conditions are referred to as factors varied within each replication (where parallel simulations are referred to as “replications“ in this script). This design mimics a matched-pair design where sampling variability within these factors is set to zero (Burton et al., 2006, pp. 4282). All simulations were run in the R programming language (R Development Core Team, 2013) using the R default (pseudo)random number generator, the “Mersenne-Twister”.

2.5 Application to real data sets

Several areas of application for IRT models and psychometric tools related to IRT have been described in the introduction to Chapter 1. However, the most typical fields of application are cognitive performance tests in psychological research and in particular intelligence tests (e.g., Schmiedek, 2005), and achievement tests in educational research and educational large-scale assessment programs (e.g., Leutner, Fleischer, Spoden, & Wirth, 2007; Adams, Wu, & Carstensen, 2007).

Intelligence is a main research area in psychology overlapping with neuroscience and genetics (e.g., Wilhelm & Engle, 2005). Intelligence scores belong to the most influential predictors for educational achievement (e.g., Rohde & Thompson, 2007), and are also a typical covariate in educational research including experimental designs (e.g., when instructional designs factors or training programs are evaluated) as well as large-scale assessments where these are applied as covariates in (latent) regression models to correctly attribute learning on the variables of interest (as an example from PISA see, e.g., Leutner, Fleischer, & Wirth, 2006). Thus, to further validate results from the simulations described in Study 1, the presented methods were applied to tests of Rasch-scalability of a recognized German intelligence test as a key area of application for IRT models in psychological research (Chapter 3.5).

Educational large-scale assessments have been established as a diagnostic instrument in the United States of America since several decades but also its impact on European education is currently growing alongside with the ongoing influence of accountability systems in education. In a recent summary concerning advantages of IRT modeling for competence assessment – a recent development in central European educational research (e.g., Hartig, Klieme, & Leutner, 2008) – Hartig and Frey (2013) argue for advantages of IRT models with respect to localizing item difficulties and individual abilities on a common scale (which is of advantage when competence levels are described or items are selected for CAT administration; Hartig, 2007; van der Linden & Pashley, 2000), the modeling of complex data structures like Multi-Matrix-Sampling (Frey, Hartig, & Rupp, 2009), the parameterization of item characteristics (as it is the case, for example, in the linear-logistic test model; Fischer, 1996), and the opportunity to take the testing context or different working strategies of the respondents into account by applying mixture distribution IRT models (e.g., Rost & von Davier, 1995). Given this attractiveness of IRT models for educational achievement tests, results from the simulations in Study 2 were complemented by the application of the presented methods to testing data from a state-wide administered educational large-scale assessment program administered in the German federal state of North Rhine-Westphalia (Chapter 4.6). This real data application contributes to a growing line of research focusing on applications of person fit analyses in educational large-scale assessments: De la Torre and Deng (2008) argued person fit analysis may complement operational procedures and quality control for high-stakes testing in the context of educational programs like K-12, Harnisch and Linn (1981) as well as Miller (1986) identified classes with a poor match between test content and instructional coverage by person fit statistics, Jacob and Levitt (2003a, 2003b) applied statistical indicators including an aggregated person fit statistic to identify and validate test score cheating by teachers, Brown and Villareal (2007) corrected aggregated score reporting by person fit analysis, and Spoden, Fleischer, and Leutner (2014) investigated person fit analysis to identify teacher rater bias by lack of conformity to a coding manual.

2.6 Overview of the next chapters

The focus of the following chapters lies on the comparative analysis of Type I error and statistical power by simulation-based methods of person fit analysis. Simulation is also applied as the primary research method as it allows comparing fit / misfit classifications based on person fit statistics with (simulated) true classifications (the two studies presented in the Chapter 3 and 4 therefore correspond to the design of a “two-stage” simulation study). Real data examples exemplify the presented approaches.

In Chapter 3, a recent statistical approach for the Rasch model (the Rasch Sampler) is adopted for person fit analysis. The usefulness of the approach with regard to Type I error recovery and statistical power is evaluated under nominal Type I error levels comparing this new approach to conventional methods. The new approach is based on the influential MCMC simulation methods in Bayesian statistics (e.g., Fox, 2010; Jackman, 2009).

In Chapter 4, MC simulation (parametric bootstrap) based on weighted ML estimators for parametric person fit analysis is examined. Previous studies found that parametric person fit statistics based on Mislevy and Bocks (1982) BS estimator outperformed statistics based on other scoring methods (Meijer & Nering, 1997; Reise, 1995) but a new approach by Schuster and Yuan (2011) offers an additional option for person fit statistics based on robust latent ability estimates. In contrast to previous studies relying on normalization formulas, MC simulation may help to recover Type I error rates correctly. Additionally, MC simulation based on the WL scoring method by Warm (1989) and designed parallel to the Bayesian method by de la Torre and Deng (2008) was evaluated.

As outlined above, real data examples are presented in both Chapter 3 and 4 to further illustrate differences in the methods described above with typical psychological or educational outcome variables.

Chapter 5 gives a discussion of findings, including contributions, limitations and practical implications.

2.7 References

Adams, R. J., Wu, M., & Carstensen, C. H. (2007). Application of Multivariate Rasch Models in International Large Scale Survey Assessments. In M. Von Davier & C. H. Carstensen (Hrsg.), *Multivariate and Mixture Distribution Rasch Models - Extensions and Applications* (pp. 271–280). New York, NY: Springer.

Armstrong, R. D., & Shi, M. (2009a). A parametric cumulative sum statistic for person fit. *Applied Psychological Measurement, 33*, 391–410.

Armstrong, R. D., & Shi, M. (2009b). Model-free CUSUM methods for person fit. *Journal of Educational Measurement, 46*, 408–428.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel-Dekker.

Béland, S., Magis, D., Raîche, G., & Talbot, N. (2010). *Three person-fit indexes with estimated ability level: A simulation study*. Paper presented at the International Objective Measurement Workshop (IOMW), Boulder, CO.

Brown, R. S., & Villareal, J. C. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing*, *7*, 1–25.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*, 4279–4292.

Conijn, J. M., Emons W. H. M., & Sijtsma, K. (2014). Statistic l_z-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement, 38*, 122–136.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person-fit for cognitive diagnostic assessment. *Journal of Educational Measurement*, *46*, 429–449.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159–177.

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person fit statistics. *Journal of Applied Measurement, 7*, 170–183.

Drasgow, F. (1982). Choice of test model for appropriateness measurement. *Applied Psychological Measurement*, *6*, 297–308.

Emons, W. H. M. (2009). Detection and diagnosis of person misfit from vectors of summed polytomous item. *Applied Psychological Measurement, 33*, 599–619.

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement, 27*, 459–478.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 personfit statistic. *Applied Psychological Measurement, 26*, 88–108.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39*, 1–35.

Fischer, G. H. (1996). Unidimensional linear logistic Rasch models. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 225–243). New York, NY: Springer.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Frey, A., Hartig, J., & Rupp, A. (2009). Booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28,* 39–53.

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.

Harnisch, D. L., & Linn, R. L. (1981). Analysis of item response patterns: Questionable test data and dissimilar curriculum practices. *Journal of Educational Measurement, 18*(3), 133–146.

Hartig, J. (2007). Skalierung und Definition von Kompetenzniveaus [Scaling and definition of proficiency levels]. In: Beck, Bärbel; Klieme, Eckhard (Hrsg.): Sprachliche Kompetenzen. Konzepte und Messung. DESI-Ergebnisse Band 1 (pp. 83-99). Weinheim: Beltz.

Hartig, J. & Frey, A. (2013). Sind Modelle der Item-Response-Theorie (IRT) das „Mittel der Wahl" für die Modellierung von Kompetenzen? [Benefits and limitations of modeling competencies by means of Item Response Theory (IRT)]. *Zeitschrift für Erziehungswissenschaft, Sonderheft 18,* 47–51.

Hartig, J., Klieme, E. & Leutner, D. (Eds.). (2008). *Assessment of competencies in educational contexts*. Göttingen: Hogrefe.

Hendrawan, I., Glas, C. A. W., & Meijer, R. R. (2005). The effect of person misfit on classification decisions. *Applied Psychological Measurement, 29,* 26–44.

Jackman, S. (2009). *Bayesian analysis for the social sciences*. Chichester, UK: John Wiley & Sons.

Jacob, B. A., & Levitt, S. D. (2003a). Rotten Apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics, 118*, 843–877.

Jacob, B. A., & Levitt S. D. (2003b). Catching cheating teachers: The results of an unusual experiment in implementing theory. In W. G. Gale & J. Rothenberg Pack (Eds.), *Brookings-Wharton Papers on Urban Affairs 2003* (S. 185–209). Washington, DC: Brookings Institution Press.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16,* 277–298.

Leutner, D., Fleischer, J., Spoden, C., & Wirth, J. (2007). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik [State-wide standardized assessments of learning between educational monitoring and individual diagnostics]. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8,* 149–167.

Leutner, D., Fleischer, J., & Wirth, J. (2006). Problemlösekompetenz als Prädiktor für zukünftige Kompetenz in Mathematik und in den Naturwissenschaften [Problem solving as a predictor for future competence in mathematics and science]. In PISA-Konsortium Deutschland (Eds.), *PISA 2003. Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (pp. 119–137). Münster: Waxmann.

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

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

Magis, D., Raîche, G., & Béland, S. (2012). On the accurate selection of asymptotic detection thresholds for infit and outfit indexes of person fit. *Paper presented at the International Objective Measurement Workshop*, Vancouver.

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3–8.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using IRT based person-fit statistics. *Psychological Methods, 8*, 72–87.

Meijer, R. R., Molenaar, L. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111–120.

Meijer, R. R., & Nering, M. L. (1997). Ability estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321–336.

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

Miller, M. D. (1986). Allocation and patterns of item response. *Journal of Educational Measurement, 23,* 147–156.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*, 725–737.

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

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

Nering, M. L., & Meijer, R. R. (1998). A comparison of the person response function and the lz person-fit statistic. *Applied Psychological Measurement, 22*, 53–69.

R Development Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

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

Reise, S. P., & Due, A. M. (1991). Test characteristics and their influence on the detection of aberrant response patterns. *Applied Psychological Measurement, 15*, 217–226.

Rizopoulos, D. (2013). *ltm - Latent trait models under IRT: Reference manual* (Ver. .9-7). Retrieved February 2012 from http://cran.r-project.org/web/packages/ltm/ltm.pdf

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

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*, 83–92.

Rost, J., & von Davier, M. (1995). Mixture distribution Rasch models. In G. Fischer & I. Molenaar (Hrsg.), *Rasch Models: Foundations, recent developments, and applications* (S. 257–268). New York, NY: Springer.

Rudner, L. M. (1983). Individual assessment accuracy. *Journal of Educational Measurement, 20*, 207–219.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Response Theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessement Modeling, 55*, 3–38.

Schmiedek, F. (2005). Item response theory and the measurement of cognitive processes. In O. Wilhelm & R. W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 265–277). Thousand Oaks, CA: Sage.

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement*, *23*, 41–53.

Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics, 36*, 720–735.

Seo, D. G., & Weiss, D. J. (2013). l_z person-fit index to identify students with achievement test data. *Educational and Psychological Measurement*, 73, 994–1016.

Sijtsma, K., & Meijer, R. R. (2001). The person response function as a tool in person-fit research. *Psychometrika, 66*, 191–208.

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

Spoden, C., Fleischer, J. & Leutner, D. (2014). Niedrige Testmodellpassung als Resultat mangelnder Auswertungsobjektivität bei der Kodierung landesweiter Vergleichsarbeiten durch Lehrkräfte [Low test model fit and teacher rater bias — results from a state-wide administered large-scale assessment of competencies]. *Journal für Mathematikdidaktik, 35(1),* 79-99.

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (2009). A Monte Carlo study of the effect of item characteristic curve estimation on the accuracy of three person-fit statistics. *Applied Psychological Measurement*, *33*, 307–324.

Tarnai, C., & Rost, J. (1990). *Identifying aberrant response patterns in the Rasch model: The Q index. Sozialwissenschaftliche Forschungsdokumentationen*. Münster: Institut für sozialwissenschaftliche Forschung e.V. .

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

van der Flier, H. (1980). *Vergelijbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets and Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267–298.

van der Linden, W. J., and Pashley, P. J. (2000). Item selection and ability estimlation in adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), Computerized adaptive testing. Theory and practice (pp. 1–25). Boston, MA: Kluwer.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Wilhelm, O., & Engle. R. (Eds.) (2005). *Understanding and measuring intelligence*. London: Sage.

Zijlstra, W. P., van der Ark, L. A., & Sijtsma, K. (2007). Outlier detection in test and questionnaire data. *Multivariate Behavioral Research, 42*, 531–555.

3 STUDY I - APPLYING THE RASCH SAMPLER FOR PERSON FIT ANALYSIS[1]

3.1 Background

The quality of individual response vectors is investigated in many fields of testing (e.g., Lamprianou, 2010). Person fit statistics (Meijer & Sijtsma, 2001) provide the test administrators with information about whether the individual response vector is likely under the test model applied. Unlikely or aberrant response vectors should be treated with caution as ability estimation may be invalid for those vectors (Meijer & Nering, 1997). For the Rasch model (Rasch, 1960) a systematic and extensive comparison of person fit statistics is given by Karabatsos (2003). Extending previous research (e.g., Li & Olejnik, 1997; Meijer, Molenaar, & Sijtsma, 1994; Reise & Due, 1991), Karabatsos (2003) found the statistical power of the person fit statistics to depend on the type of aberrancy and to rise with increasing test length and with decreasing percentage of aberrancy. Furthermore, his study reveals that four out of the five best performing person fit statistics for the Rasch model were nonparametric statistics. A parametric person fit statistic for the Rasch model depends on the predicted probability of an item response $P_i(\theta_\mathrm{p})$ given the common Rasch equation (see Chapter 1).

In contrast, nonparametric person fit statistics do not make use of the estimated item parameters and ability but compute the statistic from the response vectors and the marginals of the data matrix. Karabatsos argued that parametric statistics are "biased to be overoptimistic" (Karabatsos, 2003, p. 290) as the data set is used twice, once for the estimation of item parameters and abilities to predict the probability of item endorsement, and once again to measure the fit of the data to these predictions or to the estimated parameters –

[1] This chapter is based on:
Spoden, C., Fleischer, J. & Leutner, D. (2014). Applying the Rasch Sampler for person fit analysis under fixed nominal alpha level. *Journal of Applied Measurement, 15*, 276-291.

unless item parameters are known as it is the case in computer adaptive testing or with previously administered tests.

Further bias in person fit statistics comes from normalization attempts: For classifying an item response vector as misfitting independent from the (estimated) ability, the distribution of the statistic under the null hypothesis is needed. With both parametric and nonparametric person fit statistics the adequate approximation of a standard normal distribution was found to fail when normalization formulas were applied (Emons, Meijer, & Sijtsma, 2002; Molenaar & Hoijtink, 1990; Nering, 1995; van Krimpen-Stoop & Meijer, 1999). As a consequence, the empirical Type I error rate differs from its nominal rate and an inadequate number of response vectors is classified as misfitting.

Several authors have proposed possible adjustments to normalization formulas with emphasis on either item response models (e.g., de la Torre & Deng, 2008; Snijders, 2001) or the Rasch model (e.g., Bedrick, 1997). These adjustments were found to improve the approximation of the nominal α-level but are also restricted to certain (types of) person fit statistics (Bedrick, 1997; Snijders, 2001) or certain estimators (de la Torre & Deng, 2008). For the Rasch model it has also been recommended to replace the ability estimate by the use of basic symmetric functions (Dimitrov & Smith, 2006; Karabatsos, 2000; Molenaar & Hoijtink, 1990; Ponocny, 2000; Tarnai & Rost, 1990), which requires the item parameters and the score to be computed but not the latent ability (Baker & Kim, 2004). The exact probability of a response vector can then be enumerated for shorter test lengths under the assumption that the estimated item parameters conform the true item parameters (Molenaar & Hoijtink, 1990). Results from simulated data have shown that this approach improves the statistical power of the statistics (Dimitrov & Smith, 2006). For longer tests MC simulation has been suggested (Molenaar & Hoijtink, 1990; Ponocny, 2000), as the computation of the basic symmetric functions and the enumeration of the probability for each response vector under the given item parameters may include extensive computation (Dimitrov & Smith, 2006, p. 182; Molenaar & Hoijtink, 1990, p. 98).

MC simulation is a typical approach to solve complex numerical problems and can quite easily be implemented to create *p*-values for person fit statistics (e.g., Rizopoulos, 2013, pp. 39-42; see Chapter 4): In a MC simulation $m = 1,\ldots, M$ new response vectors are being generated under the estimated model parameters before the *p*-value of the initial response vector is estimated as the proportion of response vectors satisfying ($T_m \leq T_{obs}$) where T_m is the value of the person fit statistic for response vector *m* and T_{obs} is the value for the original data set. The usefulness of the approach depends on the statistical properties of the underlying (estimated) parameters. With the EAP estimator of the ability level, for example, de la Torre and Deng (2008) developed a shrinkage correction before adjusting the distribution of the person fit statistic by MC simulation. They found this approach to approximate the distribution of the person fit statistic more adequately than traditional methods.

A different simulation approach investigated person fit in a Bayesian framework. For example, Glas and Meijer (2003) used posterior predictive checks to investigate the power of several statistics obtained by MCMC (Gelman, Carlin, Stern, & Rubin, 2004; Jackman, 2009). With MCMC as a sequential method, a target posterior distribution is first approximated by a draw from the current state of a Markov chain and then corrected to improve the approximation at the next step of the iterative process. Glas and Meijer (2003) sampled the item and ability parameters from the current draw of the posterior distribution in a fixed interval of iterations of the Markov chain, generated new model-conform data under these parameters and estimated the Bayesian *p*-value of the statistic for the initial response vector from this reference data similar to what was described before. As an advantage to frequentist approaches, MCMC takes the uncertainty in the parameter estimation properly into account (Glas & Meijer, 2003). However, a statistical test to identify aberrant responding behavior under this Bayesian estimation was found to be conservative (de la Torre & Deng, 2008; Glas & Meijer, 2003), which is a well-known result with hypothesis tests based on posterior predictive checks (e.g., Bayarri & Castellanos, 2001).

To the knowledge of the authors of the present paper, another approach for simulating the distribution of Rasch fit statistics under constant marginals (as sufficient statistics for the item and ability parameters in the Rasch model) has been neglected for person fit so far: The Rasch Sampler (for a technical description see Verhelst, 2008; Verhelst, Hatzinger, & Mair, 2007, and the appendix) is a MCMC algorithm to sample binary data matrices with the same marginals by binomial transformations of the original data set. If the Rasch model holds, all binary data matrices with common marginals have the same probability. As Ponocny (2001) points out, Rasch (1960) was already aware of the possibility to create exact nonparametric tests for his model by enumerating these matrices, but complete enumeration was — and still is, even under enhanced computational support — impossible for reasonable matrix sizes. However, to approximate the null distribution of any fit statistic, a random sample from the collection of equally likely matrices can be simulated. Under this approach, several nonparametric fit statistics for the Rasch model that check for local dependence, differential item functioning or diverging item discriminations, among others, were developed by Ponocny (2001) and are currently implemented in the package eRm (Mair & Hatzinger, 2007b) for the statistical software R (R Development Core Team, 2011) using the Rasch Sampler.

The rationale for applying the Rasch Sampler for nonparametric tests works analogously for person fit analysis, except that each single response vector, as a partition of the data matrix, is independently analyzed. As the marginals are sufficient statistics for the estimated item parameters and the latent abilities in the Rasch model, a nonparametric simulation based on these constant marginals can be applied to build the statistics' distribution empirically. In case of person fit, complete data matrices are sampled by the Rasch Sampler, the probability of each response vector, conditional on the score *r* and the item marginals, are computed and the *p*-value of the observed response vectors is found as the proportion of response vectors in the newly generated matrices with a value of the person fit statistic smaller than or equal to the value of the statistic in the observed data (see above). For a nonparametric statistic neither item nor ability parameters have to be estimated; in case of a parametric statistic the parameters have to be estimated only once on the observed data set and can then be used to compute the statistic on the observed and any generated data matrix. The accuracy of this simulated reference distribution depends on the number of samples and therefore on the time spent on drawing new data matrices. Using a MC approach, de la Torre and Deng (2008) found convincing results for M = 1000 samples to approximate the distribution of a person fit statistic. The major advantages of the approach are that (1) the algorithm is already implemented with a user-friendly interface in R (R Development Core Team, 2011), that (2) in contrast to conventional approaches, the simulation is not based on estimated parameters, that (3) it can be applied to mostly any kind of parametric or nonparametric person fit statistics, and that (4) the same data sets generated in one simulation can be used to test person fit as well as further assumptions of the Rasch model. Of course, with regard to (4) adjustment of the Type I error rate for each single test is necessary (Kubinger & Draxler, 2007).

3.2 Purpose of this study

The usage of the Rasch Sampler for simulating reference data and generating the distribution of person fit statistics in the Rasch model is investigated. In two simulation studies, *p*-values of typical person fit statistics obtained when applying the Rasch Sampler are analyzed with regard to Type I error (false alarm rates) and statistical power (detection rates).

These quantities are compared to normalization formulas. In Simulation 1, Type I error rates are investigated for each score *r*, as results from several studies indicate that the distributional properties of the statistics were differentially affected for different levels of the latent ability (Emons et al., 2002; Nering, 1995; van Krimpen-Stoop & Meijer, 1999). In Simulation 2, both power and Type I error rates are investigated under different types of model violations in a design very similar to the one used by Dimitrov and Smith (2006). A real data example completes the analysis and illustrates the results from simulated data.

3.3 Simulation 1: Investigation of Type I error rates

To gain insights into the benefits of the Rasch Sampler approach, two person fit statistics (*U3* and l_0) with documented problems to approximate the nominal Type I error rate accurately by normalization formulas (NOR) were compared in a simulation with regard to Type I error rates. The nonparametric statistic *U3* (van der Flier, 1980; 1982) was chosen because it was found to be one of the five most powerful statistics for the Rasch model by Karabatsos (2003), and a normalization formula for *U3* was developed by van der Flier (1982). Its unstandardized formula is given by:

$$U3 = \frac{\log\left(\boldsymbol{X}_p^{max}\right) - \log\left(\boldsymbol{X}_p\right)}{\log\left(\boldsymbol{X}_p^{max}\right) - \log\left(\boldsymbol{X}_p^{min}\right)} \ , \qquad [3.1]$$

where $\boldsymbol{X}_p$ is the response vector of the original data set, $\boldsymbol{X}_p^{max}$ is the maximum possible value resulting from the Guttman vector for score *r* under test length *L* and $\boldsymbol{X}_p^{min}$ is the minimum value resulting from the anti-Guttman vector for score *r*. Van der Flier (1982) derived that:

$$U3^{\mathrm{NOR}} = ZU3 = \frac{U3 - E(U3)}{[\mathrm{Var}\ (U3)]^{1/2}} \ , \qquad [3.2]$$

was approximately normally distributed with expectation:

$$E(U3) = \frac{\sum_{i=1}^{r} \log \frac{P_i}{1-P_i} - \sum_{i=1}^{L} P_i \log \frac{P_i}{1-P_i} + \frac{\sum_{i=1}^{L} P_i\,(1-P_i) \log \frac{P_i}{1-P_i}}{\sum_{i=1}^{L} P_i\,(1-P_i)} \left(r - \sum_{i=1}^{L} P_i\right)}{\sum_{i=1}^{r} \log \frac{P_i}{1-P_i} - \sum_{i=L-r+1}^{L} \log \frac{P_i}{1-P_i}} \,, \qquad [3.3]$$

and variance given by:

$$\mathrm{Var}(U3) = \frac{\left\{ \sum_{i=1}^{L} P_i\,(1-P_i) \left(\log \frac{P_i}{1-P_i}\right)^2 - \frac{\left[\sum_{i=1}^{L} P_i\,(1-P_i) \log \frac{P_i}{1-P_i}\right]^2}{\sum_{i=1}^{L} P_i\,(1-P_i)} \right\}^{\frac{1}{2}}}{\left| \sum_{i=1}^{r} \log \frac{P_i}{1-P_i} - \sum_{i=L-r+1}^{L} \log \frac{P_i}{1-P_i} \right|} \,, \qquad [3.4]$$

The empirical distribution of $U3^{\mathrm{NOR}}$ was previously found to deviate from the standard normal distribution with larger differences for extreme values of *r* (Emons et al., 2002).

