Psychological Test and Assessment Modeling, Volume 57, 2015 (1), 3-36

Pairwise Rasch model item parameter recovery under sparse data conditions

Jörg-Henrik Heine[1] & *Christian Tarnai*[2]

Abstract

In social science research the occurrence of missing values is a prevalent issue. In addition to that the use of multi-level response formats often results in low cell frequencies, especially under the condition of higher proportions of missing values and small datasets. This in turn may lead to problems for parameter identification in probabilistic item response models, applying the Conditional Maximum Likelihood (CML) or Marginal Maximum Likelihood (MML) method. While *listwise deletion* or *pairwise deletion* usually results in a significant reduction of the sample size and may thus impair efficiency, most methods of data imputation require preceding assumptions about the data loss mechanism, which often can not be checked. Consequently an alternative non-iterative approach to item parameter recovery is introduced. This will be outlined using a minimal example. This approach, named PAIR, is based on conditional pairwise item category comparisons. In the present study data from the NEO-PI-R inventory were analyzed using three different algorithms for item parameter recovery (PAIR, CML, MML) each under a continuing rise of missing data. The resulting item parameter estimates are compared with regard to their accuracy of the point estimates as well as the size of their respective standard errors. The results are discussed comparatively under the condition of the increasing proportion of missing values by simulation. The results suggest that even at higher levels of missing data the PAIR approach leads to stable item parameter estimates.

Keywords: Rasch model, item parameter recovery, pairwise comparisons, least squares, pairwise

[1] *Correspondence concerning this article should be addressed to:* Jörg-Henrik Heine, Technische Universität München, TUM School of Education, Arcisstr. 21 | D-80333 München, Germany; email: joerg.heine@tum.de

[2] Universität der Bundeswehr München

Introduction

The objective of this paper is twofold. First, a short overview of different approaches to item parameter recovery in the framework of Item Response Theory (IRT) is given placing an emphasis on parameter estimation in the presence of missing data. Also an alternative approach to identifying the item parameters is introduced. This approach is the explicit calculation of item parameters on the basis of conditional item category frequencies, which are obtained through a pairwise comparison task. For better understanding the historical background of the method is briefly outlined as well as its relation to the fundamental assumptions of the Rasch model.

Secondly, the pairwise method is tested in comparison with other methods of item parameter recovery. For this a dataset comprising of $n = 620$ students answering to eight items of one personality facet from the German NEO-PI-R inventory (Ostendorf, 2004) is analyzed. To demonstrate the performance of the pairwise method in the context of missing values a minimal simulation scenario, based on the empirical dataset is constructed. The computations following the pairwise comparison approach were conducted using the `R` (R Development Core Team, 2014) package `pairwise` (Heine, 2014) and two other standalone software packages commonly used for estimation of the Rasch model in social sciences.

According to these two objectives, this article is organized as follows. The first section gives a general overview including some theoretical and historical aspects of methods of parameter estimation in the framework of Item Response Theory (IRT). As a rejoinder the principles of applying pairwise comparisons are derived from the basic equations given by Rasch (1960) following an approach first formalized by Choppin (1968). Further a minimal example is illustrated to demonstrate the basic principles of pairwise item parameter recovery from a practical perspective.

An "empirical section" covers the second objective in order to test the pairwise method under practical and simulated conditions. Thus the item parameters in the framework of IRT are recovered for empirical data containing no missing values at baseline first, using different methods of estimation. Thereupon artificial missing values are added to the complete baseline dataset in several steps, estimating the item parameters at every stage of missing data percentage.

Theoretical framework and history of pairwise comparisons

Since the basic formulation of the probabilistic test model by the Danish mathematician Georg Rash and its extension to multi-level, ordinal response formats by Masters (1982), several estimation methods for parameter recovery have been developed and proposed. With regard to the practical application of the model, each of these different methods have certain advantages and disadvantages. As a consequence each method must be carefully considered in each case of application, depending on the different objectives of empirical studies, their designs and the kind of inference to draw from their results. Next, a short outline to several estimation methods in the framework of Item Response Theory (IRT) is given, that are most prevalent in current applications in social sciences. The article then goes on to discuss some issues about the structure of the data matrices in general and the missing data problem in particular, concerning the choice of either method of parameter estimation in IRT. Lastly the method of pairwise comparisons as a non-iterative method of item parameter recovery is introduced, beginning with some remarks on the historical origins of this method and its parallels to measurement according to the Rasch model.

Methods of parameter estimation in Rasch models

As stated by Johnson (2007) there are basically four estimation methods for (item) parameter recovery in the framework of IRT, commonly used in social sciences. The Joint Maximum Likelihood (JML), Conditional Maximum Likelihood (CML), Marginal Maximum Likelihood (MML) and bayesian estimation with Markov Chain Monte Carlo algorithm (MCMC). Discussing several estimation methods for the Rasch model with regard of their precision and accuracy Linacre (1999) generally classified the parameter estimation methods into iterative and non-iterative methods. Following this outline, parameter estimation methods as implemented in prevalent software packages commonly used in social sciences, such as `WINMIRA` (CML) by von Davier (2001), `ConQuest` (MML) by Wu, Adams, Wilson, and Halda (2012), `R`-packages like `MixRasch` (JML) introduced by Willse (2011) and `eRm` (CML - Mair & Hatzinger, 2007), can be assigned to the iterative methods. Parameter recovery in those methods is based on maximizing the likelihood of the margins (model parameters) given the empirical data, in an iterative algorithm – usually a Newton-Raphson type (Linacre, 2004). While in JML the estimation of item and person parameters is conducted jointly, in CML and MML the structural parameters (items) are estimated in a separate step by conditioning out the incidental parameters (persons) in the first estimation step. It is usually recognized, that CML and MML lead to unbiased and consistent parameter estimates while JML has routinely been criticized for inconsistency due to the *Incidental Parameter Problem* (Neyman & Scott, 1948). Concerning this inconsistency it is usually argued that both model parameters, though assumed in JML, are not symmetrical in their nature. However, Linacre (2004) argued that consistency and unbiasedness for CML and MML estimates holds only under certain conditions, which is rarely the case in practical applications. With regard to