The statistic l_0 (Levine & Rubin, 1979) is the prominent parametric log-likelihood statistic given by

$$l_0 = \sum_{i=1}^{L} x_i \log[P_i(\theta)] + (1 - x_i) \log[1 - P_i(\theta)]. \qquad [3.5]$$

A normalization formula for *l* was developed by Drasgow, Levine and Williams (1985) as:

$$l^{\mathrm{NOR}} = l_z = \frac{l_0 - E(l_0)}{[\mathrm{Var}(l_0)]^{1/2}} \,, \qquad [3.6]$$

with

$$E(l_0) = \sum_{i=1}^{L} \{P_i(\theta) \log[P_i(\theta)] + [1 - P_i(\theta)] \log[1 - P_i(\theta)]\} \,, \qquad [3.7]$$

and

$$\mathrm{Var}(l_0) = \sum_{i=1}^{L} P_i(\theta)\,[1 - P_i(\theta)] \left[\log \frac{P_i(\theta)}{1 - P_i(\theta)}\right]^2. \qquad [3.8]$$

l^{NOR} was regarded to be a very promising person fit statistic for the Rasch model by Li and Olejnik (1997) but several studies have found its empirical distribution to deviate from the standard normal distribution, especially when estimated instead of true abilities are used (Nering, 1995; van Krimpen-Stoop & Meijer, 1999).

Type I error rates under the method NOR are used as benchmark to evaluate the new method, where the distribution of the statistic was built based on new reference data simulated by the MCMC approach implemented in the Rasch Sampler (MCMC (RS)). To prevent autocorrelations in the process of sequential draws and to make sure that the process has reached its uniform posterior target distribution, a number of initial iterations of the Markov chain known as burn-in as well as a fixed number of iterations (steps) after each simulation had been omitted before a new data matrix was generated (Verhelst et al., 2007). Under these assumptions, the simulated data matrices generated by the Rasch Sampler are drawn from a *truly* uniform distribution (Verhelst, 2008), and empirical Type I errors equal to the predefined level are expected. In this simulation, the Rasch Sampler was initiated with a burn-in of 1000 iterations and a step number of 25. A smaller number of burn-in iterations and a lower step number might be chosen to run the sampler, but we wanted to be sure of the process' convergence. For each observed data set, M = 1000 new data sets were generated to obtain p-values for the person fit statistic of the observed data. To compare the accuracy of the approach with regard to the nominal Type I error rate, the mean absolute difference (MAD), $\frac{1}{L-1}\sum_{r=1}^{L-1}|\widehat{\alpha}_{\mathrm{r}} - \alpha_{\mathrm{r}}|$, between empirical and nominal Type I error rate was computed.

With each of these two methods, item parameters b_i for the parametric statistic l_0 were estimated by conditional maximum likelihood estimation in eRm (Mair & Hatzinger, 2007a, 2007b). Ability estimates θ_{p} were obtained by empirical Bayes estimation (MAPs) from the R-package ltm (Rizopoulos, 2013).

3.3.1 Data simulation

To evaluate Type I error rates of the statistics under the Rasch model 100 replications of 50 response vectors at each level of r (except $r = 0$ and $r = L$ for which person fit cannot be evaluated meaningfully) were simulated for test lengths of L = 20, L = 40 and L = 60 items. Overall, the number of simulated response vectors was 95.000, 195.000 and 295.000. True item parameters were sampled equidistantly in the interval [−2, 2] for the 20- and in the interval [−2.75, 2.75] for the 40- and 60-items sets (Dimitrov & Smith, 2006). An R-Code (R Development Core Team, 2011) was written to generate the data, run the Rasch scaling in eRm and compute empirical *p*-values of the person fit statistics.

3.3.2 Evaluation of Type I error rate

The Type I error is estimated in data sets which contain no model violations as the percentage[2] of response vectors with a *p*-value of the person fit statistic smaller than the nominal α-level. Because low power of person fit statistics is expected under small Type I error rates, results are presented for α = .05 and α = .10.

3.3.3 Results

Figure 3.1 shows the empirical Type I error rates for statistics *U3* and l_0 under NOR and MCMC (RS) at three test lengths and two alpha levels. Results for NOR demonstrate a strong curvilinear dependency of the empirical Type I error rate on the score. With *U3* the empirical Type I error is deflated for average scores and strongly inflated for extreme scores. Note that for some extreme scores Type I error rates were not obtained due to about zero probability of this level of r. With l_0 the Type I error is quite adequate for average scores but

[2] Subsequently, results are expressed as decimals.

deflated for extreme scores. These results replicate previous findings for example by Emons et al. (2002), Nering (1995), or Reise (1995). The MCMC (RS) approach adequately reproduced the true Type I error rates of both statistics for all values of *r* and each of the three selected test lengths. Differences between nominal and empirical α-level were generally very small and very much the same for both statistics. Well-adjusted Type I error rates for MCMC (RS) and considerable differences between nominal and empirical Type I error rates found for NOR are also reflected by the MAD shown in Table 3.1.

Table 3.1
Mean absolute difference (MAD) between empirical and nominal Type I error for two approaches to generate p-values: Normalization formula (NOR) and Markov chain Monte Carlo simulation of the Rasch Sampler (MCMC (RS)).

		U3		l_0	
α	*L*	NOR	MCMC (RS)	NOR	MCMC (RS)
.05	20	.056	.004	.026	.004
	40	.093	.002	.025	.002
	60	.103	.002	.026	.002
.10	20	.127	.004	.055	.004
	40	.162	.003	.053	.003
	60	.168	.003	.053	.003

As a conclusion from Simulation 1, we may state that the estimated *p*-values based on the MCMC (RS) approach are in accordance with the nominal α-level at all test lengths and for all scores, while non-ignorable deviations from the expected Type I error rates were found for NOR, especially for extreme scores.

3.4 Simulation 2: Investigation of statistical power to detect model violations

In Simulation 2, the selected person fit statistics, the estimation methods and the initiation parameters for the Rasch Sampler were the same as in Simulation 1.

3.4.1 Data simulation

A 3 × 2 × 2 design was used with three test lengths (20, 40 and 60 items), two types of aberrant response behavior (guessing and cheating) and two types of aberrancy levels (20 % and 40 %). 36 replications of $N = 1000$ were simulated under the Rasch model. Again, true item parameters were sampled equidistantly in the interval [−2, 2] for the 20-items sets and in the interval [−2.75, 2.75] for the 40- and 60-items sets. In this study ability was sampled from N (0, 1), which is the common approach in simulation studies (e.g., Dimitrov & Smith, 2006; Glas & Meijer, 2003; Li & Olejnik, 1997; Meijer & Nering, 1997; van Krimpen-Stoop & Meijer, 1999). Aberrant response vectors were imputed for low ability respondents with $\theta <$ 0.61 (27th percentile) as described in Dimitrov and Smith (2006): Guessing was simulated by assigning a probability of .25 for a correct response on either 20 % or 40 % of the most difficult items for low ability respondents, simulating multiple-choice items with three wrong options and one correct option. Cheating was simulated by assigning a probability of .90 for a correct response on either 20 % or 40 % of the most difficult items for low ability respondents, as 100 % successful cheating might be unlikely under real testing conditions (Dimitrov & Smith, 2006).

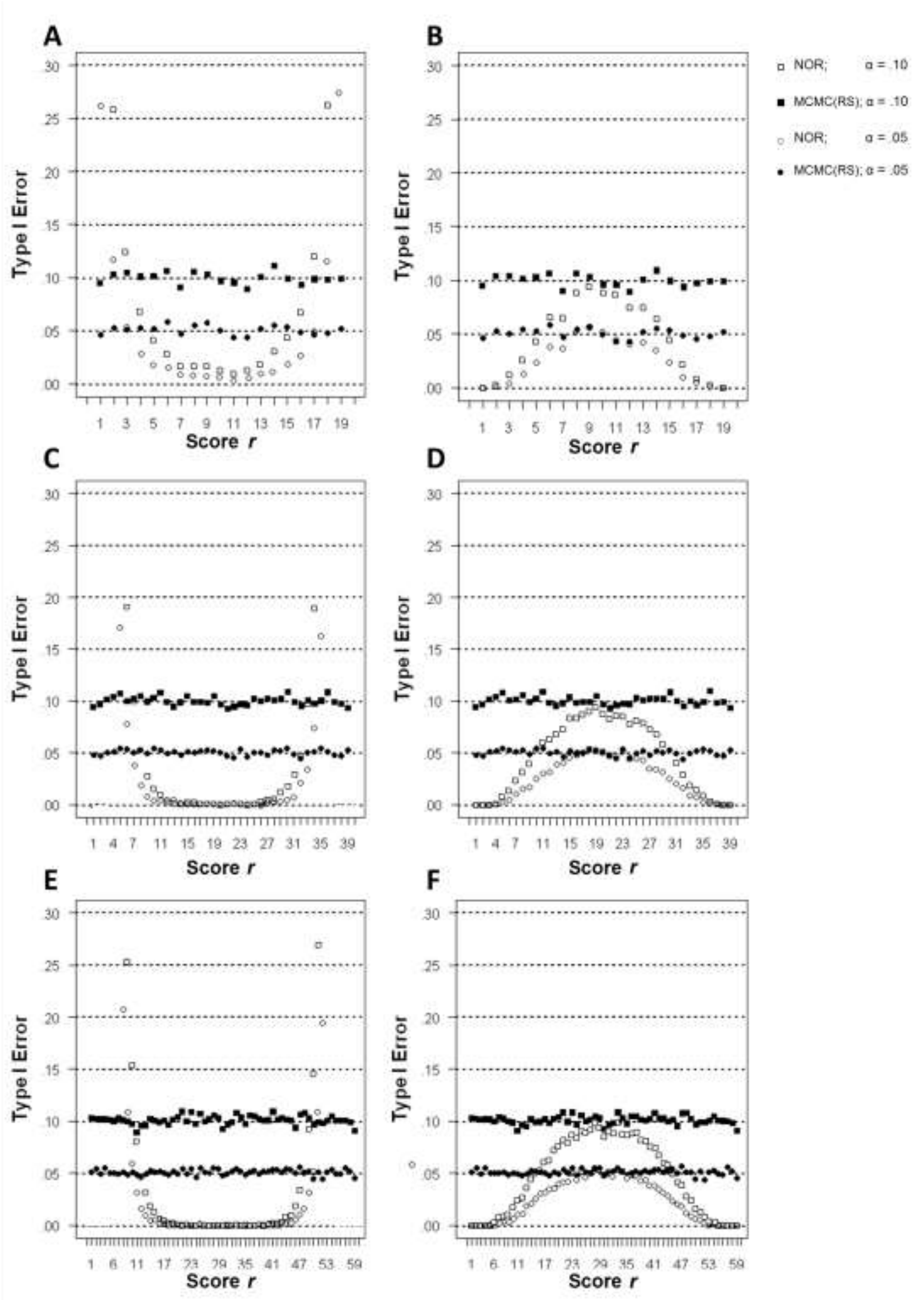

Figure 3.1. **Empirical Type I error rates of two person fit statistics and two approaches to generate *p*-values (normalization formula, NOR; Markov chain Monte Carlo simulation of the Rasch Sampler, MCMC (RS)). A: statistic *U3*, 20 items; B: statistic l_0, 20 items; C: statistic *U3*, 40 items; D: statistic l_0, 40 items; E: statistic *U3*, 60 items; F: statistic l_0, 60 items.**

3.4.2 Evaluation of statistical power and Type I error rate

Power and Type I error rate are evaluated under a fixed nominal α-level. Power is estimated in the *guessing* and the *cheating* conditions as the percentage of aberrant response vectors with a *p*-value of the person fit statistic smaller than α. The Type I error rate is estimated as the percentage of non-aberrant response vectors with a *p*-value of the person fit statistic smaller than α. The same α-levels as in Simulation 1 were used (α = .05 and α = .10).

3.4.3 Results

Power rates of the statistics are presented in Figures 3.2 and 3.3. For both cheating and guessing power increase with increasing item number, percentage of aberrancy and α-level. Cheating was generally easier to detect than guessing. In the cheating conditions, power is in most conditions highest for MCMC (RS), except for 20 items and an aberrancy rate of 20 %, where higher rates are found for *U3* under NOR. With the parametric statistic l_0, advantages of MCMC (RS) are stronger than for the nonparametric statistic *U3*. Differences between both methods also grow with decreasing item number. Over all test lengths, power rates for *U3* and l_0 are very similar under MCMC (RS), while under NOR *U3* outperforms l_0. These differences reflect inflated Type I error rates of $U3^{NOR}$ and deflated Type I error rates of l^{NOR} (Emons et al., 2002). The highest power in all cheating conditions is found for both *U3* and l_0 and in the condition of 60 items, an aberrancy rate of 40 % and α = .10, where the percentage of correctly detected vectors is near 100 % with MCMC (RS). The lowest rates are found for 20 items, an aberrancy rate of 20 % and α = .05, where rates below .35 indicate that model violations are generally hard to detect.

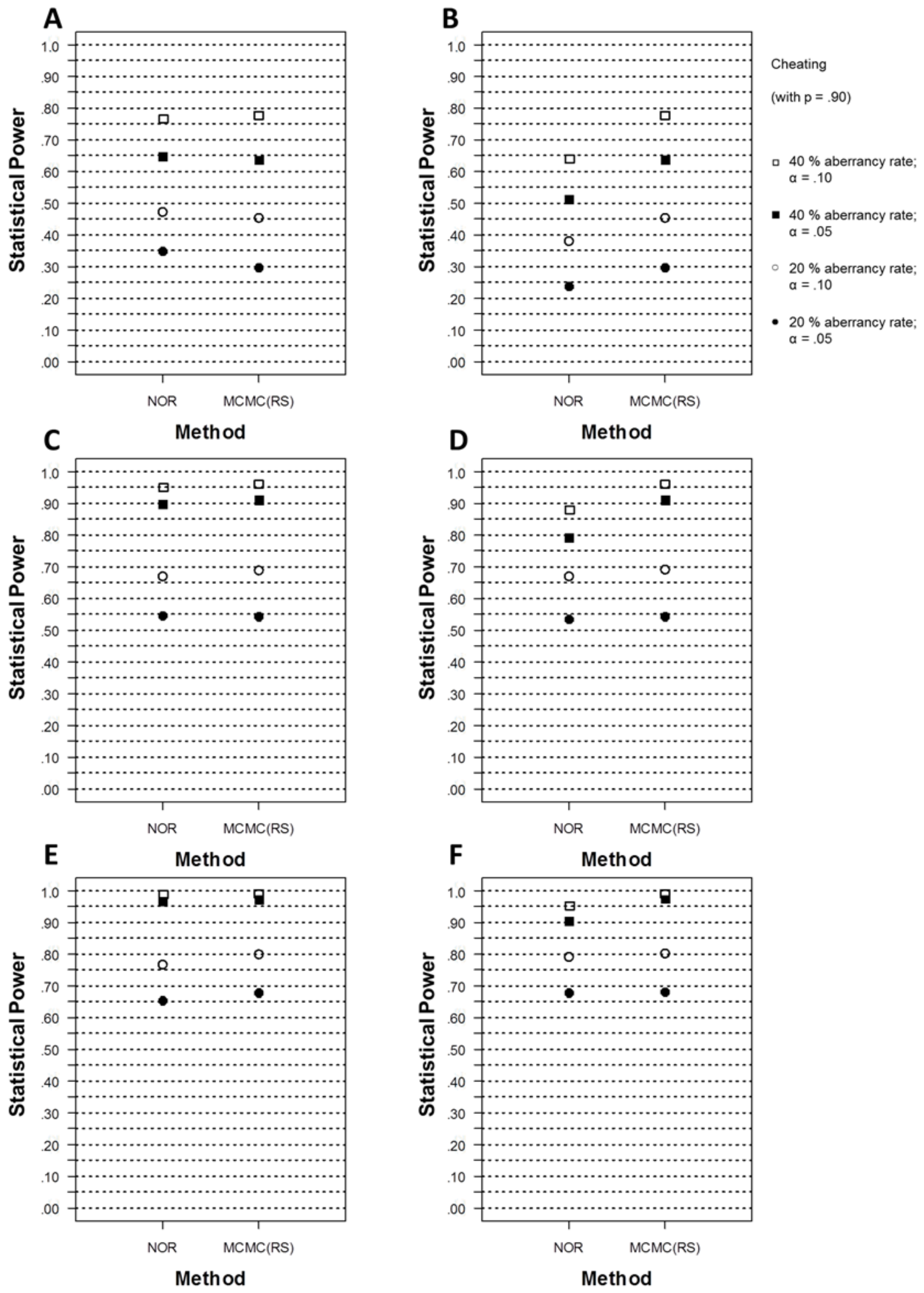

Figure 3.2. **Statistical power of two person fit statistics and two approaches to generate *p*-values (normalization formula, NOR; Markov chain Monte Carlo simulation of the Rasch Sampler, MCMC (RS)) in the condition cheating. A: statistic *U3*, 20 items; B: statistic l_0, 20 items; C: statistic *U3*, 40 items; D: statistic l_0, 40 items; E: statistic *U3*, 60 items; F: statistic l_0, 60 items.**

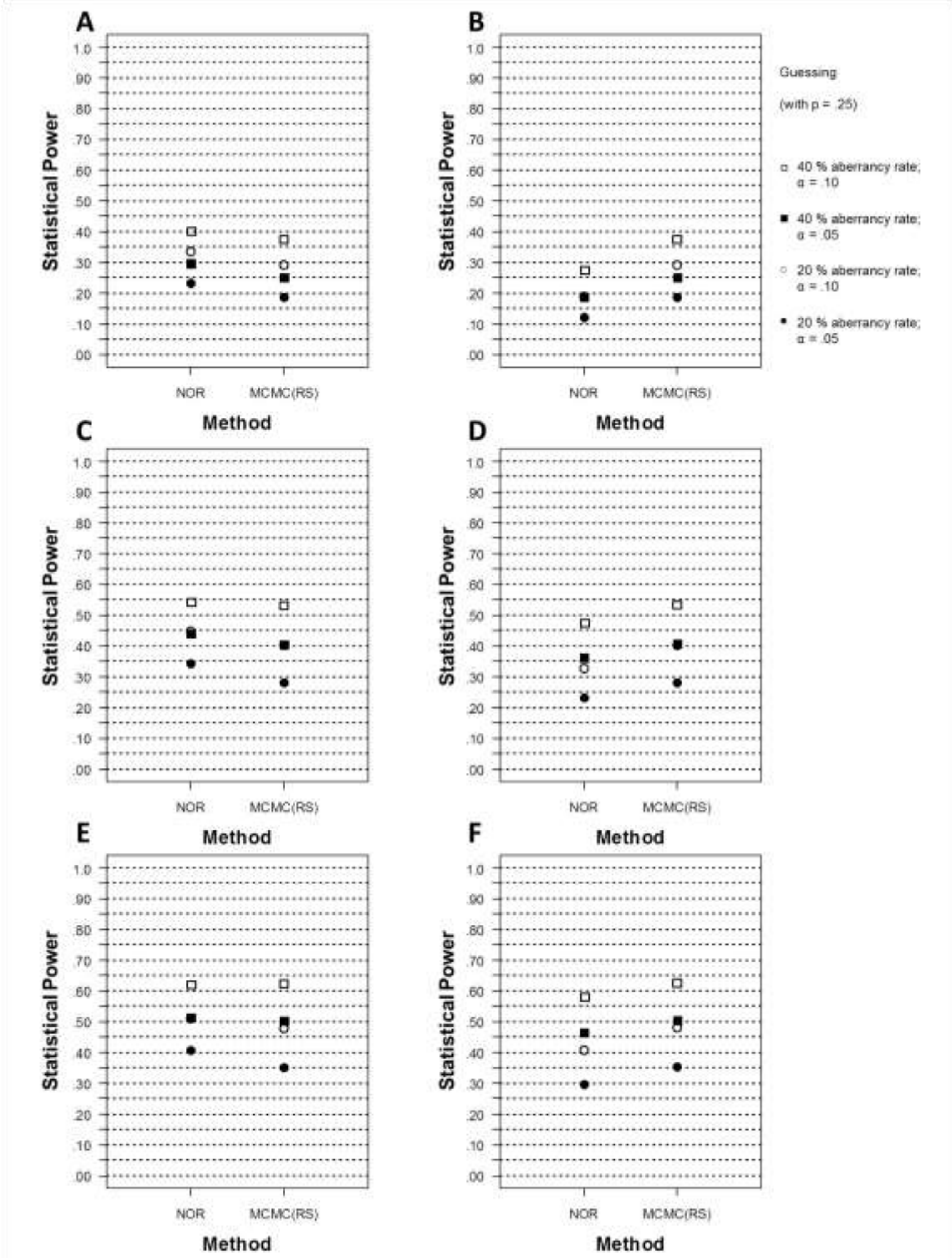

Figure 3.3. **Statistical power of two person fit statistics and two approaches to generate *p*-values (normalization formula, NOR; Markov chain Monte Carlo simulation of the Rasch Sampler, MCMC (RS)) in the condition guessing. A: statistic *U3*, 20 items; B: statistic l_0, 20 items; C: statistic *U3*, 40 items; D: statistic l_0, 40 items; E: statistic *U3*, 60 items; F: statistic l_0, 60 items.**

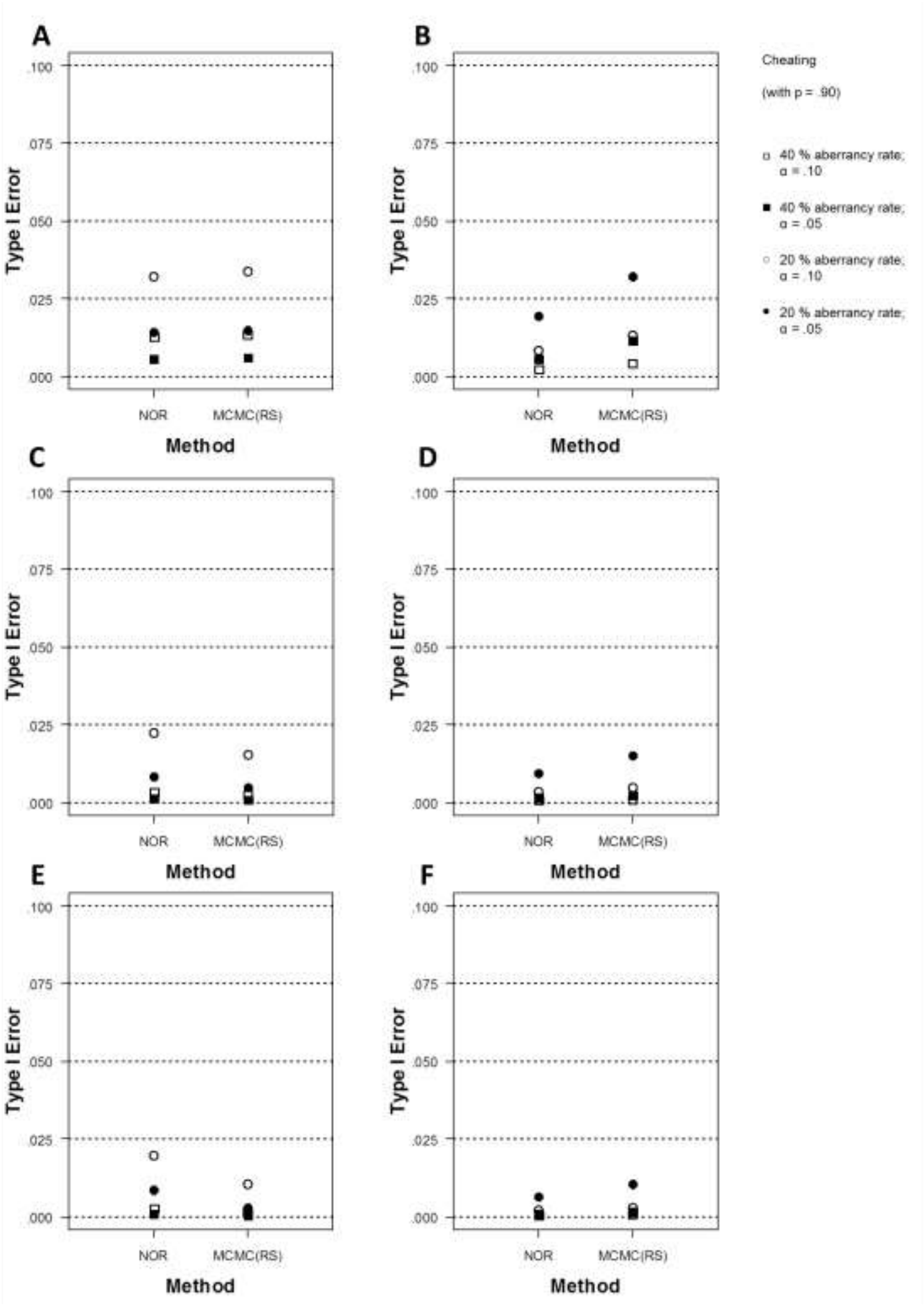

Figure 3.4. **Type I error rates of two person fit statistics and two approaches to generate *p*-values (normalization formula, NOR; Markov chain Monte Carlo simulation of the Rasch Sampler, MCMC (RS)) in the condition cheating. A: statistic *U3*, 20 items; B: statistic l_0, 20 items; C: statistic *U3*, 40 items; D: statistic l_0, 40 items; E: statistic *U3*, 60 items; F: statistic l_0, 60 items.**

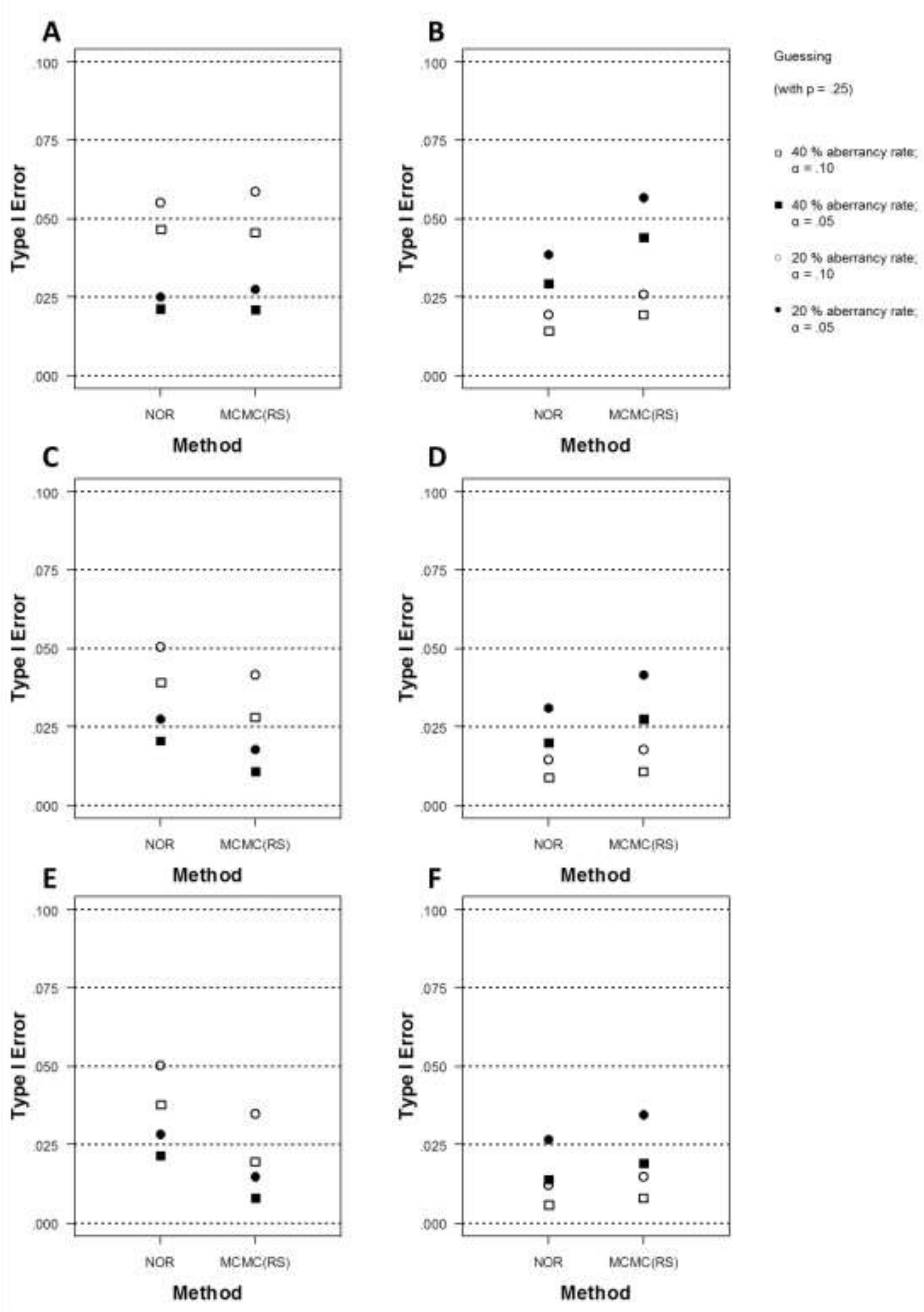

Figure 3.5. **Type I error rates of two person fit statistics and two approaches to generate *p*-values (normalization formula, NOR; Markov chain Monte Carlo simulation of the Rasch Sampler, MCMC (RS)) in the condition guessing. A: statistic *U3*, 20 items; B: statistic l_0, 20 items; C: statistic *U3*, 40 items; D: statistic l_0, 40 items; E: statistic *U3*, 60 items; F: statistic l_0, 60 items.**

In the guessing conditions, power rates of the three approaches are comparable for the *U3* statistic, except for higher rates under NOR with an aberrancy rate of 20 % and α = .05. Similar to the cheating conditions, for the parametric statistic l_0 a general advantage is found under MCMC (RS) in comparison to NOR. This result supports the impression that the power of statistic l_0 is affected by the estimation of θ (Nering, 1995; Reise, 1995). In general, the highest statistical power for both statistics in all of the guessing conditions are found in the condition 60 items, an aberrancy rate of 40 % and α = .10 with MCMC (RS), the lowest in the condition *L* = 20, an aberrancy rate of 20 % and α = .05. For test lengths *L* = 20 power is generally not outstanding in the guessing condition with at best 40 % of the violations detected, and even for longer tests at least more than 30 % of the guessing respondents remain undetected.

Under both types of model violations, Type I error rates (Figures 3.4 and 3.5) are considerably lower than the nominal α-level. Type I error rates are found to be higher in the guessing conditions, but increase with decreasing item number and decreasing percentage of aberrancy under both types of model violations. As the Type I error relates to the Rasch-conform response vectors, the effect of the aberrancy on these rates is an outcome of biased marginals (and therefore biased item parameter estimates) caused by the presence of aberrant response vectors in the data set.

3.5 Application to real data: Rasch scalability of the KFT intelligence test?

To investigate the practical significance of the results obtained by simulated data, the Rasch-scalability of Heller and Perleth’s (2000) multidimensional intelligence test (“kognitiver Fähigkeitstest”, KFT) was investigated by means of person fit statistics. In educational research contexts, the KFT is one of the most often applied testing instruments in German language to assess intelligence as a key predictor of educational success (e.g., Fischer, Labudde, Neumann, & Viiri, 2014; Möller & Bonerad, 2005; PISA-Konsortium Deutschland, 2006).

3.5.1 Research questions and method related to the real data example

The KFT is assumed to be Rasch-conform (Heller and Perleth, 2000; see also PISA-Konsortium Deutschland, 2006) but additional person fit analysis uncovers for which students Rasch ability estimates give valid information on intelligence subcomponents. The approach based on the Rasch Sampler offers an alternative method to identify misfit under a given Type I error rate of α = .05. The research question associated with this KFT data set was: Does person fit analysis support the assumption of adequate fit of the response vectors to the Rasch model and does the interpretation of this person fit information depend on the method of choice?

The studied sample is given by 382 students from Finland, 1193 students from Germany and 560 students from Switzerland participating in the trinational study *Quality of Instruction in Physics Education* (Fischer et al., 2014) funded by the German *Federal Ministery of Education and Research.* The mean age was 15.9 (.66), 45.87 % of students were females, 46.3 % were males and 7.9 % declared no information on gender. More detailed information on the sample is given in Fischer et al. (2014). The present analysis is limited to two dimensions of the KFT: According to Heller and Perleth (2000), subdimension Q2 refers to the “quantitative” part of the intelligence test. To correctly answer items from the Q2 subdimension, respondents need to select one out of five numbers which correctly completes a given column of numbers. Subdimension N2 refers to the “nonverbal” part of the intelligence test. To correctly answer items from the N2 subdimension, respondents need to select one out of five geometrical figures which correctly completes a given geometrical figure to a meaningful pairing. As outlined by Heller and Perleth (2000), the classification of these two dimensions to intelligence sub-components depends on the underlying intelligence theory but both dimensions are related to what others have described as the measurement of reasoning (Wilhelm, 2005). For both dimensions, one out of two parallel forms (A or B) was randomly administered to the students. The Q2 parallel forms consist of 20 items each. The N2 parallel forms were originally formed by 25 items but previous analyses uncovered problems with two items in form B of the test (Möller et al., 2006; Segerer et al., 2012). Subsequent to the estimation of Rasch item difficulties, the weighted and unweighted mean square item fit statistics were computed to investigate Rasch-conformity of the item sample at hand. Table 3.2 gives information on item difficulties and mean square item fit statistics (Bond & Fox, 2007, Chapter 12). Referring to these fit statistics, items from all four test forms tended to overfit, but the questionable items 29 and 34 reached mean square values indicating slight underfit (Table 3.2). This is in line with previous findings by Möller, Bonerad, and Pohlmann (2006) and Segerer, Marx, and Marx (2012). The items 29 and 34 were excluded from the person fit analysis. Response vectors with extreme raw scores (0, 1, 2, L – 2, L – 1, L) were also excluded due to a questionable interpretation of vectors with sparse correct or incorrect answers (e.g., Emons, Sijtsma, & Meijer, 2004) which reduced the actual analyzed sample sizes to N = 828 (form A) and N = 898 (form B) for dimension Q2, and N = 892 and N = 827 for dimension N2.

The person fit statistics $U3$ and l_0 were computed with p-values estimated by normalization formulas (NOR) or the MCMC simulation based on the Rasch Sampler (MCMC(RS)).

3.5.2 Results and conclusions from the real data example

Figure 3.6 and 3.7 show Venn diagrams[3] indicating the overlap of the number of item response vectors identified as misfitting by the two person fit statistics under the methods NOR and MCMC(RS). With a maximum of about 11 % of the response vectors identified as misfitting, the KFT data set again indicates rather acceptable conformity but not perfect fit to the Rasch model in this sample. It is obvious from both figures that non-ignorable differences exist between the number of identified response vectors under NOR and MCMC(RS).

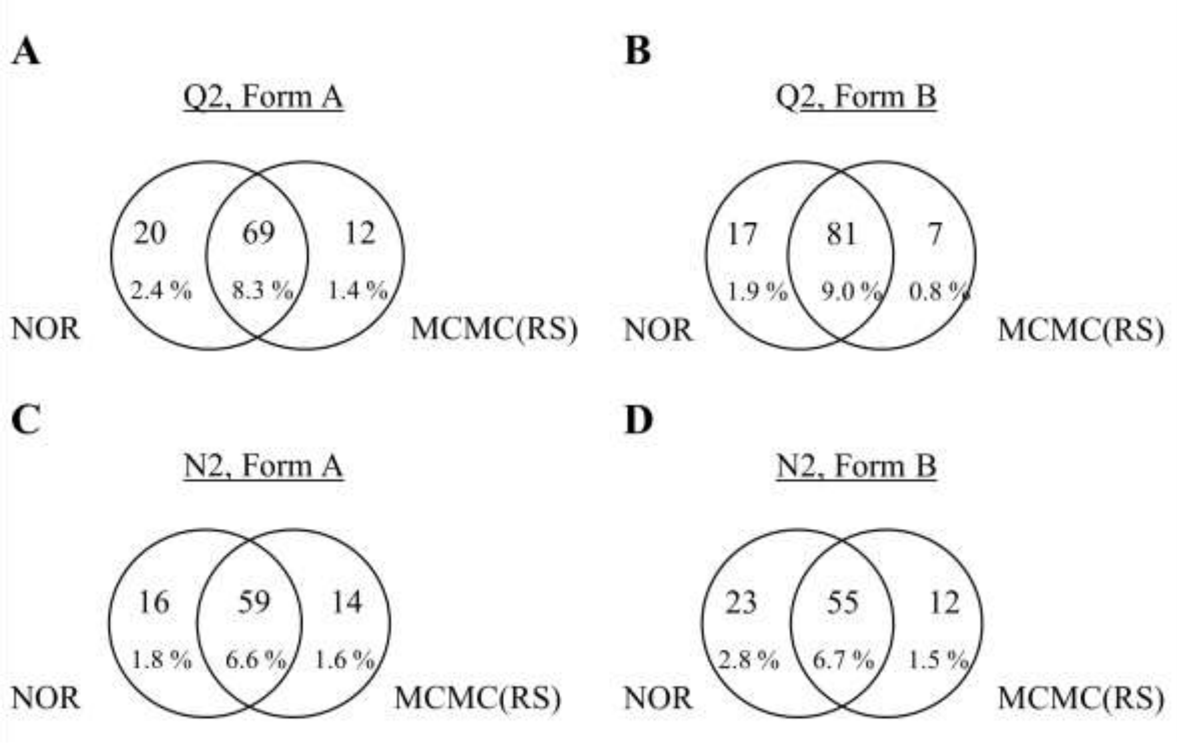

***Figure 3.6.* Venn diagrams representing the overlap of the number of item response vectors identified as misfitting by person fit statistics *U3* for the KFT intelligence test data for two approaches to generate *p*-values (normalization formula, NOR; Markov chain Monte Carlo simulation of the Rasch Sampler, MCMC(RS))**

For statistic *U3*, a substantial amount of response vectors was either solely identified as misfitting under NOR or solely identified under MCMC(RS) (Figure 3.6). Note, however, that $U3^{NOR}$ outperformed the MCMC(RS) method with regard to statistical power only for

[3] The term “Venn diagram” has been used here in reference to a similar application in Emons (2008), p. 241.

Table 3.2

Item parameters and fit statistics of the KFT Q2 and N2 dimensions

	item	item difficulty	unweighted MNSQ	weighted MNSQ	item	item difficulty	unweighted MNSQ	weighted MNSQ
	form A				form B			
dimension Q2	21	-1.557	.928	.875	21	-1.615	.975	.943
	22	-1.621	.915	.812	22	-1.541	.963	.904
	23	-1.709	.844	.644	23	-1.007	.890	.852
	24	-.580	.989	.940	24	-.386	1.012	1.011
	25	-.796	.873	.775	25	-1.373	.887	.785
	26	-1.198	.910	.912	26	-2.074	.960	.883
	27	-.443	.893	.850	27	-.705	.887	.851
	28	-1.207	.811	.698	28	.073	.907	.876
	29	-.417	.931	.901	29	-.058	.839	.806
	30	-.547	.885	.836	30	-.114	.962	.940
	31	-.007	.829	.791	31	.789	.897	.854
	32	.246	.875	.830	32	.176	.928	.914
	33	.246	.843	.782	33	.269	.881	.861
	34	1.160	.827	.777	34	1.098	.872	.811
	35	1.266	.914	.910	35	.685	.914	.895
	36	.595	.773	.721	36	2.323	1.075	1.252
	37	1.484	.833	.785	37	.568	.965	.935
	38	.730	.838	.807	38	1.178	.919	.914
	39	1.766	1.013	1.073	39	.737	.968	.935
	40	2.588	1.148	1.155	40	.977	.996	.967
	form A				form B			
dimension N2	26	-.139	1.049	1.049	26	-.984	.848	.806
	27	-.614	.874	.826	27	-1.159	.830	.715
	28	-1.412	.884	.841	28	-1.110	.848	.776
	29	-1.878	.924	.862	29*	2.127	1.247	1.913
	30	-1.308	.956	.959	30	-.635	.961	.958
	31	.017	1.100	1.098	31	-.127	.959	.950
	32	-.534	.932	.874	32	-.550	.917	.863
	33	-.565	1.002	1.044	33	-.480	.851	.808
	34	.220	1.041	1.043	34*	.658	1.152	1.239
	35	-.571	.879	.811	35	.003	.973	.955
	36	.706	.978	.964	36	-.822	.821	.748
	37	.510	.941	.911	37	-.602	.848	.814
	38	-.311	.762	.696	38	-.723	.873	.844
	39	.117	.891	.852	39	-1.252	.822	.806
	40	.484	1.022	1.017	40	-.374	.833	.782
	41	.251	.868	.833	41	.220	.914	.889
	42	.434	.901	.871	42	.545	.937	.933
	43	.251	.889	.870	43	.973	.943	.936
	44	.033	.792	.733	44	.102	.771	.729
	45	.741	.846	.828	45	-.266	.808	.759
	46	.138	.875	.826	46	.802	.953	.935
	47	1.394	.940	.913	47	.694	.983	.971
	48	.923	.874	.857	48	.617	.834	.797
	49	.333	.843	.794	49	.756	.986	.991
	50	.781	.852	.835	50	1.589	.907	.830

Notes. Items indicated with an asterisk were excluded for person fit analysis.

short test lengths and mild violations, but also suffered from serious inflation of Type I error rates for extreme raw scores. The high number of misfitting response vectors identified under NOR in this sample may therefore either represent mild forms of model violations or might indicate that many of these response vectors were incorrectly flagged. Interestingly in the light of these results, there was also a substantial amount of response vectors in each data set identified by *U3* under MCMC(RS) but ignored under NOR.

Results found for statistic l_0 (Figure 3.7) differ from those by *U3*. The majority of response vectors classified as misfitting by this statistic was identified when MCMC(RS) was applied (about 8 − 10 % of all response vectors) and at maximum one response vector was solely identified under NOR. Hence, results for l_0 suggest the usage of MCMC(RS) enhances the detection of misfitting response vectors without any risk of ignoring questionable vectors. Assuming that results with MCMC(RS) are given under a Type I error rate very close to the nominal rate, statistic *U3* and l_0 display similar statistical power for identifying misfit (81, 88, 73, and 67 response vectors identified by *U3* vs. 80, 88, 72, and 65 response vectors identified by l_0).

In summary, results from the KFT intelligence test data set support the assumption of rather adequate Rasch-conformity, matched those differences in statistical power found in Simulation 2 and illustrated the dependency of the number of identified misfitting response vectors on the particular method to determine *p*-values. Given the results from the simulation studies, fit / misfit classifications under the normalizations are assumed to represent either incorrectly flagged response vectors by *U3* or conservative person fit tests by statistic l_0. These problems are easily prevented by investigating response vectors in person fit analysis by the proposed method based on the Rasch Sampler under an accurately approximated nominal Type I error rate.

3.6 Remarks on this study

Hypothesis testing on person fit with known distributional properties of the statistic facilitate the interpretation of an individual's response vector. This strengthens the usefulness of person fit statistics to test for violations against the principles of Rasch measurement. However, obtaining exact *p*-values for person fit statistics was found to be a troublesome job (de la Torre & Deng, 2008; Emons et al., 2002; Molenaar & Hoijtink, 1990; Nering, 1995; van Krimpen-Stoop & Meijer, 1999). Furthermore, most of the developed methods that offer adequate *p*-values have not been implemented in statistical standard software. We have presented results for applying the MCMC algorithm of the Rasch Sampler as an alternative to conventional approaches. Results of Simulation 1 show that the new approach has well-adjusted Type I errors in contrast to normalization formulas. Results of Simulation 2 found the approach to be as powerful as the normalization formulas or better under most conditions. Results from the application to a real data set of item responses from a multidimensional intelligence test analyzed by person fit statistics emphasized and further exemplified results found in the simulations.

Beside its usefulness for nonparametric tests of Rasch-homogeneity proposed by Ponocny (2001), our results demonstrate further opportunities of the Rasch Sampler to generate reference data for person fit analysis of each single response vector when the enumeration of the probability for each response vector is complex. In contrast to other algorithms for the same objective (like, e.g., the one proposed by Liou & Chang, 1992), the Rasch Sampler is already implemented in the statistical software R (R Development Core Team, 2011) as one of the major statistical software packages and can easily be applied to most kinds of person fit statistics.

3.7 Appendix: Technical Aspects of the Rasch Sampler

The following description is a very concise summary of Verhelst et al. (2007) and gives insights to the technical aspects of the Rasch Sampler. For a more detailed description and proofs see Verhelst (2008).

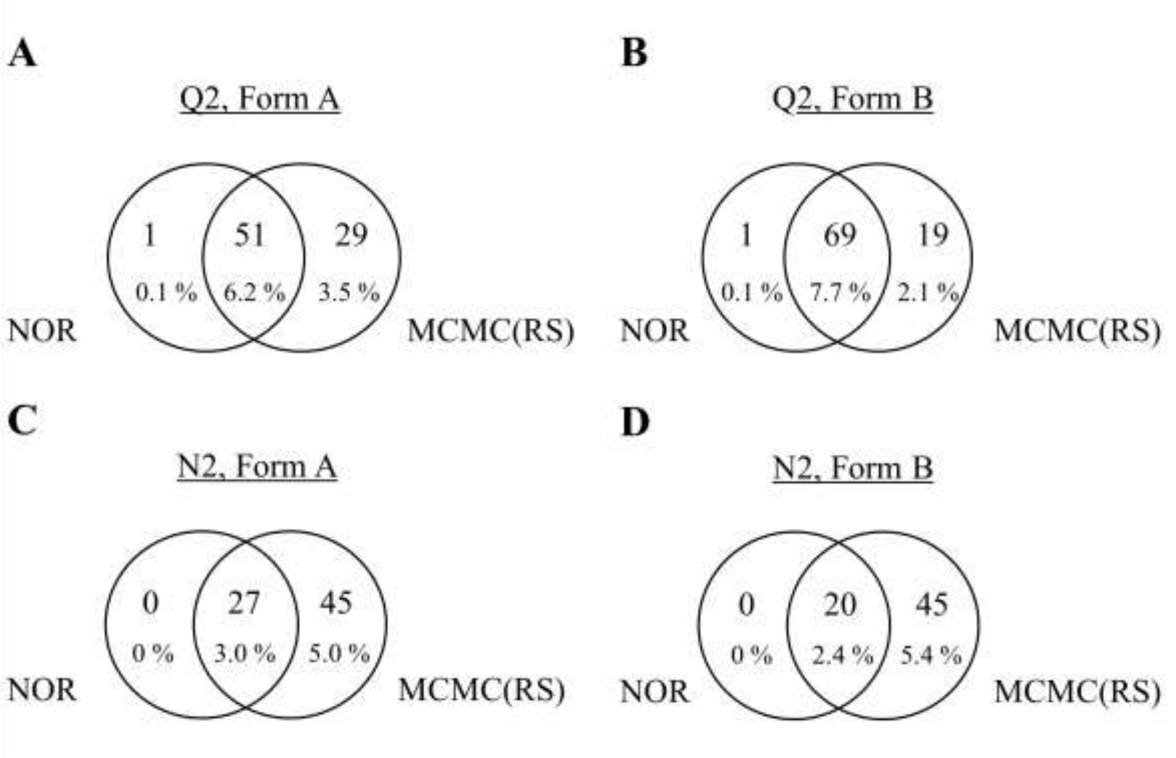

***Figure 3.7.* Venn diagrams representing the overlap of the number of item response vectors identified as misfitting by person fit statistics l_0 for the KFT intelligence test data for two approaches to generate *p*-values (normalization formula, NOR; Markov chain Monte Carlo simulation of the Rasch Sampler, MCMC(RS))**

If the Rasch model is valid, all binary matrices with the same marginals as the observed matrix have the same probability, and for a given statistic the null distribution can be found by sampling from this collection of data matrices. It is not difficult to generate new matrices with constant marginals; the difficulty arises how to sample these with equal probability. The method to generate new matrices with constant marginals is based on binomial transformations: For the observed $N \times L$ matrix $\boldsymbol{A}$ with row totals given by vector $\boldsymbol{R}$ and column totals given by vector $\boldsymbol{C}$, the sampling space, the number of all binary data matrices with these marginals, is denoted as $\Sigma_{\boldsymbol{RC}}$. Within each $N \times 2$ submatrix of $\boldsymbol{A}$ there are four possible row vectors: (0 0), (1 1), (1 0) and (0 1). A binomial transformation can only be applied to a column pair where both vectors (1 0) and (0 1) exist. Let $(e + f)$ be the number of rows with a row total of one, where e is the frequency of vector (1 0) and f is the frequency of vector (0 1) in the column pair (i,j). A binomial operation is an operation where a one is assigned to column i and a zero to column j for e of these $(e + f)$ rows, while a complementary vector is assigned to the f other rows. If the resulting matrix differs from $\boldsymbol{A}$ it will be called a binomial transform. The set of matrices constructed by binomial transformation of $\boldsymbol{A}$ it denoted by $\mathcal{A}_B^{(i,j)}(\boldsymbol{A})$.