CML estimation Linacre (2004) pointed out that results yield to consistent estimates only when extreme person score vectors (zero and perfect raw scores) are excluded from data contributing to the likelihood, which is to be maximized. Thus, by focusing on the discussion of estimation bias in JML (e.g. Wright, 1988), bias of other iterative estimation algorithms simply may have been overlooked.

In addition to these widespread and commonly discussed iterative methods, focusing on item parameter calibration, some non-iterative methods have to be mentioned. For complete data Cohen (1979) solved the iterative maximum-likelihood equations approximately and thus developed a procedure, named PROX, which can explicitly be solved – even by hand (Wright & Masters, 1982). Despite its appealingly simplicity, there is one major downside to this approach. It strongly recourses on normal distribution for both, items as well as persons. Within the framework IRT item parameter recovery, another approach has to be mentioned which splits into an iterative and a non-iterative branch. That is using pairwise comparisons of item category frequencies. This approach has already been mentioned by Rasch (1960, p. 172), suggested to him by Gustav Leunbach, although only in the form of theoretical considerations. It was Choppin (1968, 1985) who further developed it into practical application in the context of item calibration within item banks. Choppin (1985) developed Rasch's suggestion into two techniques that can be used for paired comparisons to estimate item difficulties (1) an iterative maximum likelihood approach and (2) a explicit non-iterative approach resulting in least squares estimates for item parameters (Mosteller, 1951; Garner & Engelhard, 2000). While the maximum likelihood approach of pairwise comparisons received quite some attention in literature about the Rasch model estimation (e.g. Andrich & Luo, 2003; Zwinderman, 1995; Wright & Masters, 1982), the non-iterative approach did not. However, the non-iterative approach can be very appealing as it is rather clear and shows parallels to the fundamental model assumptions formulated by Rasch (1960). Its numerical simplicity, which is illustrated in more detail later in this paper, is just another benefit of this approach. Furthermore it holds the special advantage to handle incomplete data matrices without any hassle, based on its numerical technique.

The missing data problem

Many datasets in social sciences lack completeness. According to Schnell (1997) there has been an increase in missing rates in social surveys in the last decades. When analyzing such data researchers are routinely faced with the question of how to deal with such missing data points in their data matrix under study. Following Schafer and Graham (2002), the core problem with missing data is, that most methods for data analysis are not designed for dealing with incomplete data in an appropriate manner. In some early conceptualizations regarding the phenomena of missing data structures by Rubin (1976), three basic missing data mechanisms can be distinguished. *Missing completely at random* (MCAR), *missing at random* (MAR) and *missing not at random* (MNAR) – for a comprehensive discussion of this issue see also Little and Rubin (2002). These different mechanisms have to be taken into account when preparing and analyzing data. Focusing on estimation of the data in the framework of IRT, the presence of missing data requires a decision about an appropriate algorithm for parameter estimation taking into account the missing data mechanism (Mislevy & Wu, 1988). Thus an adequate treatment of missing data, in terms of using the "right" estimation algorithm, might depend on the kind of the test, the study design and the kind of inference to draw from the estimation result and last but not least on the assumption of the underlying mechanism of missing data. In the framework of IRT scaling of achievement tests, a common practice is to treat missing data as an incorrect answer. However De Ayala, Plake, and Impara (2001) showed that this might lead to worse ability parameter estimation and consequently suggest to substitute missing responses by 0.5 (for dichotomous data).

Another strategy for dealing with missing data is to use several imputation techniques as an advanced data preparation task before applying IRT scaling procedures. Following such a procedure Finch (2008) investigated the effectiveness of different imputation techniques for the recovery of accurate sample estimates of item difficulties and discrimination parameters. Using simulated data Finch (2008) showed that applying different imputation methods resulted in varying degrees of effectiveness, depending on the different underlying mechanisms of missing data. Thus some of those imputation methods require preliminary assumptions about the missing data mechanism, which most times are hard to be verified. Others, especially model-based methods, need to be adjusted to specific subsequent research questions which address the selection of additional variables for the imputation model. This in turn might necessitate repeated IRT scaling, when different research questions arise involving other variables. To sum up, one can say, that imputation techniques as an advanced data preparation task before applying IRT scaling procedures might hold some potential pitfalls, which might be difficult to deal with.

Against the backdrop of large-scale assessments for parameter estimation in presence of missing data, which result from booklet designs, a prevalent practice is to use MML estimation (Bock & Aitkin, 1981; Bock, Mislevy, & Woodson, 1982). This usually aims at estimating population statistics on group level, based on data gathered with incomplete booklet designs (Mislevy, Beaton, Kaplan, & Sheehan, 2005) under the assumption of completely randomly missing data (MCAR) by design. If MML is used to estimate model parameters, based on the assumption of normal distribution of ability followed by an comparative analysis of ability of sub-populations, those sub-population defining variables should be taken into account for estimation by using latent regression models. This auxiliary information in turn might improve precision of parameter estimates

(Mislevy, 1988). Also Joint Maximum Likelihood estimation (JML) has proven to be robust against missing data (Linacre, 1999) and is hence applied under the assumption of a random missing mechanism. In comparison to MML even non randomly missing data is more readily accommodated by JML, while this can theoretically be problematic for MML estimation (DeMars, 2002).