In a MCMC interpretation all of these matrices, generated by binomial transformations, are regarded as states in a finite Markov chain: At a current state t, described by the matrix $\boldsymbol{A}_t$, the process can move to any other state s of a subset of these binary matrices with same marginals. This concrete subset will be called the neighborhood of the matrix $\boldsymbol{A}_t \in \sum_{\boldsymbol{RC}}$ and it is defined as:

$$\mathcal{A}_B\ (\boldsymbol{A}_t) = \bigcup_{(i,j)} \mathcal{A}_B^{(i,j)}\ (\boldsymbol{A}_t). \qquad [3.9]$$

Due to the fact that the actual choice for moving from the current state to the next is given by a random sampling from its neighborhood, the probability of each matrix to be sampled depends on the size of its neighborhood which leads to a stationary distribution that is not exactly uniform (Verhelst, 2008). To make sure that all matrices are being sampled with equal probability, Verhelst (2008) proposed to apply the Metropolis-Hastings algorithm. The resulting Markov chain has a transition matrix $\boldsymbol{Q}^*= (q^*_{ij})$ with a defined vector $\boldsymbol{\pi}$ as stationary vector. Its diagonal elements are given by q_{tt}^{*} = *1- Σ*$_{j \neq i}$ q_{st}^{*} and its off-diagonal elements are defined by:

$$q_{st}^{*} = \tau_{st} \times q_{st} \ , \qquad [3.10]$$

with

$$\tau_{st} = \begin{cases} \min\left[\left(\frac{\pi_s q_{ts}}{\pi_t q_{st}}\right), 1\right] & if\ \pi_t q_{st} > 0 \\ 1 & if\ \pi_t q_{st} = 0. \end{cases} \qquad [3.11]$$

The probabilities q_{st} can be chosen arbitrarily with the Metropolis-Hastings algorithm. For the problem considered here:

$$q_{st} = w_{st} \times [k_2(\boldsymbol{A}_t)]^{-1} \ , \qquad [3.12]$$

where

$$w_{st} = \begin{cases} \left[\binom{(e+f)_{ij}}{e_{ij}} - 1\right]^{-1} & if\ \boldsymbol{A}_s \epsilon A_B^{(i,j)}(\boldsymbol{A}_t) \\ 0 & if\ \boldsymbol{A}_s \notin A_B(\boldsymbol{A}_t) \end{cases} , \qquad [3.13]$$

gives the probability of sampling from the binomial neighborhood and $[k_2(A_t)]^{-1}$ relates the algorithm to the k_2-measure of $A_t \in \sum_{RC}$ defined as:

$k_2(\boldsymbol{A}) = \#\{(i, j) : i < j \leq k$, with $e_{ij} \times f_{ij} > 0$, for (i, j).

Turning back to Equation 3.10, a uniform distribution is induced with the Metropolis Hastings algorithm when π_s / π_t = 1. Furthermore as weights w_{st} = w_{ts} (Verhelst, 2008) and because of Equations 3.11 and 3.13 the equation simplifies to:

$$\tau_{st} = \begin{cases} \min\left[\left(\frac{k_2(\boldsymbol{A}_t)}{k_2(\boldsymbol{A}_s)}\right), 1\right] & if\ k_2(\boldsymbol{A}_s) > 0 \\ 1 & if\ k_2(\boldsymbol{A}_s) = 0. \end{cases} \qquad [3.14]$$

This leads to a relatively simple algorithm, in the words of Verhelst et al. (2007, p. 6):

“Importance Sampling and Metropolis-Hastings:

1. Select randomly a pair of columns from the $k_2(\boldsymbol{A})$ regular column pairs of $\boldsymbol{A}$.

2. Apply a random binomial operation to the selected pair, yielding $\boldsymbol{A}^*$.

    (a) If $\boldsymbol{A}^* = \boldsymbol{A}$, repeat step 2.

    (b) Otherwise,

    i. If $k_2(\boldsymbol{A}^*) \leq k_2(\boldsymbol{A})$, the new state is $\boldsymbol{A}^*$,

    ii. If $k_2(\boldsymbol{A}^*) > k_2(\boldsymbol{A})$, then the new state remains $\boldsymbol{A}$ with probability

    $1-k_2(\boldsymbol{A})/k_2(\boldsymbol{A}^*)$,

    otherwise the new state is $\boldsymbol{A}^*$.”

3.8 References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Bayarri, M. J., & Castellanos, M. E. (2001). A comparison between *p*-values for goodness-of-fit checking. In E. I. George (Ed.), *Bayesian methods with applications to science, policy, and official statistics* (pp.1–10). Luxembourg: Office for Official Publications of the European Communities.

Bedrick, E. J. (1997). Approximating the conditional distribution of person fit indexes for checking the Rasch model. *Psychometrika*, *62*, 191–199.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd Ed.). Mahwah, NJ: Lawrence Erlbaum.

de la Torre, J., & Deng, W. (2008). Improving person fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159–177.

Dimitrov, D. M., & Smith, R. M. (2006). Adjusted Rasch person-fit statistics. *Journal of Applied Measurement, 7*, 170–183.

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.

Emons, W. H. M. (2008). Nonparametric person-fit analysis of polytomous item scores. *Applied Psychological Measurement, 32*, 224–247.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement, 26*, 88–108.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person-response function in person-fit analysis. *Multivariate Behavioral Research, 39*, 1–35.

Fischer, H., Labudde, P., Neumann, K., & Viiri, J. (Eds.) (2014), *Quality of Instruction in Physics – Comparing Finland, Germany and Switzerland*. Münster: Waxmann.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman and Hall/CRC.

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.

Heller, K. A., & Perleth, C. (2001). *Kognitiver Fähigkeits-Test für 5. – 12./13. Klassen, Revision – Materialien-Manual-Koffer* [The cognitive skill test for grade 5-12/13: Revision-material-manual-case]. Göttingen: Beltz Test.

Jackman, S. (2009). *Bayesian analysis for the social sciences.* New York, NY: Wiley.

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement, 1*, 152–176.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298.

Kubinger, K. D., & Draxler, C. (2007). Probleme bei der Testkonstruktion nach dem Rasch-Modell [Some problems in calibrating an item pool according to the Rasch model]. *Diagnostica, 53*, 131–143.

Lamprianou, I. (2010). The practical application of optimal appropriateness measurement on empirical data using Rasch Models. *Journal of Applied Measurement, 11*, 409–423.

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

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

Liou, M., & Chang, C.-H. (1992). Constructing the exact significance level for a person fit statistic. *Psychometrika, 57*, 169–181.

Mair, P., & Hatzinger, R. (2007a). CML based estimation of extended Rasch models with the eRm package in R. *Psychology Science, 49*, 26–43.

Mair, P., & Hatzinger, R. (2007b). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1–20.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). The influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111–120.

Meijer, R. R., & Nering, M. L. (1997). Ability estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321–336.

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

Möller, J. & Bonerad, E.-M. (2007). Fragebogen zur habituellen Lesemotivation [Habitual Reading Motivation Questionnaire]. *Psychologie in Erziehung und Unterricht, 54*, 259–267.

Möller, J., Bonerad, E.-M. & Pohlmann, B. (2006). Antwortmuster und Itemkennwerte: ein unlösbares Item im KFT 4-12+ R [Response patterns and item characteristics: An insolvable item in the KFT 4-12+R]. *Diagnostica*, *52*, 73–75.

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

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

OECD (2004). *Learning for tomorrow's world: First results from PISA 2003.* Paris: OECD Publishing.

PISA-Konsortium Deutschland (2006) (Eds.). *PISA 2003. Dokumentation der Erhebungsinstrumente* [PISA 2003 – test documentation]. Münster: Waxmann.

Ponocny, I. (2000). Exact person fit indexes for the Rasch model for arbitrary alternatives. *Psychometrika, 65*, 29–42.

Ponocny, I. (2001). Non-parametric goodness-of-fit tests for the Rasch model. *Psychometrika*, *66*, 437–459.

R Development Core Team (2011). *R: A language and environment for statistical computing* [Computer program]. Vienna: R Foundation for Statistical Computing.

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

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

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

Rizopoulos, D. (2013). *ltm - Latent trait models under IRT: Reference manual* (Ver. .9-7). Retrieved February 2012 from http://cran.r-project.org/web/packages/ltm/ltm.pdf

Smith, R.M. (1982). *Detecting measurement disturbances with the Rasch model* (Unpublished doctoral dissertation). University of Chicago, Chicago, IL.

Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika, 66*, 331–334.

Tarnai, C., & Rost, J. (1990). *Identifying aberrant response patterns in the Rasch model: The Q index. Sozialwissenschaftliche Forschungsdokumentationen*. Münster: Institut für sozialwissenschaftliche Forschung e.V. .

van der Flier, H. (1980). *Vergelijbaarheid van individuele testprestaties* [Comparability of individual test performance]. Lisse: Swets and Zeitlinger.

van der Flier, H. (1982). Deviant response patterns and comparability of test scores. *Journal of Cross-Cultural Psychology, 13*, 267–298.

van Krimpen-Stoop, E. M. L. A., & Meijer, R. R. (1999). The null distribution of person-fit statistics for conventional and adaptive tests. *Applied Psychological Measurement, 23*, 164–180.

Verhelst, N. D. (2008). An efficient MCMC algorithm to sample binary matrices with fixed marginals. *Psychometrika, 73*, 705–728.

Verhelst, N. D., Hatzinger, R., & Mair, P. (2007). The Rasch sampler. *Journal of Statistical Software, 20*(4), 1–14.

4 STUDY II – BOOTSTRAP PERSON FIT TESTS WITH WEIGHTED ML SCORING

4.1 Background

When a test has been analyzed according to IRT modeling, aberrant and unlikely response vectors may restrict the validity of test results (e.g., Meijer, 1997; Schmitt, Chan, Sacco, McFarland, & Jennings, 1999; Schmitt, Cortina, & Whitley, 1993; Chapter 1). Therefore, one goal of test administrators is the identification of aberrant responders by means of person fit statistics (Meijer & Sijtsma, 2001). Parametric person fit statistics based on estimated item and ability parameters are regularly applied for IRT modeling. These models can be used to describe test data with a limited number of parameters which can be estimated by using, for example, MML or CML methods (Baker & Kim, 2004). Computationally efficient as well as consistent and asymptotically unbiased (see Chapter 1), these methods are an intuitive choice for item parameter estimation; for ability estimation, weighted ML methods like Warm’s (1989) estimator outperform the conventional ML estimator in terms of reducing bias.

With regard to the expected data matrix, the estimation of parametric IRT models is rather vulnerable towards disturbances like cheating, guessing, carelessness or test anxiety (see Chapter 1) that may be found in observed test data. Robust weight functions (e.g., Heritier, Cantoni, Victoria-Feser, & Copt, 2009; Maronna, Martin, & Yohai, 2006) can be applied to handle disturbances and estimate the model parameters, in particular the latent abilities (Mislevy & Bock, 1982; Schuster & Yuan, 2011; Wainer & Wright, 1980). Smith (1985) compared the usage of person fit statistics and robust scoring methods in few typical aberrancy scenarios and warned that robust scoring functions might mask diagnostically important information on the response behavior. But both concepts, person fit analysis and robust scoring methods, may also be combined: A crucial finding by Reise (1995) on the in

Chapter 3 defined parametric person fit statistic l_z (Drasgow, Levine, & Williams, 1985; Equations 3.6-3.8) reveals that the statistical power of this statistic depends on the scoring method for the latent ability θ. He found that compared to ML and EAP (Bock & Mislevy, 1982; see Chapter 1) person fit statistic l_z computed using the robust BS estimator (Mislevy & Bock, 1982), a scoring method provided by the IRT software package BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), had higher power to detect aberrancy. In the presence of downweighting aberrant responses by the robust BS scoring method also the problem of deflated Type I error levels previously determined for l_z was less pronounced. As outlined in Chapter 2 and Chapter 3, it is a well-known finding that results for the normalized person fit statistic l_z are often problematic due to an incorrect approximation of the standard normal distribution, particularly when true abilities are replaced by scoring methods to estimate θ (e.g., de la Torre & Deng, 2008; Li & Olejnik, 1997; Meijer & Nering, 1997; Reise, 1995; see Chapters 2 and 3).

The Basis for the conventional ML scoring is the log likelihood contribution of the dichotomous item response x_i determined by

$$l_i(\theta) = x_i \log[P_i(\theta)] + (1 - x_i) \log[1 - P_i(\theta)]. \qquad [4.1]$$

It is the logarithm of the probability of a single item response $P_i(\theta)^{x_i}[1 - P_i(\theta)]^{1-x_i}$. Given the local independence assumption (see Chapter 1) holds, the log likelihood contributions of L item responses are added. The ML estimator, $\hat{\theta}_{\mathrm{ML}}$, must satisfy the relationship

$$\sum_{i=1}^{L} \left(\frac{\partial l_i(\theta)}{\partial \theta} \right) = 0 \qquad [4.2]$$

where

$$\sum_{i=1}^{L}\left(\frac{\partial l_i(\theta)}{\partial\theta}\right) = \sum_{i=1}^{L}[x_i - P_i(\theta)]\frac{P'_i(\theta)}{P_i(\theta)[1 - P_i(\theta)]} \qquad [4.3]$$

and with $P'_i(\theta)$ denoting the first derivative of $P_i(\theta)$ with respect to θ (for details see Baker & Kim, 2004). Hence, to obtain $\hat{\theta}_{\mathrm{ML}}$ the first derivative of the sum of log likelihood item contributions is equated to zero (Warm, 1989). ML estimates are substantially biased outward. As an alternative to the ML estimator, Warm's (1989) WL estimator, $\hat{\theta}_{\mathrm{WL}}$, was developed to reduce the first order bias term from the ML by satisfying the equation

$$\sum_{i=1}^{L}[x_i - P_i(\theta)]\frac{P'_i(\theta)}{P_i(\theta)[1 - P_i(\theta)]} + \left\{\sum_{i=1}^{L}\frac{P'_i(\theta)\,P''_i(\theta)}{P_i(\theta)[1 - P_i(\theta)]}\right\}\frac{1}{2I(\theta)} = 0 \qquad [4.4]$$

with $P''_i(\theta)$ as the second derivative of $P_i(\theta)$ and $I(\theta)$ as the test information at the ability location given by

$$I(\theta) = \sum_{\mathrm{i}=1}^{L}\frac{P'^2_i(\theta)}{P_i(\theta)[1 - P_i(\theta)]}. \qquad [4.5]$$

(e.g., Baker & Kim, 2004). WL estimates are available for raw scores including *r* =1 and *r* = *L*.

Following Mislevy and Bock (1982, p. 725) who argued that ML is "…overly sensitive to measurement disturbances that are common in educational testing…", Schuster and Yuan (2011) recently suggested a different weight function as a new method of robust estimation of the latent ability. Reformulating Equation 4.2 to

$$\sum_{i=1}^{L} w(u_i)\left(\frac{\partial l_i(\theta)}{\partial\theta}\right) = 0 \qquad [4.6]$$

offers a weighted computation of the ML estimate. The weight $w(u_i)$ depends on a weighting function *w*, which can be chosen from several functions, and the residual u_i which defines an outlier from the expected parameter range. As a useful choice (Schuster & Yuan, 2011), residuals in the 2PLM (see Chapter 1) with item discrimination a_i, item difficulty b_i and ability θ might be defined as $u_i = a_i(\theta - b_i)$, which is also the definition used in this study. The Huber-type weight (HU) is

$$w(u) = \begin{cases} 1 & \text{for } |u| \leq TC_{\mathrm{HU}} \\ TC_{\mathrm{HU}}/|u| & \text{for } |u| > TC_{\mathrm{HU}} \end{cases} \qquad [4.7]$$

with TC_{HU} as a tuning constant for the HU typically chosen to be 1. Figure 4.1 shows the Huber-type weight for different values of TC_{HU}. Schuster and Yuan (2011) compared the ML, the BS and the HU in a simulation study. They found that the BS has the lowest bias, but the Huber-type estimator has a smaller sampling variability than the BS. In contrast to the biweight function, convergence problems are not expected with the Huber-type function which suggests the HU as an alternative to BS and also to previous attempts to define robust scoring methods.

Contrary to person fit tests by statistic l_z based on the conventional ML, the EAP and the BS scoring method (Meijer & Nering, 1997; Reise, 1995) as well as the WL scoring method (van Krimpen-Stoop & Meijer, 1999) which illustrated deflated nominal Type I error rates when the standard normal distribution was applied as theoretical distribution, the distributional properties of person fit scores computed based on HU estimates have not been investigated. Hence, this chapter first gives some additional evidence – in form of a short side note – on the (in)adequacy of the theoretical null distribution of l_z based on the robust HU scoring method and compares the results to those obtained based on the conventional (unweighted) ML method and based on true abilities. As a consequence from these results, two previously suggested adjustment options of person fit tests by l_z are revisited: a correction of the normalization formula underlying l_z (Snijders, 2001) and usage of simulation-based methods (Conijn, Emons, & Sijtsma, 2014; de la Torre & Deng, 2008;

Rizopoulos, 2013; van Krimpen-Stoop & Meijer, 1999). Subsequently, two variants of simulation-based methods for person fit tests based on weighted ML estimates are proposed and evaluated in the following with regard to Type I error rate recovery and statistical power.

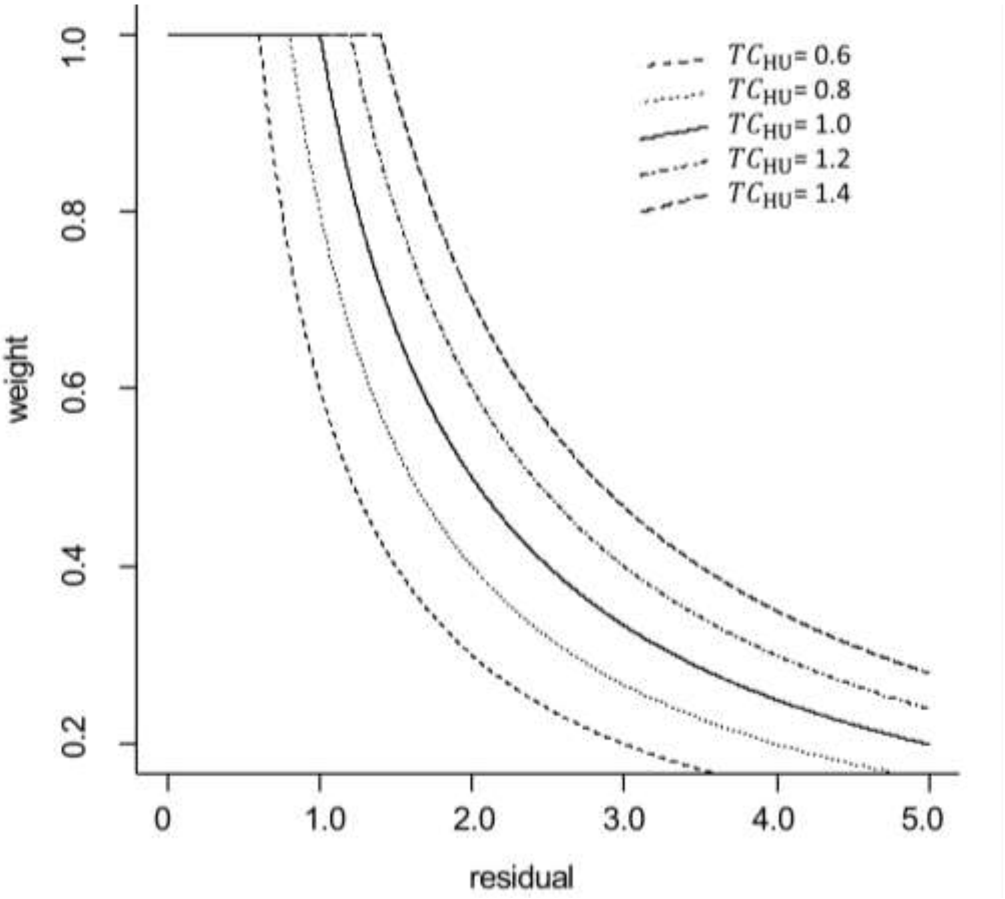

***Figure 4.1*: Huber weights for different values of TC_{HU}**

4.2 A side note on the distributional properties of person fit statistic l_z under robust HU scoring

As there is strong evidence elsewhere (Nering, 1995; Reise, 1995; van Krimpen-Stoop & Meijer, 1999) for deviations of the empirical distribution of l_z from the standard normal distribution as a theoretical sampling distribution, it was not intended to present a large simulation study to further emphasize this result under the robust HU scoring but to provide an illustration on the (in)adequacy of the normalization of l_z under conventional and robust

HU scoring by means of simulated 2PLM data. The empirical sampling distribution of l_z was computed in four scenarios distinguishing extreme and medium ability levels (θ = -2 and θ = 0) in short and medium test lengths (20 and 40 items). The item parameters were chosen as described in detail in Chapter 4.4. Figure 4.2 and Figure 4.3 present the empirical distribution (N = 10,000) of the statistic, the four moments mean, variance, skewness and kurtosis, as well as the empirical critical values for α-levels of .01, .05 and .10 based on the ML scoring method, the HU scoring method and the true ability levels (TT) to compute l_z. Based on the ML scoring method the distribution of l_z generated from simulated data had a mean and a variance different from the expected values derived from the standard normal distribution. Instead, a positive mean for θ = 0 and a strongly reduced variance were found, particularly for θ = -2 and going along with a positive kurtosis. Additionally, the distribution of l_z was negatively skewed. The critical values in the simulated distributions were higher than those under a standard normal distribution resulting in deflated Type I error rates for extreme θ. This finding is in line with previous research on this statistic (Meijer & Nering, 1997; Reise, 1995; Snijders, 2001). In particular for the extreme ability level (θ = -2), l_z based on HU approached the standard normal distribution more adequately than l_z based on ML. The mean of the l_z distribution was closer to 0 but the distribution had a variance smaller than 1, negative skewness and positive kurtosis. For l_z computed based on TT the expected values under a standard normal distribution with a mean equal to 0 and a variance equal to 1 were well approximated, but the distributions still display negative skewness and positive kurtosis, which leads to inflated Type I error rates found in previous analyses (e.g., de la Torre & Deng, 2008; Meijer & Nering, 1997; Reise, 1995; van Krimpen-Stoop & Meijer, 1999). Deviations of the empirical distribution of l_z from the standard normal distribution were generally stronger for the shorter test length. Summarizing the results, previous findings by Reise (1995) as well as Meijer and Nering (1997) for l_z based on the robust BS estimator for the latent ability were similarly found for HU. The distributions of l_z under the robust scoring methods differed from those under conventional ML scoring and were much closer to the standard normal distribution, particularly for extreme abilities. The statistic was still not perfectly normalized under the robust HU scoring method (and also TT). Additionally, the distribution of l_z was not consistent across different ability levels (Reise, 1995). Obviously, these problems related to the normalzation do not only affect the identification of response vector underfit by l_z (for which the statistic is most often applied), but also the identification of response vector overfit based on the positive tail of its distribution.

Consequently, l_z may be used as a descriptive measure of person fit but it is not advisable to apply hypothesis tests on person fit based on this statistic and the standard normal as a null distribution (even though person fit tests with statistic l_z based on HU will be less conservative than those based on ML). Two general options exist for handling the problem of incorrect normalization of l_z (and many other person fit statistics).

The first option is to adjust the normalization formula underlying l_z (Bedrick, 1997; Snijders, 2001); Snijders (2001) has presented an approach to correct the first two moments of many standardized parametric person fit statistics. The second option is to use simulation based methods (Conijn et al., 2014; de la Torre & Deng, 2008; Rizopoulos, 2013; van Krimpen-Stoop & Meijer, 1999).

The method by Snijders (2001) – in the following denoted by SNIJ – to correct the normalization is defined as follows. Referring to the (centered) general person fit statistic defined by Equation 1.18, statistic l_z takes the form of $V(\theta)$ by assuming ability estimates and selecting a weight[4]

$$v_i(\hat{\theta}) = \log \frac{P_i(\hat{\theta})}{1-P_i(\hat{\theta})}. \qquad [4.8]$$

[4] Contrary to the original notations by Snijders (2001), the correction is here defined by referring to Greek letters to prevent confusion with formulas previously presented in this script.

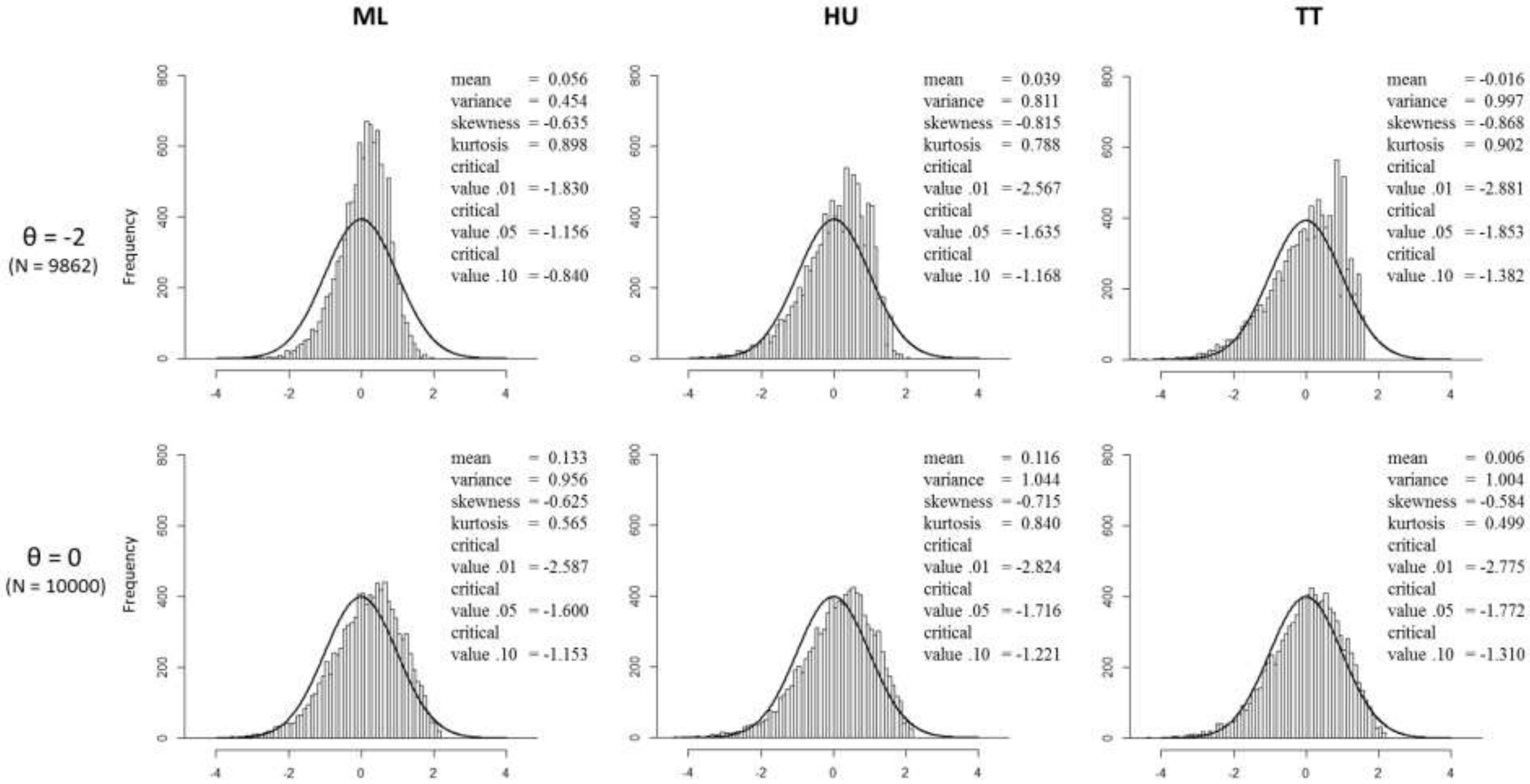

***Figure 4.2*: Distributional characteristics of l_z based on two scoring methods and true θ for 20 items**

Notes. ML = Maximum likelihood scoring method; HU = Huber-type weighted scoring method; TT = true trait level; 138 response vectors with scores r = 0 were excluded for θ = -2.

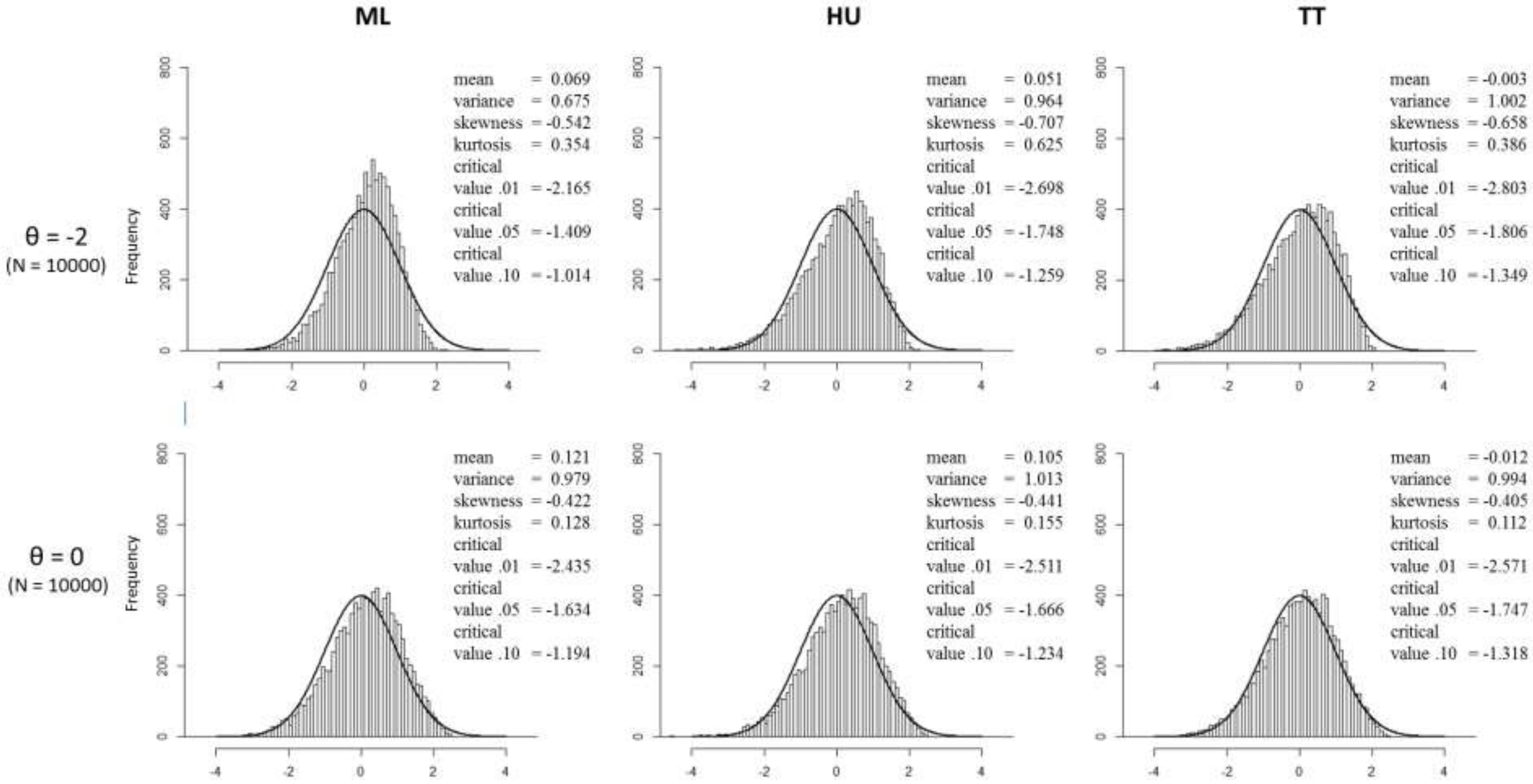

***Figure 4.3*: Distributional characteristics of l_z based on two scoring methods and true θ for 40 items**

Notes. ML = Maximum likelihood scoring method; HU = Huber-type weighted scoring method; TT = true trait level.

It is obtained that

$$l_0 - E(l_0) = V(\hat{\theta}); Var(l_0) = Var[V(\hat{\theta})] \qquad [4.9]$$

(Magis, Raîche, & Béland, 2012; Snijders, 2001). Snijders (2001) proposed to use a corrected weight

$$\tilde{v}_i(\hat{\theta}) = v_i(\hat{\theta}) - \varphi(\hat{\theta})\rho_i(\hat{\theta}) \qquad [4.9]$$

with

$$\varphi(\hat{\theta}) = \frac{\sum_{i=1}^{L} P'_i(\hat{\theta})\ v_i(\hat{\theta})}{\sum_{i=1}^{L} P'_i(\hat{\theta})\ \rho_i(\hat{\theta})} \qquad [4.10]$$

and

$$\rho_i(\hat{\theta}) = \frac{P'_i(\hat{\theta})}{P_i(\hat{\theta})[1 - P_i(\hat{\theta})]} \qquad [4.11]$$

where $P'_i(\hat{\theta})$ is defined as before. Snijders (2001) defined his method for any type of estimator satisfying

$$\rho_0(\hat{\theta}) + \sum_{i=1}^{L}[X_i - P_i(\hat{\theta})]\,\rho_i(\hat{\theta}) = 0 \qquad [4.12]$$

which is fulfilled by the ML scoring method with $\rho_0(\hat{\theta}) = 0$ (for the correction depending on $\hat{\theta}_{ML}$ and other estimators see Magis, Raîche, & Béland, 2012). The corrected expectation of the statistic is then given by

$$E\left(\tilde{V}(\hat{\theta})\right) \approx -\,\varphi_i(\hat{\theta})\,\rho_0(\hat{\theta}), \qquad [4.13]$$

the corrected variance is defined by

$$Var\left(\tilde{V}(\hat{\theta})\right) \approx \sum_{i=1}^{L} \tilde{v}^2{}_i(\hat{\theta})P_i(\hat{\theta})[1 - P_i(\hat{\theta})]. \qquad [4.14]$$

This corrected normalization was developed by Snijders (2001) to better approximate the normal distribution for l_z. Empirical analyses uncovered the variance of l_z under SNIJ to be closer to the standard normal distribution even though its mean remains to be biased and the skewness and kurtosis were similar to the original version of l_z (van Krimpen-Stoop & Meijer, 1999); for more details on the method, please refer to Snijders (2001) or Magis, Raîche, and Béland (2012).

To facilitate the interpretation of l_z, simulation-based methods are an alternative to using critical values from the theoretical distribution under a standard normal distribution (see Chapter 2). The methods evaluated here, denoted as parametric bootstrap by Conijn et al. (2014) or van Krimpen-Stoop and Meijer (1999), can be applied to different types of dichotomous and polytomous IRT models (for similar approaches see Conijn et al., 2014; de la Torre & Deng, 2008; Rizopoulos, 2013; van Krimpen-Stoop & Meijer, 1999). Van Krimpen-Stoop and Meijer (1999) evaluated the usefulness of a parametric bootstrap by assuming θ equalled $\hat{\theta}$ estimated by the WL. They found that the simulated distributions of significance probabilities for l_z did not differ significantly from the uniform distribution for paper-and-pencil test designs, but were not in accordance with the uniform distribution for CAT which indicates that MC simulation is not useful for person fit analysis in this form of test administrations. As a limitation of their study, the authors did not systematically vary the underlying abilty level of the respondents but drew true ability levels from a standard normal distribution which does not allow studying empirical Type I error rates systematically for different θ.

More recently, de la Torre and Deng (2008) proposed a similar parametric bootstrap approach relying on an adjusted, shrinkage-corrected EAP scoring method. Shrinkage describes the effect that Bayesian estimators like the EAP regress to the expectation of the a priori distribution (Baker & Kim, 2004, Chapter 7). To correct the shrinkage effect and account for test unreliability, de la Torre and Deng (2008) proposed to apply the following adjustment to improve Type I error recovery and statistical power of person fit statistics. As the amount of shrinkage is inversely related to $I(\theta)$, the authors refer to the reliability estimate

$$\mathrm{rel}(\theta) = \frac{1}{1 + 1/I(\theta)} \qquad [4.15]$$

(for details see also Wainer et al., 2001). To find the shrinkage-adjusted latent ability, θ is first estimated by the EAP, $\hat{\theta}_{\mathrm{EAP}}$, before it is corrected relative to the (un-)reliability at the latent ability by

$$\hat{\theta}_{\mathrm{EAP}_{\mathrm{adj}}} = \frac{\hat{\theta}_{\mathrm{EAP}}}{\mathrm{rel}\left(\hat{\theta}_{\mathrm{EAP}}\right)} \qquad [4.16]$$

A simulation-based approach is utilized to generate the distribution of this adjusted statistic and to address the problem of inaccurate representations of the standard normal distribution of the applied person fit statistic l_z by normalization formulas described above. For each initial response vector from the original data set, new latent ability values θ_{new} were simulated and a large number of response vectors were generated according to $\hat{\theta}$ and the item parameters (see Chapter 3). For each of these generated response vectors, abilities were again estimated according to the adjusted approach described above and an adjusted null distribution for the person fit statistic was determined. The *p*-value of the person fit statistic was then found as the proportion of response vectors with a person fit value less than or equal to the value of the response vector from the original data. A decision on the fit can be made under a given nominal α-level. De la Torre and Deng (2008) figured out that this approach, though being computationally intensive, was slightly more exact with regard to *p*-values than person fit analysis under alternative Bayesian estimation methods and SNIJ. Conijn et al.

(2014) found the method to be useful for the analysis of multiscale measures with polytomous item scoring.

4.3 Purpose of this study

Referring in particular to the method by de la Torre and Deng (2008), this method is based on the idea to first improve scoring estimates to determine a more accurate representation of the true ability of the respondent, and then – distrusting the adequacy of standardized forms of the statistic – to determine person fit given a likewise adjusted simulated distribution under a nominal Type I error rate. As an alternative to the Bayesian method proposed by de la Torre and Deng (2008), it is reasonable to assume that the combination of bias-reducing weighted ML scoring methods and simulation is also beneficial with regard to recovery of nominal Type I error rates and statistical power of l_z. The conventional ML and the robust HU estimates do not provide finite estimates for scores $r = 0$ or $r = L$, unlike the EAP or the WL. As such response vectors cannot be interpreted meaningfully with regard to model fit, this fact may not degrade the usefulness of person fit analysis based on ML or HU in general. However, de la Torre and Deng (2008) proposed to estimate ability levels for simulated response vectors to determine the reference distribution of l_z which requires finite ability estimates for all response vectors to compute the person fit statistic. Differences between the scoring methods are therefore accommodated in this study by applying two methods for simulating the reference distribution of l_z. Bootstrap variant I (B^I) based on the ML or the HU scoring methods is implemented by the following steps (compare the descriptions by de la Torre & Deng, 2008, Rizopoulos, 2013, or van Krimpen-Stoop & Meijer, 1999):

1. For a fixed examinee, estimate θ by the ML or HU scoring method ($\hat{\theta}_{\mathrm{ML}}$ or $\hat{\theta}_{\mathrm{HU}}$).

2. Compute l_z for each response vector given $\hat{\theta}_{\mathrm{ML}}$, respectively $\hat{\theta}_{\mathrm{HU}}$, and the item parameters.
3. Simulate $m = 1,\ldots, M$ new ability values θ_{new} from $N(\hat{\theta}_{\mathrm{ML}}, \mathrm{SE}[\hat{\theta}_{\mathrm{ML}}])$, respectively $N(\hat{\theta}_{\mathrm{HU}}, \mathrm{SE}[\hat{\theta}_{\mathrm{HU}}])$.
4. Simulate M new response vectors given θ_{new} and the item parameters.
5. Compute l_z for each simulated response vector given θ_{new} and the item parameters ($l_{\mathrm{z}_{\mathrm{new}}}$).
6. Compute the p-value of l_z, as the proportion of $l_{\mathrm{z}_{\mathrm{new}}}$ based on the simulated response vectors less than or equal to l_z based on the original response vector.
7. Decide on person fit using a fixed nominal α-level.

For bootstrap variant II (B^{II}) based on the WL, steps 1-4 and 6-7 were the same as for variant A but the simulation scheme was varied in the following way:

4b. Estimate $\hat{\theta}_{\mathrm{new}}$ by the WL scoring method ($\hat{\theta}_{\mathrm{WL}}$) for each simulated response vector given the item parameters.

5. Compute l_z for each simulated response vector given $\hat{\theta}_{\mathrm{new}}$ and the item parameters ($l_{\mathrm{z}_{\mathrm{new}}}$).

Please note that B^{I} is similar to the methods proposed by Rizopoulos (2013) and van Krimpen-Stoop and Meijer (1999) while B^{II} is similar to the one proposed by de la Torre and Deng (2008). Obviously, estimating $\hat{\theta}_{\mathrm{new}}$ by the WL scoring method for each simulated response vector increases the computational demand of B^{II}.

To investigate the usefulness of these approaches, Type I error (false alarm rate) under model conform data and statistical power to detect aberrancy (detection rates) were explored in two simulation studies. Simulation 1 investigated the usefulness of the proposed methods for the recovery of nominal Type I error rates. Simulation 2 investigated the statistical power of these methods for misfit detection by l_z. A real data large scale assessment example illustrates the results from simulated data.

4.4 Simulation 1: Investigation of Type I error rates

To gain insights into the benefits of the simulation-based approach, person fit analysis by l_z were compared in a simulation with regard to empirical Type I error rates. Results on Type I error under the established $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$ and the SNIJ method were used as benchmark to evaluate $\mathrm{B}^{\mathrm{I}}_{\mathrm{ML}}$, $\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}}$and $\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$. Type I error rates for $\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}}$ were evaluated given five values of the tuning constant: TC_{HU} = 1.4, TC_{HU} = 1.2, TC_{HU} = 1.0, TC_{HU} = 0.8, and TC_{HU} = 0.6.

To compare the accuracy of the several methods with regard to the nominal Type I error rates, the MAD between empirical and nominal Type I error rate was again computed.

4.4.1 Data simulation

Even though the Rasch model is the most typical IRT model and is, for example, often applied in large-scale assessments of competencies (see Chapter 4.6), the 2PLM model is a more flexible model and displays adequate fit in most empirical data sets. The 3PLM (or the 4PLM) may often be over-parameterized and, hence, non-convergence or poorly estimated parameters of the lower (and upper) asymptote may be obtained (for a discussion see Baker & Kim, 2004, Chapter 4). Thus, for the simulation presented here, 2 × 2 × 5 = 20 conditions were varied under a 2PLM. Two test lengths, L = 20 and L = 40, were assumed. This restriction to short and medium size test lengths was made as person fit analysis in short test lengths is the most challenging scenario for person fit statistics, and also considering previous findings that l_z followed the standard normal distribution rather adequately for longer tests (say, 80 items; Drasgow et al., 1985). Item difficulty parameters were selected to be distributed with equal distances in parameter ranges equal to those used in Study 1, [-2, 2] for L = 20 and [-2.75, 2.75] for L = 40. As this simulation focused on the influence of scoring methods on person fit analysis, these item difficulties were treated as known. Discrimination parameters were not randomly sampled as it was done in previous studies. Instead, for a test lengths of L = 20 the same sequence of four different values was replicated in the way that a_1 = 1.15, a_2 = 0.85, a_3 = 1.30, a_4 = 0.70, a_5 = 1.15, a_6 = 0.85, a_7 = 1.30, a_8 = 0.70,…, a_{17} = 1.15, a_{18} = 0.85, a_{19} = 1.30, a_{20} = 0.70 for a test with low mean discrimination (in the following $\bar{a}_i$ = 1.0), and a_1 = 1.65, a_2 = 1.35, a_3 = 1.80, a_4 = 1.20, a_5 = 1.65, a_6 = 1.35, a_7 = 1.80, a_8 = 1.20,…, a_{17} = 1.65, a_{18} = 1.35, a_{19} = 1.80, a_{20} = 1.20 for a test with high mean discrimination ($\bar{a}_i$ = 1.5). For test length L = 40, these sequences were doubled. Please note that the variance of the discrimination parameters is thereby very close to the discrimination parameter variance randomly sampled from N (1.0, 0.25), respectively N (1.5, 0.25). The selected discrimination levels are realistic given that items with lower discrimination, say a_i = 0.5, are usually excluded from IRT-based tests data, and items with higher discrimination, say a_i = 2.0, are rarely found in real data. Response data was simulated under the above mentioned parameters for five ability levels (θ = -2, θ = -1, θ = 0, θ = 1 and θ = 2). Response vectors with r = 0 and r = L were excluded as person fit cannot be evaluated meaningfully for these scores. For l_z based on the computer intensive MC simulation, the number of new response vectors generated in the simulation was M = 600. Due to the high computing time of the MC method underlying each scoring method, the number of response vectors generated for each test length and ability was restricted to N = 5000.

The statistical software R (R Development Core Team, 2011) was used for data generation and analysis. R-Code snippets from the R-packages irtoys (Partchev, 2011), ltm (Rizopoulos, 2013), and the appendices of Schuster and Yuan (2011) as well as Magis et al. (2012) were applied to compute ability estimates, the MC methods and the SNIJ correction.

4.4.2 Evaluation of Type I error rate

Type I error rates of the statistic were obtained as described in Chapter 3. Results are presented for Type I error rates of α = .01, α = .05 and α = .10.