A different strategy to cope with incomplete data occasionally encountered in literature is to treat the missing data as parameters and to maximize the complete data likelihood over the missing data and parameters. Although this approach can be useful under particular missing data problems, Little and Rubin (1983) pointed out that this is in general no reliable approach to analyzing data matrices with missing data points. In a more particular case this is simple from a computation perspective, however, only if the number of missing values is small

As a general conclusion one can say that in theory for most algorithms it is not a prerequisite to have a perfect complete data matrix with no missing data point in order to conduct parameter estimation. So scaling following probabilistic theory could be straightforward with incomplete data. However in practice the presence of missing data might have an impact on item fit and the model validity of the Rasch model (Hohensinn & Kubinger, 2011). Furthermore, especially when working with small datasets containing a relative big amount of missing data, the occurrence of empty cells with multistage response formats might rise (Andrich & Luo, 2003). As a consequence thereof problems within the M-step of the EM algorithm might occur when using marginal methods to find the maximum of the likelihood (Watanabe & Yamaguchi, 2004, p. 18).

As an alternative approach to item parameter recovery under the condition of missing data, the pairwise method, as briefly mentioned above, could be taken into consideration.

Origins and early applications of pairwise comparisons

When scaling psychometric inventories within the framework of IRT, a fundamental goal is to identify some items (stimuli) contributing to a latent trait to be measured in an objective manner. According to the Rasch model, the latent trait would have to be measured in a *specific objective* manner i.e. by constructing a metric scale which satisfies the epistemological principle of generalization of scientific statements (Fischer, 1988). This in turn allows quantitative comparisons for both, items and persons, which meets scientific demands. Following a definition given by Rasch (1966), such [pairwise] comparisons can be seen as the foundation of scientific activity in general. The basic principle in applying pairwise comparisons is to judge several objects presented in pairs on whether either of the two objects is preferred over the other within one rating task.

The fundamental goal in applying the method of pairwise comparisons is to derive a ranking, or even metric system, for k objects if only (subjective) pairwise comparisons of the objects presented are possible simultaneously. To get a full rank order given k objects, $\binom{k}{2}$ pairwise rating tasks must be carried out and, as the simplest way of finding the final rank order of the objects, the number of "wins" and "losses" for each of the k objects must be evaluated.

An early origin of the method of pairwise comparisons, in the sense of solving such a practical ranking problem, can historically be located in the beginning of the 20th century in a non-psychometric context – more specifically in the ranking of chess tournaments (e.g. Ahrens, 1901; Drobny, 1900, 1901; Tietz, 1900a, 1900b). It resulted from the necessity to obtain rather objective and fair rankings of some competitors in tournaments where two competitors were playing each other respectively, either losing, winning or tying (Drobny, 1900; Landau, 1914). As pointed out by David (1988, p. 9) the historical origins of the method of paired comparisons can even be traced back to Fechner (1860). In his book *Elemente der Psychophysik* [elements of psychophysics] Fechner reported some results studying the relation between human perception and physical measures. He found that the point of just noticeable difference in intensity of two physical stimuli can be expressed as the odds ratio of a stimulus strength and the difference in strength of the preceding stimulus. The crucial point here was the finding that this relation between difference in stimulus strength and human perception can be generalized over several human perceptual modalities and might thus fulfill the requirements of scientific (objective) measurement as mentioned above.

Those early applications of pairwise comparisons arose from the necessity to find (specific) objective rankings of either stimulus severity as in the case of psychophysics as proposed by Fechner, or to rank playing strength as in any sports competitions. However, both applications share the same goal which also complies with the principles and fundamental assumptions for parameter estimation in the framework of IRT. In the context of ranking chess tournaments it was Zermelo (1929) who first put this ranking problem into the mathematical context of maximizing the likelihood of a probability distribution for the assumed playing strength of each chess player. With regard to the ranking of tournaments Kendall (1955) showed that the ranking can be improved by reallocating the score of each player by adding the scores of the players they had beaten, which is numerical equivalent to powering the matrix of pairwise comparisons. As already proved by Kendall (1955) repeatedly powering the matrix of pairwise comparisons will stabilize the rankings (David, 1987) and as powering proceeds, the final ordering converges with the ordering given by the eigenvector of the largest eigenvalue derived from the comparison matrix (see also: David, 1971; Garner & Engelhard, 2000; Saaty, 2008).

In the framework of the Rasch model the term "specific objective" can be defined as the metric order of the items along their difficulties, as expressed in the item parameters of the model, being invariant with respect to any particular group measured (Rasch, 1960). Furthermore, as repeatedly pointed out by Georg Rasch, specific objectivity can be expressed in terms of pairwise comparisons (Rasch, 1966; Rasch & Wright, 1979). Thus pairwise comparisons and measurement applying the Rasch model share generally the same goal, namely constructing a common invariant even metric (ranking) scale. How to derive the systematical pairwise comparison of item category frequencies from the basic equation of the Rasch model, following the approach proposed by Choppin (1968, 1985), will be shown in a more formalized way in the following section.