4.4.3 Results

Table 4.1 gives information on the Type I error rate of person fit statistics l_z based on the different simulation-based methods presented above plus SNIJ on three different α-levels. Across all conditions, the empirical Type I error rates of l_z under $\mathrm{B}_{\mathrm{ML}}^{\mathrm{I}}$were substantially deflated, in particular for extreme abilities. With a decrease of the tuning constant (increased downweigthing of aberrant item responses), empirical Type I error rates for $\mathrm{B}_{\mathrm{HU}}^{\mathrm{I}}$ were more in accordance with the nominal rates. The best recovery of nominal Type I error rates for $\mathrm{B}_{\mathrm{HU}}^{\mathrm{I}}$ was found when a rather low value for the tuning constant (TC_{HU}= .80 or TC_{HU}= .60) was selected. In line with results by de la Torre and Deng (2008), nominal α-levels for statistic l_z were well recovered for $\mathrm{B}_{\mathrm{EAP}_{\mathrm{adj}}}^{\mathrm{II}}$. Across each of the test lengths, discrimination and ability levels, results for $\mathrm{B}_{\mathrm{WL}}^{\mathrm{II}}$ were very similar to those obtained for $\mathrm{B}_{\mathrm{EAP}_{\mathrm{adj}}}^{\mathrm{II}}$. Type I error rates for method SNIJ (and the ML scoring method) were inflated under a Type I error rate of α = .01, slightly inflated under a Type I error rate of α = .05 but slightly deflated under a Type I error rate of α = .10. These patterns for SNIJ were similarly found under each discrimination level, test length and across each ability level in the simulation. Previously, de la Torre and Deng (2008) have presented similar findings on method $\mathrm{B}_{\mathrm{EAP}_{\mathrm{adj}}}^{\mathrm{II}}$ and SNIJ.

Figure 4.4 shows the MAD aggregated across two test lengths (L = 20 and L = 40) and two discrimination levels ($\bar{a}_i$ = 1.0 and $\bar{a}_i$ = 1.5) for each of the presented methods and each of the five ability levels. The highest discrepancy between nominal and empirical Type I errors measured by the MAD were found for $\mathrm{B}_{\mathrm{ML}}^{\mathrm{I}}$and for $\mathrm{B}_{\mathrm{HU}}^{\mathrm{I}}$ with high values of the tuning

Table 4.1

Type I error rates for person fit statistic l_z based on eight bootstrap methods and SNIJ

L	$\bar{a}_i$	θ	N	$\mathrm{B}^{\mathrm{I}}_{\mathrm{ML}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.4}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.2}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.0}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{0.8}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{0.6}}$	$\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$	$\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$	SNIJ
α = .01												
20	1	-2	4943	.001	.002	.002	.003	.005	.005	.004	.004	.022
		-1	5000	.004	.007	.007	.009	.009	.013	.010	.010	.015
		0	5000	.007	.008	.009	.010	.012	.014	.012	.013	.016
		1	5000	.003	.007	.008	.011	.014	.017	.012	.013	.020
		2	4926	.001	.003	.003	.005	.007	.008	.008	.006	.025
	1.5	-2	4754	.002	.005	.006	.007	.008	.008	.009	.007	.027
		-1	5000	.005	.009	.011	.011	.013	.016	.011	.012	.021
		0	5000	.007	.008	.009	.011	.011	.013	.012	.012	.017
		1	4999	.004	.008	.010	.011	.013	.013	.012	.011	.018
		2	4749	.000	.003	.004	.005	.007	.008	.009	.007	.028
40	1	-2	5000	.003	.007	.009	.012	.012	.018	.012	.012	.022
		-1	5000	.008	.009	.010	.011	.013	.013	.013	.013	.017
		0	5000	.010	.011	.012	.010	.013	.013	.013	.015	.018
		1	5000	.005	.006	.007	.008	.008	.012	.009	.010	.014
		2	5000	.003	.006	.007	.008	.010	.013	.010	.008	.018
	1.5	-2	4999	.003	.006	.006	.009	.011	.014	.009	.009	.018
		-1	5000	.006	.007	.008	.009	.009	.009	.010	.009	.015
		0	5000	.008	.008	.009	.009	.010	.011	.010	.011	.017
		1	5000	.005	.006	.008	.008	.008	.010	.009	.009	.015
		2	5000	.005	.008	.010	.012	.013	.016	.011	.011	.020
α =.05												
20	1	-2	4943	.008	.024	.028	.032	.035	.045	.045	.041	.062
		-1	5000	.026	.033	.034	.041	.044	.051	.051	.053	.053
		0	5000	.036	.040	.041	.041	.045	.052	.054	.057	.050
		1	5000	.027	.035	.041	.043	.046	.053	.056	.055	.056
		2	4926	.010	.031	.035	.040	.044	.049	.051	.048	.070
	1.5	-2	4754	.015	.034	.035	.042	.045	.050	.051	.046	.071
		-1	5000	.029	.043	.044	.047	.051	.054	.055	.055	.056
		0	5000	.037	.042	.042	.045	.048	.052	.053	.054	.054
		1	4999	.028	.039	.041	.041	.048	.053	.051	.052	.053
		2	4749	.013	.034	.035	.042	.047	.050	.047	.042	.063
40	1	-2	5000	.027	.043	.047	.050	.057	.064	.058	.061	.067
		-1	5000	.035	.041	.042	.044	.046	.053	.051	.053	.053
		0	5000	.046	.049	.048	.050	.050	.054	.056	.057	.057
		1	5000	.039	.044	.045	.048	.050	.054	.055	.055	.055
		2	5000	.020	.034	.037	.042	.050	.057	.046	.047	.052
	1.5	-2	4999	.025	.040	.041	.044	.049	.054	.051	.050	.054
		-1	5000	.034	.038	.040	.040	.043	.047	.048	.047	.048
		0	5000	.042	.043	.044	.045	.047	.050	.052	.053	.053
		1	5000	.037	.042	.042	.041	.044	.046	.048	.050	.049
		2	5000	.027	.040	.045	.047	.053	.056	.053	.053	.057

Table 4.1 continued

α =.10												
20	1	-2	4943	.028	.064	.068	.077	.088	.097	.098	.098	.103
		-1	5000	.064	.075	.080	.086	.094	.103	.110	.109	.094
		0	5000	.079	.084	.085	.088	.093	.102	.105	.112	.091
		1	5000	.062	.072	.078	.083	.087	.098	.105	.106	.094
		2	4926	.036	.067	.071	.081	.085	.094	.101	.099	.110
	1.5	-2	4754	.041	.078	.081	.087	.094	.105	.102	.095	.097
		-1	5000	.069	.081	.085	.088	.092	.099	.104	.107	.090
		0	5000	.081	.086	.089	.091	.095	.102	.107	.108	.093
		1	4999	.066	.079	.086	.088	.094	.100	.104	.103	.087
		2	4749	.037	.079	.085	.092	.098	.105	.106	.101	.108
40	1	-2	5000	.066	.088	.091	.098	.104	.113	.111	.112	.104
		-1	5000	.072	.080	.083	.087	.087	.094	.097	.098	.086
		0	5000	.092	.094	.095	.096	.099	.105	.106	.111	.099
		1	5000	.081	.090	.093	.094	.100	.105	.109	.110	.097
		2	5000	.050	.073	.079	.084	.092	.103	.098	.102	.091
	1.5	-2	4999	.061	.079	.081	.087	.092	.097	.100	.101	.087
		-1	5000	.078	.085	.082	.089	.088	.093	.098	.099	.088
		0	5000	.086	.087	.090	.089	.091	.093	.105	.105	.094
		1	5000	.076	.078	.082	.085	.087	.091	.096	.098	.084
		2	5000	.063	.083	.087	.092	.099	.104	.103	.102	.091

constant. For $\mathrm{B}^{\mathrm{II}}_{\mathrm{HU}}$ with low values of the tuning constant as well as for each of the methods $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$, $\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$ and SNIJ, differences between nominal and empirical Type I errors measured by the MAD were very small with a slight preference for $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$as the best method in terms of nominal Type I error rate recovery.

In summary, the results supported the usage of $\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}}$ with substantial downweighting of aberrant scores in comparison to $\mathrm{B}^{\mathrm{I}}_{\mathrm{ML}}$ to prevent deflation of nominal α-levels, even though the particular value of TC_{HU} to achieve the best recovery of nominal Type I error rates differed slightly depending on the conditions of the simulation (the test characteristics) and the nominal α-levels applied. The results also re-emphasized the ecouraging results for $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$ presented in de la Torre and Deng (2008) and indicated that nominal α-levels were accurately recovered by method $\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$. As previously described by others (de la Torre &

Deng, 2008, Snijders, 2001; van Krimpen-Stoop & Meijer, 1999), SNIJ was found to be a well-defined alternative to the usage of MC simulation except for a small α-level.

4.5 Simulation 2: Investigation of statistical power to detect model violations

In Simulation 2, the statistical power to detect misfit by l_z based on the previously described simulation-based methods was investigated in a simulation study. The scoring methods underlying the computation of l_z were the same as in Simulation 1. Again, the $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$ and the SNIJ method were used as benchmark to evaluate the usefulness of the methods $\mathrm{B}^{\mathrm{I}}_{\mathrm{ML}}$, $\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}}$ (with the same tuning constants as selected in Simulation 1) and $\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$.

4.5.1 Data simulation

To compare the scoring methods with regard to statistical power, 2PLM item response data was simulated for 2 × 2 × 2 × 3 cells in a cross-factorial design. The test lengths were again set to either $L = 20$ or $L = 40$ which represents small to medium size lengths, the mean discrimination levels were again set to $\bar{a}_i = 1.0$ and $\bar{a}_i = 1.5$ displaying lower and higher discriminating power, and the item difficulty and item discrimination parameters as well as the parameters for the MC simulation were selected as in Simulation 1. Two types of misfit, representing spuriously high and spuriously low response vectors, were simulated: Cheating (spuriously high) was simulated by assigning a probability of .90 for a correct response on 20 % of the most difficult items; test anxiety (spuriously low) was simulated by assigning a probability of .25 for a correct response on 20 % of the easiest items. For example, when 20 items were simulated and cheating was induced on the most difficult items, the item parameters from the 17th to the 20th item ($b_{17} = 1.37$, $b_{18} = 1.58$, $b_{19} = 1.79$ and $b_{20} = 2.00$) were set to $P = .90$, irrespective of the original item difficulty and the original ability.

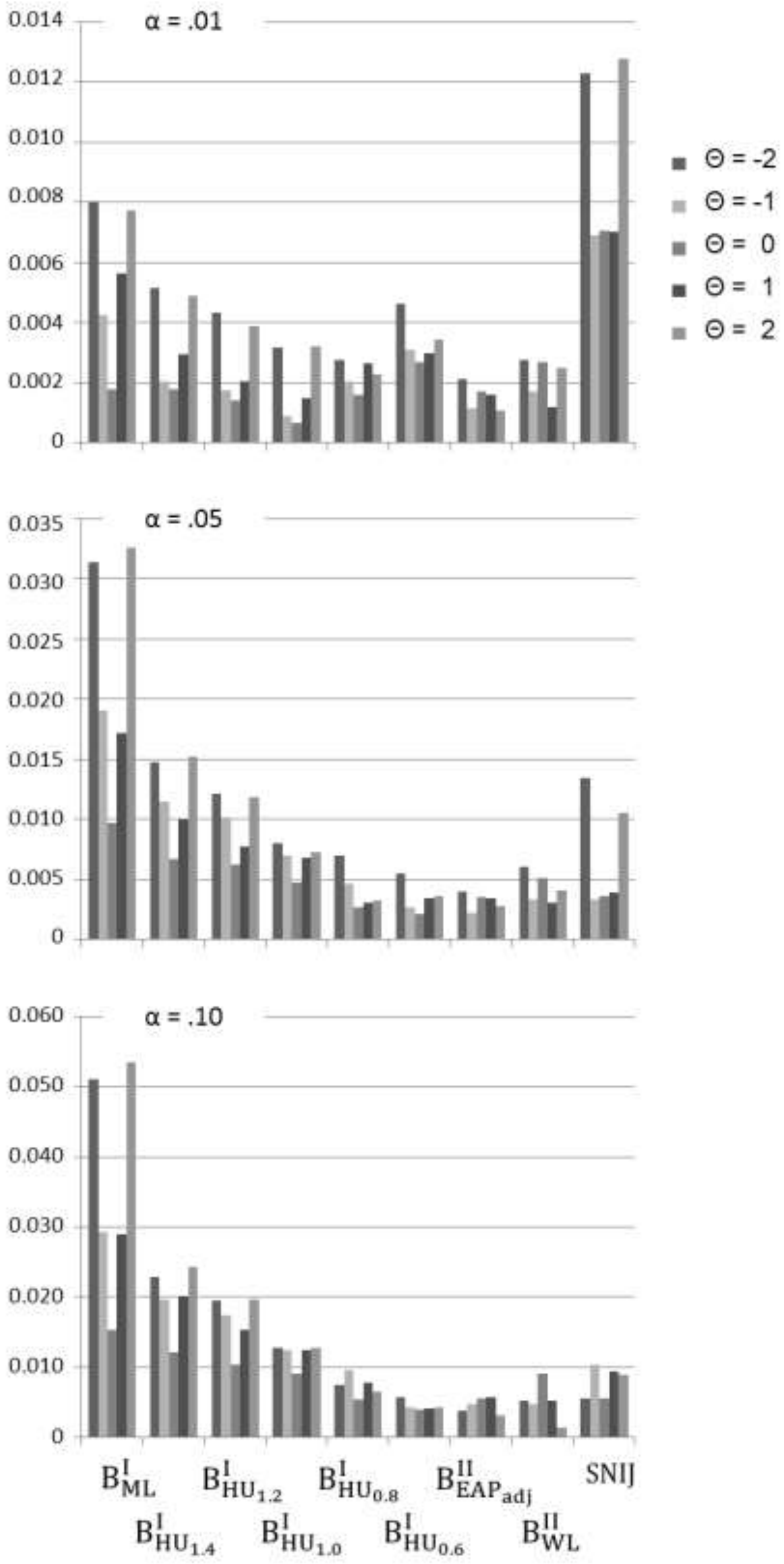

***Figure 4.4*: Mean absolute difference (MAD) between empirical and nominal Type I error for eight bootstrap methods and SNIJ aggregated across two test lengths (*L* = 20 and *L* = 40) and two mean discrimination levels ($\bar{a}_i$ = 1.0 and $\bar{a}_i$ = 1.5)**

For both aberrancy types and both aberrancy levels, three latent abilities were analyzed. As cheating is likely to occur with low ability respondents, the abilities for cheating were $\theta = -2$, $\theta = -1$ and $\theta = 0$; as test anxiety is assumed to impair scores of otherwise competent respondents, the abilities considered for test anxiety were $\theta = 0$, $\theta = 1$ and $\theta = 2$.

4.5.2 Evaluation of statistical power

Statistical power in the *cheating* and *test anxiety* conditions is evaluated as described in Chapter 3. Results are presented for nominal Type I error rates of $\alpha = .01$, $\alpha = .05$ and $\alpha = .10$. Please remember that following de la Torre and Deng (2008), the statistical power for different methods is not comparable in a strict manner as the methods differ in their empirical Type I errors.

4.5.3 Results

Table 4.2 and Table 4.3 provide information on the statistical power of person fit statistic l_z based on the methods presented above in the 24 cells of the simulation. In most conditions, statistical power was slightly higher compared to, for example, the analysis by de la Torre and Deng (2008), which might be attributed to a larger range of item difficulties (e.g., Meijer, Molenaar, & Sijtsma, 1994; Reise & Due, 1991), the absence of a 3PLM guessing parameter (e.g., Meijer & Nering, 1997; Reise & Due, 1991) and a different number of items being affected by misfit. Power generally increased for longer test lengths, higher item discrimination and for extreme compared to medium abilities. Cheating (Table 4.2) was easier to detect than test anxiety (Table 4.3). In the most favorable conditions for misfit detection (higher item number, higher item discrimination and extreme abilities) almost all response vectors were identified by each of the presented methods. Hence, lower nominal

Type I error rates than in Study 1 may be utilized to identify differences in the detection rates of the several methods (which was the reason for including a nominal Type I error rate of α = .01 in this study). But for medium ability levels, tests of length L = 20 were too short to reach acceptable statistical power, say .70 or .80, to detect misfit, particularly when the discrimination level was low.

The following results were found in comparison of the presented methods: The statistical power of $\mathrm{B}_{\mathrm{ML}}^{\mathrm{I}}$ was lowest in all conditions while for $\mathrm{B}_{\mathrm{HU}}^{\mathrm{I}}$ the statistical power increased with decreasing tuning constant of the HU estimate and was similar to $\mathrm{B}_{\mathrm{ML}}^{\mathrm{I}}$ for high values of the tuning constant (like TC_{HU}= 1.20 or TC_{HU}= 1.40), and close to $\mathrm{B}_{\mathrm{EAP}_{\mathrm{adj}}}^{\mathrm{II}}$ and $\mathrm{B}_{\mathrm{WL}}^{\mathrm{II}}$for low values of the tuning constant (like TC_{HU}= 0.60 or TC_{HU}= 0.80). Comparing the methods $\mathrm{B}_{\mathrm{EAP}_{\mathrm{adj}}}^{\mathrm{II}}$, $\mathrm{B}_{\mathrm{WL}}^{\mathrm{II}}$and SNIJ, it was difficult to evaluate which of the three methods performed best across all conditions. Under lower nominal Type I error rates the highest power to detect misfit was found for SNIJ while for higher nominal Type I error rates the statistical power of SNIJ was outperfomed by $\mathrm{B}_{\mathrm{EAP}_{\mathrm{adj}}}^{\mathrm{II}}$ and $\mathrm{B}_{\mathrm{WL}}^{\mathrm{II}}$, two methods hard to differentiate in terms of power. Similar patterns on the statistical power of $\mathrm{B}_{\mathrm{EAP}_{\mathrm{adj}}}^{\mathrm{II}}$ and SNIJ were also found by others (e.g., de la Torre & Deng, 2008) and are expected given that the SNIJ method had inflated empirical Type I error rates under a lower nominal α-levels but deflated Type I error rates under higher nominal α-levels.

4.6 Application to real data: Conclusions on the validity of educational large-scale assessment results for students with disabilities by person fit analysis

To further demonstrate the usefulness of the methods described in this chapter and to investigate implications for the analysis of real data, four samples of students with disabilities participating in the North Rhine-Westphalian educational large-scale assessment program *lernstand8* (Leutner, Fleischer, Spoden, & Wirth, 2007) were analyzed by means of person fit statistics testing for Rasch-conformity of the response vectors. The four data sets included samples of English and German language reading comprehension test data from the 2007 assessment, and two samples of mathematics test data from the 2007 and the 2008 mathematics assessment of *lernstand8*. Ignoring some peculiarities of the North Rhine-Westphalian curriculum at this point, the two reading comprehension tests were constructed rather similar to those reading comprehension tests administered in international large-scale assessments like *Deutsch-Englisch-Schülerleistungen-International* (e.g., Beck & Klieme, 2007) or PISA (OECD, 2009, 2010). The mathematics tests focused on the student's ability to communicate mathematical concepts in the 2007 assessment and on their ability to apply mathematical tools such as pocket calculators, circles, ruler and dynamic geometry software in the 2008 assessment; information on the design of the mathematics assessments including descriptions of the competence scales, sample items and coding manuals are given in Heymann and Pallack (2007) or Spoden, Fleischer, and Leutner (2010).

The students attended special education schools focusing on different types of disabilities, impairments and developmental disadvantages including learning and perceptual disabilities as well as problems in emotional and affective development (Table 4.4). Each of the schools participated voluntarily in the assessment and received comprehensive feedback on the competence structure of their students. Please note that due to protection of data privacy the individual impairment of each student is strictly unknown to the test administrators.

Table 4.2

Statistical power of person fit statistic l_z based on eight bootstrap methods and SNIJ to detect cheating

L	$\bar{a}_i$	θ	N	$\mathrm{B}^{\mathrm{I}}_{\mathrm{ML}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.4}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.2}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.0}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{0.8}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{0.6}}$	$\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$	$\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$	SNIJ
α =.01												
20	1	-2	5000	.839	.899	.905	.910	.918	.922	.924	.924	.951
		-1	5000	.601	.641	.653	.675	.696	.715	.706	.721	.778
		0	5000	.224	.254	.268	.295	.342	.406	.360	.356	.452
	1.5	-2	5000	.983	.991	.990	.990	.990	.991	.992	.992	.996
		-1	5000	.908	.922	.936	.935	.942	.947	.935	.936	.964
		0	5000	.456	.501	.521	.551	.584	.610	.587	.591	.706
40	1	-2	5000	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		-1	5000	.986	.988	.987	.990	.990	.992	.991	.991	.994
		0	5000	.813	.830	.831	.846	.866	.882	.891	.894	.934
	1.5	-2	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		-1	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0	5000	.984	.986	.987	.987	.989	.991	.991	.991	.997
α =.05												
20	1	-2	5000	.969	.979	.980	.982	.982	.983	.986	.986	.986
		-1	5000	.876	.887	.892	.900	.909	.921	.922	.926	.921
		0	5000	.548	.578	.591	.614	.656	.702	.691	.689	.698
	1.5	-2	5000	.996	.997	.997	.997	.997	.997	.998	.999	.998
		-1	5000	.981	.985	.987	.987	.989	.988	.987	.989	.988
		0	5000	.785	.803	.810	.830	.850	.857	.873	.874	.880
40	1	-2	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		-1	5000	.999	.998	.999	.999	.999	.999	1.000	.999	.999
		0	5000	.966	.966	.967	.970	.973	.976	.984	.984	.984
	1.5	-2	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		-1	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0	5000	.999	.999	.999	.999	.999	.999	.999	.999	.999
α =.10												
20	1	-2	5000	.987	.992	.992	.992	.992	.991	.994	.994	.992
		-1	5000	.947	.949	.952	.953	.958	.961	.963	.964	.956
		0	5000	.715	.735	.747	.763	.790	.821	.835	.832	.810
	1.5	-2	5000	.999	1.000	.999	.999	.999	.999	1.000	1.000	1.000
		-1	5000	.993	.994	.995	.995	.995	.996	.995	.996	.994
		0	5000	.901	.905	.909	.918	.928	.933	.943	.942	.930
40	1	-2	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		-1	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0	5000	.990	.989	.990	.991	.991	.992	.996	.995	.995
	1.5	-2	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		-1	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		0	5000	.999	1.000	.999	1.000	1.000	1.000	1.000	1.000	1.000

Table 4.3

Statistical power of person fit statistic l_z based on eight bootstrap methods and SNIJ to detect test anxiety

L	$\bar{a}_i$	θ	N	$\mathrm{B}^{\mathrm{I}}_{\mathrm{ML}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.4}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.2}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{1.0}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{0.8}}$	$\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}_{0.6}}$	$\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$	$\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$	SNIJ
α =.01												
20	1	0	4999	.166	.187	.198	.215	.241	.278	.261	.264	.330
		1	5000	.457	.501	.512	.529	.541	.568	.566	.573	.629
		2	5000	.622	.720	.730	.745	.751	.756	.785	.772	.847
	1.5	0	5000	.352	.381	.390	.412	.435	.454	.447	.455	.557
		1	5000	.760	.797	.805	.809	.815	.818	.807	.810	.859
		2	4997	.900	.936	.936	.936	.939	.937	.946	.942	.964
40	1	0	5000	.641	.656	.661	.656	.677	.695	.716	.726	.781
		1	5000	.918	.930	.931	.935	.940	.945	.935	.932	.953
		2	5000	.973	.983	.985	.984	.985	.985	.983	.985	.990
	1.5	0	5000	.910	.915	.914	.919	.923	.928	.930	.930	.958
		1	5000	.992	.995	.995	.995	.995	.995	.995	.995	.997
		2	5000	.999	.999	.999	.999	.999	.999	.999	.999	.999
α =.05												
20	1	0	4999	.435	.462	.465	.483	.511	.540	.552	.556	.553
		1	5000	.742	.765	.776	.780	.789	.798	.803	.810	.803
		2	5000	.859	.899	.905	.909	.909	.913	.926	.926	.929
	1.5	0	5000	.665	.679	.683	.687	.702	.717	.742	.748	.744
		1	5000	.913	.925	.927	.928	.932	.934	.936	.936	.936
		2	4997	.964	.976	.977	.977	.977	.977	.982	.981	.984
40	1	0	5000	.879	.878	.878	.878	.884	.886	.909	.911	.912
		1	5000	.980	.982	.983	.984	.985	.986	.984	.985	.984
		2	5000	.996	.998	.998	.998	.998	.997	.998	.998	.998
	1.5	0	5000	.982	.980	.983	.983	.984	.984	.987	.987	.988
		1	5000	.999	.999	.999	.999	1.000	.999	1.000	.999	1.000
		2	5000	.999	.999	.999	.999	.999	.999	.999	.999	.999
α =.10												
20	1	0	4999	.607	.617	.627	.638	.652	.674	.700	.704	.671
		1	5000	.849	.861	.865	.868	.874	.881	.888	.891	.871
		2	5000	.929	.946	.950	.951	.949	.950	.964	.962	.959
	1.5	0	5000	.799	.804	.809	.813	.820	.828	.847	.849	.828
		1	5000	.954	.960	.962	.962	.964	.964	.965	.966	.960
		2	4997	.982	.986	.986	.986	.986	.986	.992	.991	.989
40	1	0	5000	.940	.938	.938	.938	.940	.940	.954	.956	.950
		1	5000	.993	.994	.994	.995	.995	.995	.996	.996	.994
		2	5000	.998	.999	.999	.999	.999	.999	.999	.999	.999
	1.5	0	5000	.993	.994	.994	.994	.994	.994	.995	.996	.995
		1	5000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		2	5000	.999	.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000

4.6.1 Research questions and method related to the real data example

As Engelhard (2009) argued the analysis of person fit may be useful in this context to provide some insights on whether instructional differences or modifications of test practice influences students to complete the test and to evaluate whether valid inferences may be drawn on the student abilities in the particular competence area. Engelhard (2009) presented a conceptual framework for item equivalence and fit analysis when assessing the competence of students with disabilities. The analysis presented here is restricted to the identification of individual misfit by person fit statistics as one of the components of this framework.

Methods obtaining an inaccurate recovery of nominal Type I error rates in Simulation 1 (like $\mathrm{B}^{\mathrm{I}}_{\mathrm{ML}}$) were not included in the following analyses. Thus, person fit analysis under the proposed methods – HU scoring (TC_{HU} = 0.8) or WL scoring method in combination with a parametric bootstrap – was contrasted to the $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$ method under a nominal Type I error rate of α = .05. Though several Rasch-specific person fit statistics exist (see Chapter 3), the bootstrap variants illustrated in this chapter represent more flexible methods which can be applied to various types of IRT models. Additionally, in contrast to the nonparametric method for testing hypothesis on person fit based on sampling from marginals of a given data matrix presented in Chapter 3, the methods illustrated in this chapter also facilitate person fit testing under known item parameters estimated in the sample of students without disabilities. Hartig and Frey (2012) emphasized the importance of item difficulties in competence assessments to derive criterion-referenced descriptions and interpretations of the competence scales and to validate the test scores. Concerning the feedback of test results presented to the teachers (e.g., Leutner et al. 2007), qualitative differences in the competence of two students may also be illustrated by contrasting the responses of these students on items of diverging difficulty and different cognitive demands (Hartig & Frey, 2012). However, these benefits of criterion-referenced testing in general and competence assessment in particular are degraded by item bias. Person fit analysis is a method to explore item-invariance by examining whether students respond as expected to the test items (Engelhard, 2009). Following Wright (1984, p. 285) who opposed that item bias might be uniformly present or absent among group members and outlined that "…removing the bias … will have to be done on the individual level of the much more useful person fit analyses…", person fit tests were applied to identify unlikely response vectors under the estimated item parameters which constitute the competence scale. Thus, the research question associated with the data was: Does person fit analysis indicate adequate fit of the response vectors of students with disabilities and does the interpretation of this person fit information depend on the method of choice?

Item parameters for the assessments were estimated under the Rasch model with an overall sample size of about 190,000 students and, following Leutner et al. (2007), sufficiently adequate fit of the model was satisfied according to a weighted mean square item fit statistics in the range of $0.8 - 1.2$ (see also Bond & Fox, 2007, Chapter 12). In each of the four assessments one out of two booklets containing items with lower mean difficulty was administered to the subsamples of students with disabilities. The mean WL, HU and EAP ability estimates are given in Table 4.4 and illustrate that the present analysis of person fit focused on misfit in the lower tail of the ability distributions. Students with extreme raw scores (0, 1, 2, L – 2, L – 1, L) were again excluded from the analysis (see Chapter 3). Table 4.4 also gives the remaining sample size in each of the four assessments.

4.6.2 Results and conclusions from the real data example

Differences in the number of response vectors classified as misfitting by the person fit statistics l_z and the methods $\mathrm{B}_{\mathrm{WL}}^{\mathrm{II}}$, $\mathrm{B}_{\mathrm{HU}}^{\mathrm{I}}$ and $\mathrm{B}_{\mathrm{EAP}_{\mathrm{adj}}}^{\mathrm{II}}$ are presented by means of Venn diagrams. Figure 4.5 (A) shows a Venn diagram of the number of respondents classified as misfitting by the person fit statistics l_z and the three previously mentioned methods for the 2007 English

Table 4.4

Descriptives on student and item sample information from the state-wide administered large-scale assessment data sets for students with disabilities

focus	read ENG 2007		read GER 2007		math 2007		math 2008	
	schools	students	schools	students	schools	students	schools	students
emotional and behavioral development / developmental delay	17	193	18	198	17	192	23	277
hearing impairment	-	-	-	-	-	-	1	7
learning disabilities	1	15	2	20	2	21	2	75
learning disabilities, emotional and behavioral development	1	3	3	50	3	29	1	16
learning disabilities, speech and language impairment	-	-	2	25	2	25	1	5
learning disabilities, speech and language impairment, emotional and behavioral development	-	-	2	24	2	21	3	65
physical development / motor development	2	14	2	14	2	15	5	24
speech and language impairment	7	214	8	240	7	201	9	226
visual impairment	1	10	1	10	1	10	2	10
N	376		533		452		525	
$\hat{\theta}_{\text{WL}}$	-0.64 (1.26)		-0.31 (0.62)		-1.24 (1.26)		-1.33 (1.4)	
$\hat{\theta}_{\text{HU}}$	-0.70 (1.35)		-0.33 (0.67)		-1.35 (1.94)		-1.75 (3.02)	
$\hat{\theta}_{\text{EAP}}$	-0.56 (1.02)		-0.28 (0.56)		-0.87 (0.86)		-0.97 (0.98)	
L	31		30		20		23	

Notes. read ENG 2007 = reading comprehension (English language) 2007 assessment; read GER 2007 = reading comprehension (German language) 2007 assessment; math 2007 = mathematics 2007 assessment; math 2008 = mathematics 2008 assessment; *N* = remaining sample size after exclusion of extreme raw scores (0, 1, 2, L – 2, L – 1, L); $\hat{\theta}_{\text{WL}}$ = WL ability estimates (mean, sd); $\hat{\theta}_{\text{HU}}$ = HU ability estimates (mean, sd); $\hat{\theta}_{\text{EAP}}$ = EAP ability estimates (mean, sd); *L* = item number

language reading comprehension assessment. The highest overall number of response vectors classified as misfitting by any of the three approaches was about 15 % by $\text{B}^{\text{II}}_{\text{EAP}_{\text{adj}}}$. Results demonstrate most of the response vectors to be identified by each of the three methods. However, there was also a substantial amount of response vectors ignored by $\text{B}^{\text{I}}_{\text{HU}}$ but identified by $\text{B}^{\text{II}}_{\text{WL}}$ and $\text{B}^{\text{II}}_{\text{EAP}_{\text{adj}}}$, and also few response vectors ignored by all methods except $\text{B}^{\text{II}}_{\text{EAP}_{\text{adj}}}$.

Figure 4.5 (B) shows a Venn diagram of the number of respondents classified as misfitting by the person fit statistics l_z and the three previously mentioned methods for the 2007 German language reading comprehension assessment. The results differed slighty from those of the first sample: The overall percentage of response vectors classified as misfitting was lower with at most 10 % of the response vectors identified by $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$. The percentage of response vectors ignored by $\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}}$ but identified under $\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$ and especially by $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$ was similar to the first example.

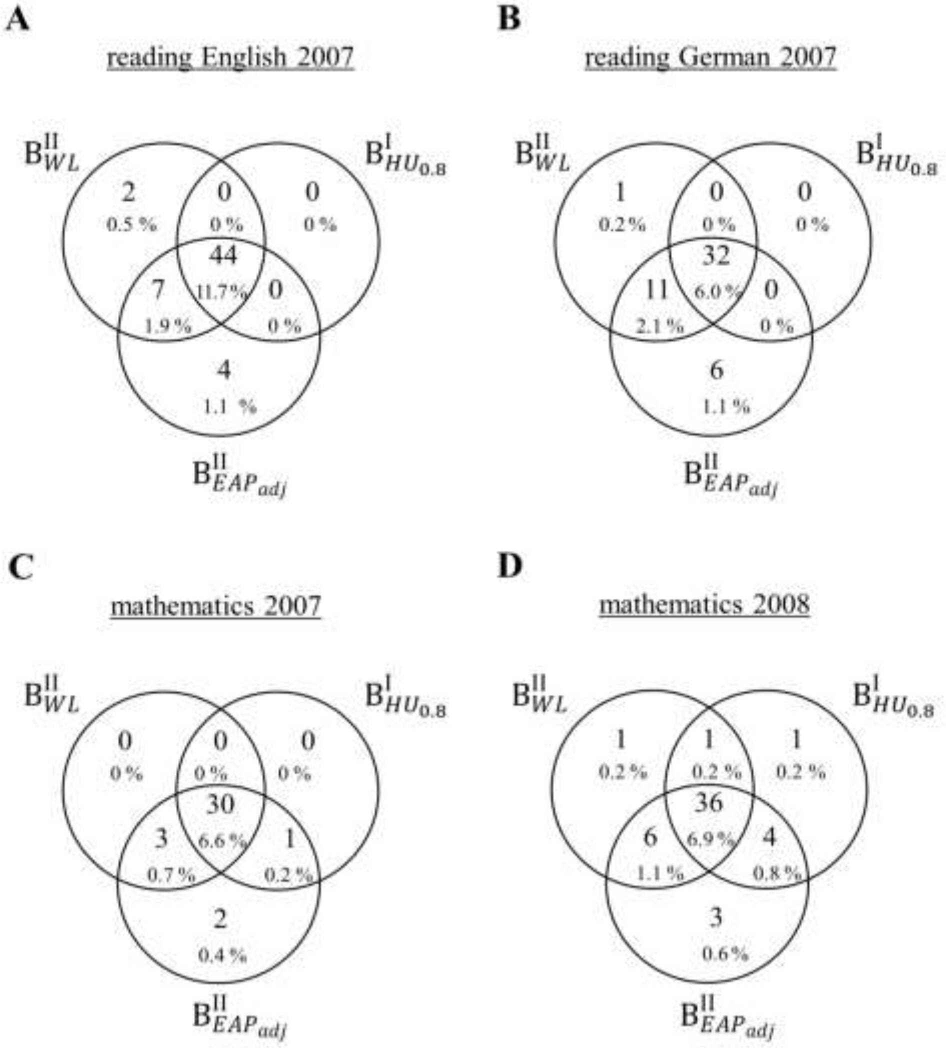
A
reading English 2007
$\mathrm{B}^{\mathrm{II}}_{WL}$
$\mathrm{B}^{\mathrm{I}}_{HU_{0.8}}$
2
0.5 %
0
0 %
0
0 %
44
11.7%
7
1.9 %
0
0 %
4
1.1 %
$\mathrm{B}^{\mathrm{II}}_{EAP_{adj}}$
B
reading German 2007
$\mathrm{B}^{\mathrm{II}}_{WL}$
$\mathrm{B}^{\mathrm{I}}_{HU_{0.8}}$
1
0.2 %
0
0 %
0
0 %
32
6.0 %
11
2.1 %
0
0 %
6
1.1 %
$\mathrm{B}^{\mathrm{II}}_{EAP_{adj}}$
C
mathematics 2007
$\mathrm{B}^{\mathrm{II}}_{WL}$
$\mathrm{B}^{\mathrm{I}}_{HU_{0.8}}$
0
0 %
0
0 %
0
0 %
30
6.6 %
3
0.7 %
1
0.2 %
2
0.4 %
$\mathrm{B}^{\mathrm{II}}_{EAP_{adj}}$
D
mathematics 2008
$\mathrm{B}^{\mathrm{II}}_{WL}$
$\mathrm{B}^{\mathrm{I}}_{HU_{0.8}}$
1
0.2 %
1
0.2 %
1
0.2 %
36
6.9 %
6
1.1 %
4
0.8 %
3
0.6 %
$\mathrm{B}^{\mathrm{II}}_{EAP_{adj}}$

Figure 4.5. **Venn diagrams representing the overlap of the number of item response vectors of students with disabilities from large-scale assessment data classified as misfitting by statistic *l_z* based on three bootstrap methods**

Figure 4.5 (C) shows a Venn diagram of the number of respondents classified as misfitting by the person fit statistics l_z and the previously mentioned methods for the 2007 mathematics assessment. Results illustrate that the number of respondents identified is generally low (at most less than 8 %). There were no substantial differences between the three methods with 30 response vectors identified as misfitting by each of the three methods and at maximum 36 response vectors identified by $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$ (with an overlap of 33 response vectors with $\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$).

Figure 4.5 (D) shows a Venn diagram of the number of respondents classified as misfitting by the person fit statistics l_z and the previously mentioned methods for the 2008 mathematics assessment data set. The maximum percentage of respondents identified as misfitting (by method $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$) was clearly below 10 %. Again, most of the response vectors were identified by each of the three methods but there were also few response vectors ignored by either method $\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}}$ or method $\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$.

Inspection of the item response vectors of these students by person fit statistics allows detecting validity problems underlying their scores in the presence of test modifications and accommodations (e.g., Engelhard, 2009). Summarizing the results obtained from real data, the response vectors from four samples of students with disabilities displayed – contrary to what might be expected given the potential instructional differences and the test modification students with disabilities receive – quite adequate fit to the Rasch model parameters estimated for the complete sample in the large-scale mathematics assessment (about 190,000 students), probably with exception of the 2007 English language reading assessment. Given that the selected methods all displayed rather accurate recovery of nominal Type I error rates in Simulation 1, differences between misfit rates under the three selected methods were generally low in this real data application. Slightly more response vectors were classified as misfitting by $\mathrm{B}^{\mathrm{II}}_{\mathrm{EAP}_{\mathrm{adj}}}$ and $\mathrm{B}^{\mathrm{II}}_{\mathrm{WL}}$ compared to $\mathrm{B}^{\mathrm{I}}_{\mathrm{HU}}$ but independent of the selected simulation method, adequate person fit was diagnosed for more than 90% of the analyzed response vectors in the German reading comprehension and the two mathematics assessment data sets. As an important implication for teachers and educational administrators, these results suggest educational large-scale assessments to provide valid information on the competence of the large majority of participating students with disabilities.

4.7 Remarks on this study

As aberrancy and disturbances in test data are a common finding in psychological and educational assessments, person fit statistics offer a psychometric tool to identify aberrant responses and initiate further inspection of questionable response vectors. The combination of improved trait level estimation – here defined by weighted ML scoring methods – and MC simulation was proposed to facilitate the interpretation of l_z by accurate statistical tests on person fit. Two different variants of simulation-based methods were investigated in simulated data and in real data from a state-wide administered large-scale assessment: The first bootstrap variant characterized particularly by robust HU scoring displayed deflated empirical Type I error rates under weak downweighting of aberrant item responses, and rather adequate recovery of nominal Type I error rates under strong downweighting of aberrant item responses. The appropriate choice of the tuning constant to achieve best recovery of Type I error rates was not the default value for TC_{HU} proposed by Schuster and Yuan (2011) and probably needs to be uncovered depending on the test characteristics. It is a matter of discussion whether item response scoring with strong downweighting of unexpected responses (TC_{HU} = 0.8 and TC_{HU} = 0.6) is generally appropriate. With regard to robust scoring methods, the analyses presented here were restricted to HU as other types of robust scoring (BS, Mislevy and Bock, 1982; AMT robustified Jackknife estimate and WIM estimation scheme[5], Wainer & Wright, 1980) demonstrated drawbacks. For example, prevention of convergence problems of the BS method by Mislevy and Bock (1982) for unexpected response vectors with sparse correct responses is required. Beyond typical IRT models, methods to robustify item and ability estimation like the one described by Bafumi, Gelman, Park, and Kaplan (2005) based on MCMC may originate more accurate estimates underlying the computation of a person fit statistic and may therefore enhance its statistical power.

The second bootstrap variant characterized by Warms (1989) weighted likelihood scoring and additionally estimating θ for simulated response vectors confirmed well-recovered Type I error rates and may serve alternatively to the EAP-based method proposed by de la Torre and Deng (2008). In contrast to the latter method which requires some autonomous programming to correct the primary EAP measures, the WL is directly available from many IRT software packages including ConQuest (Adams, Wu, & Wilson, 2012), Winmira (von Davier, 1997) and several R packages like irtoys (Partchev, 2011). Alternatively, interpolation methods like spline interpolation included in the eRm package (Mair & Hatzinger, 2007) may help to determine finite ability estimates for scores $r = 0$ and $r = L$ which allows this bootstrap variant to be applied based on ML or HU scoring. Limited results from additional analyses available from the author upon request indicate that the second Bootstrap variant generally outperformed the first in establishing an adequate reference distribution (i.e., the distribution of significance probabilities approached the expected uniform distribution adequately; for details on this distribution see van Krimpen-Stoop & Meijer, 1999). In line with previous analyses (e.g., de la Torre & Deng, 2008; van Krimpen-Stoop & Meijer, 1999), results from this study also demonstrated the usefulness of

[5] Mislevy and Bock (1982) used the acronym AMT for the Sine M-estimator and the acronym WIM for an estimation scheme by Benjamin Wright and Ronald Mead.

the correction by Snijders (2001) as an alternative to simulation-based methods considering Type I error recovery and statistical power in most of the studied conditions.

This study focused on scoring methods and followed other studies (e.g., de la Torre & Deng, 2008; Reise, 1995) in applying known item parameters which eliminates error from the item parameters estimation step (see Chapter 1) and assumes more precision than actually exists. This decision was made to concentrate on the effects of scoring methods but results might therefore be biased overoptimistic; for example, statistical power is most likely lower when estimated item parameters are applied. Additionally, differences between different IRT models were ignored and it needs to be further evaluated to which extent aberrant responses affect several IRT models disparately and how person fit statistics perform under these different models. Meijer and Nering (1997) found that detection of misfitting response vectors for low θ levels was higher for 2PLM-conform data compared to 3PLM-conform data which indicates that guessing complicates the detection of other types of misfit. But just the opposite was found for high θ levels, which complicates the interpretation of these results. Please note that person fit tests for the 3PLM (and 4PLM) IRT model were not investigated in the simulations for the reasons outlined in Chapter 4.4 and also to facilitate the comparison of the several methods as the lower (and upper) asymptote are incorporated by computing EAP and WLE estimates but not by computing the robust HU estimates. Also the BILOG-MG software package (Zimowski et al., 1996) referred to in the beginning of this chapter does not incorporate the lower asymptote when computing BS scores (du Toit, 2003). An additional simulation study focusing on a direct comparison of the method by de la Torre and Deng (2008) and the parallel method based on the WL scoring for the 3PLM is currently under way (Spoden, in prep.).

Bearing each of these aspects in mind, researchers and practioners interested in applying person fit statistics will easily identify the main advantage of the presented simulation-based methods; the flexibility of the parametric bootstrap to be applied to many types of IRT models, including multiscale tests and polytomous item formats (Conijn et al., 2014). Restrictions of this flexibility, for example with regard to CAT designs (van Krimpen-Stoop and Meijer, 1999) or to the computationally demanding analyses of large-scale assessment data (see Chapter 5.3), need to be systematically evaluated in future studies. The analysis of real data presented in this study has at least given an illustration of the usefulness of bootstrap person fit tests to examine the equivalence of student measurement across accommodations and test modifications for students with disabilities in a state-wide administered large-scale assessment.

4.8 References

Adams, R., Wu, M., & Wilson, M. (2012). Conquest 3.0. [Computer program]. Melbourne: Australian Council for Educational Research.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*, 171–187.

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Beck, B., & Klieme, E., (Eds.) (2007). *Sprachliche Kompetenzen. Konzepte und Messung. DESI-Ergebnisse Band 1* [Linguistic competencies – construct and their measurement. Results from DESI Volume 1]. Weinheim: Beltz Pädagogik.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental measurement in the human sciences* (2nd Ed.). Mahwah, NJ: Lawrence Erlbaum.

Conijn, J. M., Emons W. H. M., & Sijtsma, K. (2014). Statistic l_z-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement, 38*, 122–136.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159–177.

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

du Toit, M. (Ed.). (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International.

Engelhard, G. (2009). Using item response theory and model data fit to conceptualize differential item and person functioning for students with disabilities. *Educational and Psychological Measurement, 69*, 585–602.

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.

Hartig, J., & Frey, A. (2012). Konstruktvalidierung und Skalenbschreibung in der Kompetenzdiagnostik durch die Vorhersage von Aufgabenschwierigkeiten [Using the prediction of item difficulties for construct validation and model-based proficiency scaling]. *Psychologische Rundschau, 63*, 43–49.

Heritier, S., Cantoni, E., Victoria-Feser, M.-P., & Copt, S. (2009). *Robust methods in biostatistics*. Hoboken, NJ: Wiley-Blackwell.

Heymann, H.W., & Pallack, A. (2007). Aufgabenkonstruktion für die Lernstandserhebung Mathematik [Item construction for state-wide standardized assessments of mathematics learning]. In Ministerium für Schule und Weiterbildung des Landes Nordrhein-Westfalen (Hrsg.), *Lernstandserhebungen Mathematik in Nordrhein-Westfalen. Impulse zum Umgang mit zentralen Tests* (S. 14–46). Stuttgart: Klett.

Leutner, D., Fleischer, J., Spoden, C., & Wirth, J. (2007). Landesweite Lernstandserhebungen zwischen Bildungsmonitoring und Individualdiagnostik [State-wide standardized assessments of learning between educational monitoring and individual diagnostics]. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8*, 149–167.

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

Magis, D., Raîche, G., & Béland, S. (2012). A didactic presentation of Snijders's *l(z)** index of person fit with emphasis on response model selection and ability estimation. *Journal of Educational and Behavioral Statistics, 37,* 57–81.

Mair, P., & Hatzinger, R. (2007b). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software, 20*(9), 1–20.

Maronna, R. A., Martin, R. D., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. New York, NY: Wiley.

Meijer, R. R. (1997). Person-fit and criterion-related validity: An extension of the Schmitt, Cortina and Whitney study. *Applied Psychological Measurement, 21*, 99–113.

Meijer, R. R., Molenaar, I. W., & Sijtsma, K. (1994). Influence of test and person characteristics on nonparametric appropriateness measurement. *Applied Psychological Measurement, 18*, 111–120.

Meijer, R. R., & Nering, M. L. (1997). Ability estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321–336.

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

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*, 725–737.

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

OECD (2004). *Problem Solving for Tomorrow's World. First Measures of Cross-Curricular Competencies from PISA 2003.* Paris: OECD.

Organisation for Economic Co-operation and Development (OECD). (2009). *PISA 2009 Assessment Framework – Key Competencies in Reading, Mathematics and Science*. Paris: OECD.

Organisation for Economic Co-operation and Development (OECD). (2010). *PISA 2009 results: What students know and can do. Student performance in reading, mathematics and science. Volume 1*. Paris: OECD.

Partchev, I. (2011). *Simple interface to the estimation and plotting of IRT models* [Reference Manual]. Retrieved from http://cran.r-project.org/web/packages/irtoys/irtoys.pdf

R Development Core Team (2011). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

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

Reise, S. P. (1995). Scoring methods and the detection of person misfit in a personality assessment context. *Applied Psychological Measurement, 19*, 213–229.