Deriving the pairwise method

The equation of the logistic model formulated by Rasch (1960), formalizes the probability of a person to answer in any but two item categories $p\left(x_{vi}\right), x \in \{0,1\}$, as a logistic function of the ability θ_v of a person v answering to an item i with difficulty σ_i (see equation 1).

$$p\left(x_{vi}\right) = \frac{e^{x_{vi}\left(\theta_v - \sigma_i\right)}}{1 + e^{\left(\theta_v - \sigma_i\right)}} \tag{1}$$

Equation (1) can be "split" into two equations formalizing the probability for each of the two answer categories respectively. So the probability of a correct response is (2) and the probability of a wrong response is given by (3).

$$p\left(x_{vi} = 1\right) = \frac{e^{\left(\theta_v - \sigma_i\right)}}{1 + e^{\left(\theta_v - \sigma_i\right)}} \tag{2}$$

$$p\left(x_{vi} = 0\right) = \frac{1}{1 + e^{\left(\theta_v - \sigma_i\right)}} \tag{3}$$

In continuation let us assume that a sample of n persons v (with $v = 1...n$), has to answer on two items i and j. Based on the core assumption of conditional stochastic independence answering to i and j given the person parameter in the Rasch model, the probabilities of the four possible outcomes answering on the two items achieving a sum score of either 0, 1 or 2 respectively, may be calculated by combining the probabilities of the single possible outcomes answering to each item respectively by multiplication. Thus using (2) and (3) to express the probabilities of any of the three possible raw score outcomes results in (4) to (7).

$$p\left(x_{vi}=0,x_{vj}=0\right)=\frac{1}{1+e^{\left(\theta_v-\sigma_i\right)}}\times\frac{1}{1+e^{\left(\theta_v-\sigma_j\right)}} \tag{4}$$

$$p\left(x_{vi}=1,x_{vj}=0\right)=\frac{e^{\left(\theta_v-\sigma_i\right)}}{1+e^{\left(\theta_v-\sigma_i\right)}}\times\frac{1}{1+e^{\left(\theta_v-\sigma_j\right)}} \tag{5}$$

$$p\left(x_{vi}=0,x_{vj}=1\right)=\frac{1}{1+e^{\left(\theta_v-\sigma_i\right)}}\times\frac{e^{\left(\theta_v-\sigma_j\right)}}{1+e^{\left(\theta_v-\sigma_j\right)}} \tag{6}$$

$$p\left(x_{vi}=1,x_{vj}=1\right)=\frac{e^{\left(\theta_v-\sigma_i\right)}}{1+e^{\left(\theta_v-\sigma_i\right)}}\times\frac{e^{\left(\theta_v-\sigma_j\right)}}{1+e^{\left(\theta_v-\sigma_j\right)}} \tag{7}$$

To further estimate the differences in difficulty of the two items *i* and *j* we now focus on the persons being right on item *i* while being wrong on item *j* (5) and the other way round (6), following the approach expressed by Choppin (1968). This is an obvious approach to follow, as one would not expect to get any information regarding the differences in item difficulty out of a correct or wrong answer on both items respectively. The joint probability to achieve a sum score of 1 answering on both items can the expressed by equation 8.

$$p\left(x_{vi}+x_{vj}=1\right)=\frac{e^{\left(\theta_v-\sigma_i\right)}}{\left(1+e^{\left(\theta_v-\sigma_i\right)}\right)\times\left(1+e^{\left(\theta_v-\sigma_j\right)}\right)}+\frac{e^{\left(\theta_v-\sigma_j\right)}}{\left(1+e^{\left(\theta_v-\sigma_i\right)}\right)\times\left(1+e^{\left(\theta_v-\sigma_j\right)}\right)} \tag{8}$$

The conditional probability to score 1 on item *i* given the sum score on both items is $x_{vi}+x_{vj}=1$, that is $p\left(x_{vi}=1|x_{vi}+x_{vj}=1\right)$, can be expressed as a fraction of (8) and (5). Thus taking (5) as the enumerator and (8) as the denominator results in (9).

$$p\left(x_{vi}=1|x_{vi}+x_{vj}=1\right)=\frac{\frac{e^{\left(\theta_v-\sigma_i\right)}}{1+e^{\left(\theta_v-\sigma_i\right)}}\times\frac{1}{1+e^{\left(\theta_v-\sigma_j\right)}}}{\frac{e^{\left(\theta_v-\sigma_i\right)}}{\left(1+e^{\left(\theta_v-\sigma_i\right)}\right)\times\left(1+e^{\left(\theta_v-\sigma_j\right)}\right)}+\frac{e^{\left(\theta_v-\sigma_j\right)}}{\left(1+e^{\left(\theta_v-\sigma_i\right)}\right)\times\left(1+e^{\left(\theta_v-\sigma_j\right)}\right)}} \tag{9}$$

By analogy to the possibility of eliminating the item parameter out of the equation in oder to estimate person parameter in the Rasch model, as expressed in Kubinger (1988, p. 40)

– see also (Fischer, 1974), the fraction in (9) can be reduced with some mathematical operations to (10), canceling out the person parameter θ_v.

$$p\left(x_{vi} = 1 | x_{vi} + x_{vj} = 1\right) = \frac{e^{\sigma_i}}{e^{\sigma_i} + e^{\sigma_j}} \tag{10}$$

By analogy to equation (10), the conditional probability to score 1 on item j given the sum score on both items is $x_{vi} + x_{vj} = 1$ can be expressed as (11).

$$p\left(x_{vj} = 1 | x_{vi} + x_{vj} = 1\right) = \frac{e^{\sigma_j}}{e^{\sigma_i} + e^{\sigma_j}} \tag{11}$$

Both probabilities given in (10) and (11) can be estimated from empirical data. That is counting the number or respondents $f_{i,j}$ who answered right on item i while being wrong on item j restricted to the subset of $n_{i,j} = f_{i,j} + f_{j,i}$ respondents who achieved a sum score of 1 answering on both items and further counting the number or respondents $f_{j,i}$ being right on item j while being wrong on item i restricted to the subset of $n_{i,j} = f_{i,j} + f_{j,i}$.

These relations can be expressed in (12) and (13).

$$\frac{f_{i,j}}{n_{i,j}} = \frac{e^{\sigma_i}}{e^{\sigma_i} + e^{\sigma_j}} \tag{12}$$

$$\frac{f_{j,i}}{n_{i,j}} = \frac{e^{\sigma_j}}{e^{\sigma_i} + e^{\sigma_j}} \tag{13}$$

Both relations expressed in (12) and (13) can be rewritten, given $n_{i,j} = f_{i,j} + f_{j,i}$ for a person sample of size n. Thus the ratio of the conditional category response frequencies taken from the person sample can be used to estimate the ratio of item difficulties for the population respectively, as expressed in (14) and (15).

$$\frac{f_{i,j}}{f_{j,i}} = \frac{\widehat{e^{\sigma_i}}}{e^{\sigma_j}} \tag{14}$$

$$\frac{f_{j,i}}{f_{i,j}} = \frac{\widehat{e^{\sigma_j}}}{e^{\sigma_i}} \tag{15}$$

Taking the natural logarithm over the equations (14) and (15) results in two reciprocal equations (16) and (17), expressing the estimates of the differences of the item difficulties as a function of the log odds ratios of $f_{i,j}$ and $f_{j,i}$ respectively.

$$\ln\left(\frac{f_{i,j}}{f_{j,i}}\right) = \widehat{\sigma_j - \sigma_i} \tag{16}$$

$$\ln\left(\frac{f_{j,i}}{f_{i,j}}\right) = \widehat{\sigma_i - \sigma_j} \tag{17}$$

Likewise to the application of the pairwise comparison approach in a non-psychometric context, as outlined above, $\binom{k}{2}$ pairwise (item) rating tasks $f_{j,i}$ and $f_{i,j}$ must be carried out with k being the number of items, to get a full metric ranking for the k items. To allow for an unambiguous identification of the item difficulty parameters that are based on the relations expressed in equations (16) and (17), an additional restriction must be introduced. In the framework of IRT that is usually the assumption that the estimated item difficulty parameters sum to zero.

With respect to any existing missing values in the underlying data matrix used for estimation, it has to be noted that the occurrence of missing data points does not necessarily cause any problems in determining $f_{j,i}$ and $f_{i,j}$. This had also been mentioned by Choppin (1985, p. 34). In the same sense, only individuals being exposed to any pair of items i, j (and submitting a valid response on them), contribute to the estimation of the respective $f_{j,i}$ and $f_{i,j}$, while others don't. The estimates of the item parameters base only on the relative ratios of the response category frequencies, regardless of what persons these frequencies arise – see also (Rasch, 1977). Thus, with regard to incomplete data matrices, the minimal prerequisite in terms of completeness can be defined as follows. To successfully estimate all $f_{j,i}$ and $f_{i,j}$ of all items on one common scale, it must be ensured that the observed data matrix cannot be split into two independent data sets due to item omission, by simply rearranging the persons and items. This would be the case if a subset of people n_1 completely omits a subset of the items i_1 while another subset of people n_2 completely omits the subset of items i_2 (with $i_1 \neq i_2$ and $n_1 \neq n_2$). Such a minimal prerequisite for parameter estimation using the PAIR algorithm matches a definition already given by Fischer (1981) with regard to the mere existence of any (item) parameters to be estimated based on a given data matrix – see also (Rost, 2004, p. 317).

Taking into account the numerical technique within the PAIR algorithm, as derived above, obviously no standard errors result from the calculation process that way. In contrast to maximum likelihood based approaches of item parameter estimation returning standard errors based on the assumptions of likelihood theory, other methods have to be applied. The estimation of standard errors, as implemented in the package `pairwise` (Heine, 2014), is realized by bootstrapping the standard errors. That is repeated calculation of item parameters for subsamples, which are drawn from the data matrix being analyzed. This approach is described in a more detailed manner by Brennan (2001, p. 185) in the framework of the Generalizability Theory first discussed in a 1972 monograph by Cronbach, Gleser, Nanda, and Rajaratnam (1972) – see also (Cronbach, Linn, Brennan, & Haertel, 1997). Using bootstrapped standard errors will also mean adopting a technique that is widely used for estimating variances by analyzing data that is collected by using complex sampling designs (Rust, 1985; Rust & Rao, 1996).

For the purpose of practical implementation of the formalized derivation given above, the conditional item category response frequencies as well as the log odds ratios should be arranged in a matrix format. A simple example in the next section will illustrate this.

The pairwise method – a simple example

Let's suppose *M* is a data matrix that contains the answers of eight persons on four dichotomous items. The data is coded in the usual $0,1$ manner, with 0 representing a wrong answer and 1 standing for the correct answer, while a "NA" entry represents a missing value on the response variable. This matrix *M* might look like the one depicted in figure 1.

	Item 1	*Item* 2	*Item* 3	*Item* 4
Person 1	1	1	1	0
Person 2	0	1	0	1
Person 3	1	0	0	1
Person 4	1	0	0	0
Person 5	*NA*	1	1	*NA*
Person 6	1	1	*NA*	1
Person 7	*NA*	0	0	*NA*
Person 8	1	*NA*	0	0

Figure 1: Simple example data matrix *M*.