Rizopoulos, D. (2013). *ltm - Latent trait models under IRT: Reference manual* (Ver. .9-7). Retrieved February 2012 from http://cran.r-project.org/web/packages/ltm/ltm.pdf

Schmitt, N., Chan, D., Sacco, J. M., McFarland, L. A., & Jennings, D. (1999). Correlates of person fit and effect of person fit on test validity. *Applied Psychological Measurement, 23*, 41–53.

Schmitt, N. S., Cortina, J. M., & Whitney, D. J. (1993). Appropriateness fit and criterion-related validity. *Applied Psychological Measurement, 17*, 143–15.

Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics, 36*, 720–735.

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

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

Spoden, C., Fleischer, J., & Leutner, D. (2010). Lernstandserhebungen im Fach Mathematik: Zum differenzierten Umgang mit Herausforderungen [State-wide standardized assessments of mathematics learning: handling the challenges]. In Deutscher Verein zur Förderung des mathematischen und naturwissenschaftlichen Unterrichts e.V.: *Lehrerkompetenzen in der Mathematik-Lehrerausbildung*. Verlag Klaus Seeberger.

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

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B. III, Rosa, K., Nelson, L., et al. (2001). Augmented scores—“Borrowing strength” to compute score based on small numbers of items. In D. Thissen, & H. Wainer (Eds.), *Test scoring* (pp. 343–388). Mahwah, NJ: Erlbaum.

Wainer, H., & Wright, B. D. (1980). Robust estimation of aility in the Rasch model. *Psychometrika, 45*, 373–391.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Wright, B. D. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3(1),* 281–288.

von Davier, M. (1997). WINMIRA - program description and recent enhancements. *Methods of Psychological Research – Online, 2 (2)*, 25–28.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). *Bilog MG: Multiple-group IRT analysis and test maintenance for binary* items [Computer program]. Chicago, IL: Scientific Software International, Inc.

5 GENERAL DISCUSSION

In the script at hand, simulation-based approaches for person fit analysis were evaluated. Before discussing implications of the results from the two studies in detail, some consideration concerning the underlying methodological framework is needed. The analyses presented in Chapters 3 and 4 rely heavily on the method of stochastic simulation, and its validity for the analysis of person fit depends on the simulation scheme. Burton, Altman, Royston, and Holder (2006) proposed to apply realistic scenarios and use parameter ranges similar to real data when planning a simulation study. The realization of these recommendations is facilitated by the fact that person fit has been studied intensively in simulated data for several decades (e.g., beginning systematically at the latest with Levine & Rubin, 1979). In the meantime, various simulation schemes have been established. The manifold nature of aberrant response behaviors in testing situations still cannot be simulated; but the analysis of simulated data allows "archetypes" of aberrancy to be studied (e.g., Meijer, 1996) and diagnostically useful decisions to be made regarding the application of these statistics (e.g., Rupp, 2013). The usage of the presented approaches of person fit analysis to psychologically and educationally relevant real data sets completes the analyses and underlines the validity of key findings from the simulation studies.

5.1 Summary of findings

Enhancing the statistical power of person fit analysis (under a given nominal Type I error level) is a constant aim for researchers interested in this psychometric tool. The inaccurate standardization of person fit statistics (which implies either low statistical power to detect misfit or high percentages of incorrectly flagged respondents; e.g., Emons, Meijer, & Sijtsma, 2002; Li & Olejnik, 1997; Molenaar & Hoijtink, 1990; Nering, 1995; van Krimpen-

Stoop & Meijer, 1999) impairs statistical approaches to correctly identify aberrant responding and has therefore been identified as a major challenge in this field of research. This challenge has been addressed in two studies by simulation-based methods sampling data matrices with fixed marginals (Chapter 3) or sampling response vectors under the item parameters and weighted ML ability estimates (Chapter 4).

In Study 1, the Rasch Sampler, an MCMC algorithm for sampling data with given marginals, was applied for simulating the distribution of person fit statistics under the Rasch model and compared to normalized statistics. The results illustrated advantages of the new approach regarding the recovery of nominal Type I error rates and the statistical power (similar or higher compared to the normalized statistics). As several other Rasch-specific person fit statistics exist, it was concluded that the approach is especially useful due to its applicability to determining significance probabilities for about any type of person fit statistic or when person fit statistics are combined with other tests performed to check the underlying assumptions of the Rasch model (for local dependency, multidimensionality, subgroup-invariance, item fit etc.; Mair, Hatzinger, & Maier, 2012).

In Study 2, Type I error and statistical power of person fit statistics based on MC simulation (parametric bootstrap) and weighted ML scoring methods were evaluated by means of two simulation studies. Previously proposed methods for person fit statistics were considered as a benchmark. Results for a first bootstrap variant relying on the robust HU scoring method indicated that robust scoring improves the statistical power but a satisfactory recovery of nominal Type I error rates requires to have the “right touch“ to select the tuning constant in a reasonable way. Results for a second bootstrap variant, designed parallel to the Bayesian method by de la Torre and Deng (2008) but based on the WL scoring method by Warm (1989), were promising with regard to Type I error recovery and statistical power. Compared to the approach presented in Study 1, each of these methods may serve as a more flexible approach for the facilitated interpretation of person fit results for different types of IRT models.

Real data examples from an intelligence test and educational achievement tests further illustrated the usefulness and challenges of person fit statistics based on the Rasch Sampler (Chapter 3) and the parametric bootstrap variants in combination with weighted maximum likelihood scoring methods (Chapter 4). Different results were obtained to some extent depending on whether a conventional or the newly proposed method was used to classify response vectors as fitting or misfitting. Note with regard to the validity of our results that the data sets used for the analyses at hand are typical examples for IRT-based assessments in psychology and educational sciences and that some of the most often applied person fit statistics by psychometricians and educational researchers were used in these real data examples (see Chapter 2).

The outcomes from these two studies contribute to the usage of computer-intensive methods to facilitate the interpretation of person fit analysis. Subsequently, implications of these findings for person fit research and the application of these statistics (primarily in psychological and educational contexts) are outlined.

5.2 Contributions

The results from Studies 1 and 2 (Chapters 3 and 4) presented in this script have documented encouraging results with fairly well-recovered Type I error rates and mostly similar or even higher statistical power compared to other available methods. These results have illustrated MC simulation to improve model fit decisions, even though its application may be problematic under some conditions (see Chapter 5.3). The usage of MC simulation based on the Rasch Sampler (Chapter 3) has not been analyzed before in their implications for person fit analysis. The application of the Rasch Sampler algorithm for this purpose especially widens the options for the interpretation of powerful unstandardized (non)parametric person fit statistics identified by Karabatsos (2003), but the approach may also be applied to mostly any kind of either parametric or nonparametric person fit statistics (see Chapter 3). Additionally, the analysis of person fit with *p*-values determined based on the Rasch Sampler may easily be combined with other statistical tests on Rasch model fit and is well-grounded in the theory of Rasch modeling (e.g., Ponocny, 2001), which might be an important feature for users interested in obtaining “objective measures”.

The major advantage of the proposed approaches in Study 2 (Chapter 4) is their flexibility to be applied to many types of IRT models. Though previous studies gave some indication on results of person fit analysis based on the WL (van Krimpen-Stoop & Meijer, 1999) as well as the robust BS scoring method (Reise, 1995; Meijer & Nering, 1997) and also considered parametric bootstrapping (van Krimpen-Stoop & Meijer, 1999), these studies were limited as differences between bootstrap person fit tests depending on the underlying scoring method were not investigated and, in addition, the bootstrap approach was analyzed in a rather simplified test data design. De la Torre and Deng (2008) had presented a promising approach based on Bayesian scoring methods but had excluded ML scoring methods; this method was extended to ML methods by the results from Study 2. The study also integrates methods from the growing field of robust methods in statistics and psychometrics (e.g., Bafumi, Gelman, Park, & Kaplan, 2005; Magis & De Boeck, 2012).

Finally, the implementation in user-friendly software is a major advantage of the approaches presented in Chapters 3 and 4 compared to other recently developed methods. As Rupp (2013, p. 27) has noticed, researchers and practitioners are often forced to apply a bunch of different isolated programs to perform person fit analyses. This does not apply to the proposed methods implemented in the R (R Development Core Team, 2013) software.

The approaches presented here may also easily be interpreted and communicated to practitioners, at least with regard to basic principles.

5.3 Limitations

Various factors affect Type I error and statistical power of person fit statistics (see Chapter 2); any simulation study therefore needs to focus on a limited number of conditions to evaluate the properties of person fit statistics. The simulations in this script were limited to model-conform data (to evaluate Type I error rates) and two types of aberrancy (to study statistical power to detect misfit). As outlined above, this small number of conditions does not reflect the manifold nature of disturbances assumed to be found in psychological and especially in educational data (Haladyna, 2004; Meijer, 1996), but both of these misfit types were inspired by what was previously assumed to be realistic simulation schemes (e.g., de la Torre & Deng, 2008; Levine & Rubin, 1979; Meijer & Nering, 1997) and were studied in a large number of different conditions. Thus, a reasonable empirical basis was established to derive recommendations on the usage of the proposed methods. Also the concentration on few prominent parametric statistics out of a large variety of indices (Karabatsos, 2003) belongs to the restrictions imposed to cover typical scenarios for person fit analysis. This decision was made in reference to the documented power of the statistics and its prevalence in research and software packages (see Chapter 2). Various statistics previously found to be extremely powerful suffer from drawbacks. For example, cut values for fit / misfit classification by four out of five of the best performing person fit statistics in Rasch-scaled data are not clear (Karabatsos, 2003).

Though the results presented in this script illustrate the usefulness of MC simulation, person fit analysis may not always benefit from this method to obtain correct *p*-values. For example, van Krimpen-Stoop and Meijer (1999) did not find MC simulation for person fit analysis to correctly approximate the theoretical (uniform) distribution when being applied to CAT data. At the current stage of computational speed, usage of MC simulation may remain causing problems in educational large-scale assessments as one of the main areas of application for IRT models (e.g., Rudner, Bracey & Skaggs, 1996). Item parameter estimation for large-scale assessments is generally computer-intensive due to high sample sizes, incomplete designs (multi-matrix designs) or several covariates being included in (latent) regression models. The current implementation of the Rasch Sampler is not feasible for such samples by reasons of limitation to complete matrices of size 4096 × 128 (Verhelst, Hatzinger, & Mair, 2007). The parametric bootstrap based on item parameters and ability estimates (Chapter 4) may – depending on the implementation carried out and the size of the simulated reference distribution – cause high computation times. Though likewise computationally intensive, it needs to be further evaluated whether Bayesian MC simulation might be beneficial under some conditions as the relevant parameters are already sampled in the MCMC iteration process (e.g., Glas & Meijer, 2003). Conservative tests were found in a Bayesian approach to person fit based on MCMC and posterior predictive checks (de la Torre & Deng, 2008; Glas & Meijer, 2003). Whether these can be adapted for accurate *p*-values either by “calibration” of posterior predictive checks (Steinbakk & Storvik, 2009) or by other types of adjustments like the one proposed by de la Torre and Deng (2008) for Bayesian estimators needs to be investigated.

5.4 Recommendations for future research

Parameter fine-tuning of the MC simulation methods has not been discussed in the two simulation studies presented in Chapter 3 and 4. The accuracy and computational burden of the presented methods depend on the form and size of the simulated reference distribution. It might prove beneficial to investigate the assumptions underlying this simulation (Rizopoulos,

2013) and the minimum required sample size as rules-of-thumbs are currently predominant (de la Torre & Deng, 2008; Glas & Meijer, 2003).

Fine-tuning should also be reviewed concerning the robust scoring method presented in Simulation Study 2 where a predefined residual measure and a few different values of the tuning constant (see Chapter 4) were selected to compute the HU scores as a useful choice out of a selection of robust scoring methods (e.g., Mislevy & Bock, 1982; Schuster & Yuan, 2011; Wainer & Wright, 1980). Future studies may include a more thorough systematic experimental variation of these parameters under different robust scoring methods to evaluate the influence of the choice of these parameters on the bias of latent ability estimates or the statistical power of person fit statistics.

Referring to large-scale assessments, Brown and Villareal (2007) proposed weighting respondents according to a person fit statistic when concentrating on aggregated score reports for relevant subgroups like different countries, different school tracks or different gender; limited studies had previously demonstrated that misfit may indeed be unequally distributed across such groups (e.g., Meijer & de Leeuw, 1993; Meijer & van Krimpen-Stoop, 2001). Person fit analysis for this purpose has been ignored with regard to the two large-scale assessment data sets presented in this script and it needs to be evaluated whether aggregated scores differ substantially depending on which of the scoring methods used in Study 2 is selected to compute the person fit statistic.

As an alternative to simulation-based methods, the Snijders' (2001) correction has given promising results in Study 2. This method can be applied to many parametric person fit statistics for dichotomous items satisfying the "general form" described in Equation 1.18. It involves a correction of the first two moments of the statistic but the statistic is still skewed after correction which has consequences especially for short tests (de la Torre & Deng, 2008; van Krimpen-Stoop & Meijer, 1999). Modifications on Snijders' (2001) method may include a correction of skewness (e.g., von Davier & Molenaar, 2003) and may provide users with accurate *p*-values based on the normal distributions even for short tests and also for smaller α-levels. For polytomous items however, it is not possible to write Drasgow, Levine and Williams' (1985) generalization of statistic l_z in the form of Equation 1.18 required for the Snijders correction (van Krimpen-Stoop & Meijer, 2002, p. 167). MC simulation is certainly the method of choice for test administrations containing polytomous item scoring (e.g., Conijn, Emons & Sijtsma, 2014).

Recommendations for future research may also include the application of subcomponents of the analyzed methods to other psychometric tools. For example, as person fit and item fit analysis only differ by the dimension of the data matrix (either rows or columns) being analyzed (e.g., Reise, 1990), it seems likely to discuss implications for item fit analysis. Item fit test statistics based on the Rasch Sampler have already been implemented in R (Mair et al., 2012) but were – to the knowledge of the author – not been intensively evaluated (but see Koller, 2010, for other types of nonparametric tests based on the Rasch Sampler). Whether the computation of robust measurement models (e.g., outlined in the model by Bafumi et al., 2005, pp. 178-179) or the application of robust scoring methods has any substantial impact on item fit needs to be systematically investigated by means of simulation.

5.5 Implications for practitioners

The availability of flexible and free software packages to execute the statistical procedures is of special importance for its application. Note that R-packages exist to initiate the Rasch Sampler presented in Chapter 3. A parametric bootstrap procedure for person fit analysis, similar to the one proposed in Chapter 4 and based on a conventional scoring method, is already implemented in the R-package 'ltm' (Rizopoulos, 2006). Also WL estimates applied in Chapter 4 are available in several R-packages and software code for the robust ability estimates is presented by Schuster and Yuan (2011). When person fit statistics are used, practitioners may also take advantage of the flexibility of software like R (R Development Core Team, 2013) which allows programming and modification of such methods with reasonable effort, thereby enhancing computational speed and receiving reliable inferences in shorter time periods. Limitations concerning certain areas of application (e.g., CAT or large-scale assessments) have been described above.

One of the key questions by practitioners related to person fit analysis is not empirically tractable, particularly, what to do when misfitting response vectors have been identified (e.g., Rupp, 2013). According to the suggestions by Smith (1985, p. 434) test administrators may then decide to (1) report different ability estimates for each subtests or subdimensions (which probably demonstrate higher model conformity), (2) first eliminate questionable item responses (e.g., from unreached items at the end of a booklet) and then adjust the response vector before the ability is re-estimated, (3) not report abilities and retest the individual, or (4) accept negligible error in the ability estimation process (whether the error really is negligible may be indicated according to the standard error associated with $\hat{\theta}$). Smith (1985) also offered the usage of robust scoring methods as a fifth option but proposed to combine maximum likelihood ability estimation and person fit analysis as using "…a robust estimator can mask important information about the person being measured, e.g., a particular content area in which the individual is deficient.“ (p. 434).

It certainly depends on the testing context which of these actions should be realized (Meijer & Sijtsma, 2001). Reporting different $\hat{\theta}$s or retesting respondents may go along with difficulties in communicating the results to practitioners and respondents not familiar with measurement issues; additionally, retesting is not always possible. Eliminating questionable item responses needs justification. For example, eliminating responses to items at the end of a booklet is questionable as long as the test is not speeded or long and stressful. It might also include numerous scalings to be run until an “appropriate” sample of item responses is calibrated. And accepting negligible error in the $\hat{\theta}$ estimation is only adequate in low-stakes testing.

Rupp (2013, pp. 31-32) argued to directly model the influence of aberrancy as part of a more flexible mixture IRT model where respondents are assigned into different classes according to their response patterns. This approach may reduce the number of aberrant response patterns but does not prevent aberrancy from the model predictions as the number of different classes is usually low compared to the several types of misfit often found in test or questionnaire data. Response vectors may therefore also deviate from the class-specific model parameters (von Davier & Molenaar, 2003).

A comprehensive analysis of person fit goes beyond the simple identification of aberrancy but includes specifying the type of misfit and localizing its origin (Emons, Sijtsma, & Meijer, 2004, 2005), as well as considering collateral information, either on a qualitative basis (Meijer, Egberink, Emons, & Sijtsma, 2008) or by regressing misfit on covariates (Conijn, Emons, van Assen, & Sijtsma, 2011; Reise, 2000; Woods, 2008; Woods, Oltmanns, & Turkheimer, 2008). Cui and Leighton (2009) argued that

> “To find the actual causes of misfits, additional information about students’ response processes, such as students’ verbal reports, eye tracking information, and reaction time … is needed. This type of information provides relatively detailed pictures of how students actually solve items on tests, which has the potential to help understand the reasons for misfits so that the results from person-fit statistics can be interpreted substantially and meaningfully." (p. 447).

Combining person fit with collateral information like those outlined above by Cui and Leighton (2009) provides insights which response vectors should be diagnosed in more detail (see also Emons, Glas, Meijer, & Sijtsma, 2003, p. 476; Meijer, 2003, p. 85) and which respondents probably need to be retested.

5.6 References

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*, 171–187.

Brown, R. S., & Villareal, J. C. (2007). Correcting for person misfit in aggregated score reporting. *International Journal of Testing, 7*, 1–25.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*, 4279–4292.

Conijn, J. M., Emons W. H. M., & Sijtsma, K. (2014). Statistic l_z-based person-fit methods for noncognitive multiscale measures. *Applied Psychological Measurement, 38*, 122–136.

Conijn, J. M., Emons, W. H. M., van Assen, M. A. L. M., & Sijtsma, K. (2011). On the usefulness of a multilevel logistic regression approach to person-fit analysis. *Multivariate Behavioral Research, 46*, 365–388.

Cui, Y., & Leighton, J. P. (2009). The hierarchy consistency index: Evaluating person-fit for cognitive diagnostic assessment. *Journal of Educational Measurement, 46*, 429–449.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement, 45*, 159–177.

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

Emons, W. H. M., Glas, C. A. W., Meijer, R. R., & Sijtsma, K. (2003). Person fit in order-restricted latent class models. *Applied Psychological Measurement, 27*, 459–478.

Emons, W. H. M., Meijer, R. R., & Sijtsma, K. (2002). Comparing simulated and theoretical sampling distributions of the U3 person-fit statistic. *Applied Psychological Measurement, 26*, 88–108.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2004). Testing hypotheses about the person response function in person-fit analysis. *Multivariate Behavioral Research, 39*, 1–35.

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

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.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, NJ: Lawrence Erlbaum Associates.

Karabatsos, G. (2003). Comparing the aberrant response detection performance of thirty-six person-fit statistics. *Applied Measurement in Education, 16*, 277–298.

Koller, I. (2010*). Item response models in practice: Testing the Rasch model in small samples and comparing different models for measuring change*. (Unpublished doctorial dissertation). Alpen-Adria-Universität Klagenfurt: Klagenfurt.

Levine, M. V., & Rubin, D. B. (1979). Measuring the appropriateness of multiple-choice test scores. *Journal of Educational Statistics, 4*, 269–29.

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

Magis, D., & De Boeck, P. (2012). A robust outlier approach to prevent Type I error inflation in DIF. *Educational and Psychological Measurement, 72*, 291–311.

Mair, P., Hatzinger, R., & Maier, M. J. (2012). *Package 'eRm'. Reference Manual*. (Ver. .15-1). Retrieved September 2012 from http://cran.r-project.org/web/packages/eRm/eRm.pdf

Meijer, R. R. (1996). Person-fit research: An introduction. *Applied Measurement in Education, 9*, 3–8.

Meijer, R. R. (2003). Diagnosing item score patterns on a test using item response theory-based person-fit statistics. *Psychological Methods, 8*, 72–87.

Meijer, R. R., & de Leeuw, E. D. (1993). Person fit in survey research: The detection of respondents with unexpected response patterns. In J. H. L. Oud, & R.A.W. van Blokland-Vogelesang, *Advances in longitudinal and multivariate analyses in the behavioral sciences* (pp. 235–245). Nijmegen: ITS.

Meijer, R. R., Egberink, I. J. L., Emons, W. H. M., & Sijtsma, K. (2008). Detection and validation of unscalable item score patterns using item response theory: An illustration with Harter's self-perception profile for children. *Journal of Personality Assessment, 90*, 227–238.

Meijer, R. R., & Nering, M. L. (1997). Ability estimation for nonfitting response vectors. *Applied Psychological Measurement, 21*, 321–336.

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

Meijer, R. R., & van Krimpen-Stoop, E. W. L. A. (2001). Person fit across subgroups: An achievement testing example. In A. Boomsma, M. A. J. van Duijn, & T. A. B. Snijders (Eds.), *Essays on item response theory* (pp. 377–390). New York, NY: Springer.

Mislevy, R. J., & Bock, R. D. (1982). Biweight estimates of latent ability. *Educational and Psychological Measurement, 42*, 725–737.

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

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

Ponocny, I. (2001). Non-parametric goodness-of-fit tests for the Rasch model. *Psychometrika, 66*, 437–459.

R Development Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Reise, S. P. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Maesurement, 14*, 127–137.

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

Reise, S. P. (2000). Using multilevel logistic regression to evaluate person-fit in IRT models. *Multivariate Behavioral Research, 35*, 543–568.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17*(5), 1–25.

Rizopoulos, D. (2013). *ltm - Latent trait models under IRT: Reference manual* (Ver. .9-7). Retrieved February 2012 from http://cran.r-project.org/web/packages/ltm/ltm.pdf

Rudner, L. M., Bracey, G., & Skaggs, G. (1996). The use of a person-fit statistic with one high quality achievement test. *Applied Measurement in Education ,9*, 91–109.

Rupp, A. A. (2013). A systematic review of the methodology for person fit research in Item Response Theory: Lessons about generalizability of inferences from the design of simulation studies. *Psychological Test and Assessement Modeling, 55*, 3–38.

Schuster, C., & Yuan, K.-H. (2011). Robust estimation of latent ability in item response models. *Journal of Educational and Behavioral Statistics, 36*, 720–735.

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

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

Steinbakk, G. H., & Storvik, G. O. (2009). Posterior predictive "p"-values in Bayesian hierarchical models. *Scandinavian Journal of Statistics, 36*, 320–336.

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

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

Verhelst, N. D., Hatzinger, R., & Mair, P. (2007). The Rasch sampler. *Journal of Statistical Software, 20(4)*, 1–14.

von Davier, M., & Molenaar, I. (2003). A person-fit index for polytomous Rasch models, latent class models, and their mixture generalizations. *Psychometrika, 68*, 213–228.

Wainer, H., & Wright, B.D. (1980). Robust estimation of aility in the Rasch model. *Psychometrika, 45*, 373–391.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Woods, C. M. (2008). Monte Carlo evaluation of two-level logistic regression for assessing person fit. *Multivariate Behavioral Research, 43*, 50–76.

Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2008). Detection of aberrant responding on a personality scale in a military sample: An application of evaluating person fit with two-level logistic regression. *Psychological Assessment, 20*, 159–168.

Der Lebenslauf ist in der Online-Version aus Gründen des Datenschutzes nicht enthalten.