Please note, that there are some missing data points in this simple example data matrix, which corresponds to a missing proportion of 18.75 %.

The first step when applying the pairwise comparison algorithm according to Choppin (1968) is to count the conditional category response frequencies for each item. In case of dichotomies that is counting the number of persons who got item i right under the condition of having answered wrong to item j for every item $k(i \neq j)$. Based on the simple example data matrix in figure 1 this results in a symmetrical pairwise comparison matrix C with entries $f_{i,j}$ and $f_{j,i}$, which is depicted in figure 2.

$$C_{f_{i,j}} = \begin{bmatrix} 0 & 2 & 3 & 3 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 2 & 0 \end{bmatrix}$$

Figure 2: Pairwise comparison matrix C with entries $f_{i,j}$ and $f_{j,i}$.

$$Z_{data} = \begin{array}{r|cccccccc} & I1.0 & I1.1 & I2.0 & I2.1 & I3.0 & I3.1 & I4.0 & I4.1 \\ P1 & 0 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \\ P2 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ P3 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ P4 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ P5 & \mathbf{0} & \mathbf{0} & 0 & 1 & 0 & 1 & \mathbf{0} & \mathbf{0} \\ P6 & 0 & 1 & 0 & 1 & \mathbf{0} & \mathbf{0} & 0 & 1 \\ P7 & \mathbf{0} & \mathbf{0} & 1 & 0 & 1 & 0 & \mathbf{0} & \mathbf{0} \\ P8 & 0 & 1 & \mathbf{0} & \mathbf{0} & 1 & 0 & 1 & 0 \end{array}$$

Figure 3: Item category indicator matrix Z derived from data matrix M.

It should be noted, that the matrix C has an analogy to the Burt matrix B (Burt, 1950), used in correspondence analysis (Greenacre, 1984; Blasius, 2001). The Burt matrix is computed based on the (item category) indicator matrix Z depicted in figure 3. Z is an straightforward representation of the data matrix M using a binary indicator (column) vector for each item category indicating whether the respective category was chosen (coded 1) or not (coded 0) over all persons and items in M. Special attention has to be directed to the change in representing missing data points from M in Z. While missing data points in M carry no numerical information regarding any chosen category of the respective item, this is different for Z. Here missing data points are coded in that manner having zero entries in both item category columns of a respective item in Z. The straightforward approach is to keep the information that none of the offered categories $(0,1)$ was chosen by the person for the respective item. Computing the Burt matrix B is simply multiplying the transposed indicator matrix Z by itself, that is $B = Z^T Z$. Contrary to the Burt Matrix B depicted in figure 4 (computed based on Z from M), some columns and rows are skipped in matrix C.

$$B_{f_{i,j}} = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 5 & 2 & 2 & 3 & 1 & 3 & 2 \\ 0 & 2 & 3 & 0 & 3 & 0 & 1 & 1 \\ 1 & 2 & 0 & 4 & 1 & 2 & 1 & 2 \\ 1 & 3 & 3 & 1 & 5 & 0 & 2 & 2 \\ 0 & 1 & 0 & 2 & 0 & 2 & 1 & 0 \\ 0 & 3 & 1 & 1 & 2 & 1 & 3 & 0 \\ 1 & 2 & 1 & 2 & 2 & 0 & 0 & 3 \end{bmatrix}$$

Figure 4: Burt matrix B computed from indicator super matrix Z based on responses in M.

Conferring to this simple example C could be derived out of B by just skipping rows 1, 3, 5 and 7 and columns 2, 4, 6 and 8. Naturally the diagonal of this Matrix C is filled with zeros, as there is no meaningful comparison of one item with itself. But note that that there are also off diagonal zeros to be found in this small example, which could cause numerical problems in the following steps of computation if not adequately incorporated.

$$C^3_{f_{i,j}} = \begin{bmatrix} 8 & 15 & 28 & 26 \\ 7 & 6 & 14 & 11 \\ 1 & 2 & 4 & 6 \\ 8 & 8 & 17 & 9 \end{bmatrix}$$

Figure 5: Power of 3 for pairwise comparison matrix C.

In the second step of applying the pairwise comparison algorithm, a so called positive reciprocal matrix with entries $f_{i,j}/f_{j,i}$ and $f_{j,i}/f_{i,j}$ must be formed. In case of off diagonal zeros in the antecedent matrix C, Choppin (1985) showed algebraically that instead of using the direct comparisons in matrix C, the square of C can be used. As already mentioned above in this paper, even higher powers of the matrix might be used as discussed in prior research by other authors (David, 1971; Garner & Engelhard, 2000; Kendall, 1955; Saaty, 2008). However, with regard to the question which power of the pairwise comparison matrix to take, one has to consider *’how far one would wish to go on practical grounds’* (Kendall, 1955, p. 50). Taking into account that higher potencies of the comparison matrix can lead to quite large numerical entries soon, it can be expected that this may be associated with problems in sense of numerical inaccuracies when calculating the log odds ratios of these entries in the further steps of the algorithm. Thus, as powering the comparison matrix is only necessary as far as off diagonal zeros vanish, this can be seen as a sensible criteria to which degree the comparison matrix should be powered. Under the general condition that all items are proper linked in a homogeneous scale with an intermediate number of items, no off diagonal zeros should occur when using at least the power of 3 from C. This is true for the small example data as can be seen in figure 5.

Forming the reciprocal matrix based on C^3 along the diagonal leads to the matrix D shown in figure 6.

$$D = \begin{bmatrix} {}^{8}/_{8} & {}^{7}/_{15} & {}^{1}/_{28} & {}^{8}/_{26} \\ {}^{15}/_{7} & {}^{6}/_{6} & {}^{2}/_{14} & {}^{8}/_{11} \\ {}^{28}/_{1} & {}^{14}/_{2} & {}^{4}/_{4} & {}^{17}/_{6} \\ {}^{26}/_{8} & {}^{11}/_{8} & {}^{6}/_{17} & {}^{9}/_{9} \end{bmatrix}$$

Figure 6: Reciprocal matrix D, formed from C^3.

To identify the final item difficulty parameters for the four items based on the relations ${}^{f_{i,j}}/_{f_{j,i}}$ and ${}^{f_{j,i}}/_{f_{i,j}}$ given in matrix D, the natural logarithm of the entries in D are taken and their row means are computed. This implies constraining the estimated item parameters to sum to zero for the purpose of model identification.

In order to analyze a data matrix that comprises polytomous items with more than two answer categories, the procedure of pairwise comparison basically follows the same underlying principle. To estimate the item category difficulties, each item in the data matrix is dummy coded forming a $0,1$ coded item–category data (super) matrix. This matrix is then in turn analyzed in the manner described in the simple example above. Again, the first step in the pairwise algorithm is to form the symmetrical pairwise comparison matrix C.

Now, analyzing the dummy coded item–category data matrix, the algorithm is therefore extended to the comparison of answer frequencies for each category of each item. In this case, the pairwise comparison matrix C_{fi,fj_c} with entries f_{i_c,j_c} represents the number of respondents who answered to item i in category c and to item j in category $c-1$ thus expanding Choppin's conditional pairwise algorithm to polytomous item response formats. As there is no meaningful comparison to the answer category frequencies within one item with itself, there are symmetrical square areas with size $(m-1) \times (m-1)$, with

$$C_{f_{i_c j_c}} = \begin{bmatrix} 0 & 0 & 0 & 104 & 279 & 105 & 87 & 355 & 59 \\ 0 & 0 & 0 & 36 & 135 & 128 & 32 & 177 & 95 \\ 0 & 0 & 0 & 13 & 37 & 26 & 6 & 52 & 32 \\ 20 & 279 & 135 & 0 & 0 & 0 & 60 & 336 & 66 \\ 7 & 105 & 128 & 0 & 0 & 0 & 16 & 152 & 84 \\ 1 & 22 & 22 & 0 & 0 & 0 & 3 & 23 & 32 \\ 21 & 355 & 177 & 94 & 336 & 152 & 0 & 0 & 0 \\ 6 & 59 & 95 & 10 & 66 & 84 & 0 & 0 & 0 \\ 2 & 9 & 17 & 4 & 9 & 14 & 0 & 0 & 0 \end{bmatrix}$$

Figure 7: Pairwise comparison matrix C for three items with $m = 4$ answer categories (coded from 0 to 3) with entries for conditional category frequencies f_{i_c,j_c}.

m being the number of answer categories, containing zero entries along the diagonal of the resulting pairwise comparison matrix C. Figure 7 illustrates how such a matrix C might look like for three items with $m = 4$ answer categories. Every group of three columns (from left to right) holds the answer category frequencies f_{i_c,j_c} for every item i as a result of the comparison of answer categories i_c (for $c = 1$ to $c = m-1$) with answer categories j_c (for $c = 0$ to $c = m-2$) with every item j (with $i \neq j$), which are arranged in groups of three rows (from top to bottom).

Applying the pairwise method

In the following section the method of pairwise comparisons for item parameter estimation in the framework of IRT is applied to an empirical dataset. The results of the estimation of the non-iterative PAIR algorithm as outlined above and implemented in the `R` package `pairwise`, are therefore compared with the results of estimation applying three other procedures. That are (1) Item parameter estimation with MML as implemented in the software package `ConQuest` (2) Item parameter estimation with CML as implemented in the software package `WINMIRA` and (3) applying CML estimation with `WINMIRA` on the prior imputed data using `SPSS` for the imputation task.

Data and Scale

The data was collected at the *Universität der Bundeswehr München* (Germany) as part of lectures in methodology of the social sciences between 1999 and 2002. Altogether

787 male students took part in a survey covering various aspects of (1) the general situation of studies at the university and (2) comprising several inventories related to personality and occupational interests. Thereof 620 students answered questions of the NEO-PI-R inventory (Ostendorf, 2004) by filling out the paper pencil form of the inventory. The data matrix used for the subsequent analysis thus comprised $n = 620$ persons responding to eight questions of the personality scale *Self-Consciousness* (N4). The German formulations for the eight items and the respective translations in square brackets are depicted in table 1.

Table 1: Item formulations for the personality facet Self-Consciousness (N4) in their German version as used in the present study.

polarity	name	text
+	N017	Im Umgang mit anderen befürchte ich häufig, dass ich unangenehm auffallen könnte. [*In dealing with others, I often fear that I could attract negative attention.*]
+	N047	Manchmal war mir etwas so peinlich, dass ich mich am liebsten versteckt hätte. [*Sometimes I felt so embarrassed, that I would have liked to hide away.*]
-	N077	Es bringt mich nicht besonders in Verlegenheit, wenn andere mich verspotten und lächerlich machen. [*I don't feel particularly embarrassed when others mock me or make fun of me.*]
-	N107	Ich bin selten verlegen, wenn ich unter Leuten bin. [*I am rarely embarrassed when I'm around people.*]
+	N137	Ich fühle mich anderen oft unterlegen. [*I often feel inferior to others.*]
-	N162	In Gegenwart meiner Chefs oder anderer Autoritäten fühle ich mich wohl. [*In the presence of my bosses or other authority figures I feel comfortable.*]
+	N167	Wenn ich einer Person etwas falsches gesagt oder angetan habe, kann ich es kaum ertragen, ihr noch einmal zu begegnen. [*If I have said or done something wrong to a person, I can hardly bear to meet that person again.*]
+	N217	Wenn meine Bekannten dummen Unfug treiben, so ist mir das peinlich. [*Whenever my friends engage in any mischief, I feel embarrassed.*]

The eight questions had to be answered on a five point Likert answer scale which was marked with integer numbers ranging from −2 (completely untrue) to 2 (completely true). In the preceding written instructions those integer numbers were assigned to verbal descriptors, which are shown in table 2.

Table 2: Integer marks of the answer categories for the eight items of the personality facet Self-Consciousness (N4) and their respective verbal descriptors.

integer mark	verbal descriptor
	kreuzen Sie an ... [*tick ...*]
-2	... wenn Sie der Aussage auf keinen Fall zustimmen oder wenn Sie meinen, daß die Aussage für Sie **VÖLLIG UNZUTREFFEND** ist. [... *if you absolutely disagree with the statement in any case, or if you think this statement* ***DOES NOT APPLY TO YOU AT ALL***.]
-1	... wenn Sie der Aussage nicht zustimmen oder wenn Sie meinen, daß die Aussage für Sie **UNZUTREFFEND** ist. [... *if you don't agree or if you think or if you think this statement* ***DOES NOT APPLY*** *to you.*]
0	... wenn die Aussage weder richtig noch falsch ist oder wenn Sie meinen, daß die Aussage für Sie **WEDER ZUTREFFEND NOCH UNZUTREFFEND** ist. [... *if the statement is neither right nor wrong or if you feel unsure about* ***WHETHER OR NOT THE STATEMENT APPLIES*** *to you.*]
1	... wenn Sie der Aussage zustimmen oder wenn Sie meinen, daß die Aussage auf Sie **ZUTRIFFT**. [... *If you agree with the statement or if you think that the statement* ***APPLIES*** *to you.*]
2	... wenn Sie der Aussage vollkommen zustimmen oder wenn Sie meinen, daß die Aussage auf Sie **VÖLLIG ZUTRIFFT**. [*if you absolutely agree with the statement or if you feel that the statement* ***ABSOLUTELY APPLIES*** *to you.*]

Negative formulated items, in regard to the measured construct, were reversed in polarity as described in the instructions manual of the NEO-PI-R inventory prior to further analysis. Due to sparse use of some of the five answer categories in some items, the two categories 3 and 4 where collapsed. The resulting raw item category frequencies are depicted in table 3.

Method

The pairwise method was tested against other typically applied methods of item parameter estimation in the field of social sciences. For this purpose the analysis dataset was

Table 3: Frequencies of collapsed answer categories for eight items of the personality facet Self-Consciousness (N4); $n = 620$.

	N017	N047	N077	N107	N137	N162	N167	N217
missing	0	0	0	0	0	0	0	0
score 0	123	44	40	73	170	34	69	49
score 1	257	162	188	318	320	199	270	211
score 2	216	306	322	215	123	362	254	331
score 3	24	108	70	14	7	25	27	29

first estimated at baseline (containing no missing values) with three different software programs incorporating there different types of item parameter estimation methods in the framework of Item Response Theory (IRT). The underlying estimation model was the Partial Credit Model (PCM – Masters, 1982) for a polytomous response format.

The methods used were (1) PAIR algorithm as described in the previous sections, implemented in the R–package `pairwise` (Heine, 2014), (2) CML estimation implemented in the stand alone software package `WINMIRA` (von Davier, 2001), and (3) MML estimation as implemented in the software package `ConQuest` by Wu et al. (2012). Additionally a fourth procedure of parameter estimation with missing data was incorporated by imputing the missing dataset at several stages of missingness prior to CML estimation performed with `WINMIRA`. The data imputation task was performed using the fully conditional specification (FCS), that is an iterative Markov Chain Monte Carlo (MCMC) method which is appropriate when the data have an arbitrary (monotone or non-monotone) missing pattern as implemented in SPSS (2010).

Before drawing any inferences from empirical data based on the assumptions of the Rasch model the global model fit for the empirical data should be tested using a reliable model fit test (Heene, Draxler, Ziegler, & Bühner, 2011). A frequently used test – which is also implemented in the standalone software WINMIRA – is the Pearson χ^2 - statistic. This test is based on the differences between the observed and expected (response) pattern frequencies given the model assumptions. As a consequence from using multilevel response formats, which in turn leads to large number of possible patterns, the asymptotic conditions for the interpretation of the χ^2 - statistic are often not met in most cases of empirical research (e.g. Bühner, 2006, p. 341). As a solution von Davier (1997) proposed a parametric bootstrap method to generate the sampling distribution of the test statistic. However Heene et al. (2011) showed that this procedure might fail to reject non-fitting data. Hence a graphical model check was used for the present study to test whether the empirical data at baseline satisfies the assumed model being estimated. Additionally local model violations were examined based on item fit estimates.

A respective functionality which is available in the `R` package `pairwise` (Heine, 2014) was used for these tests – the graphical model check as well as the evaluation of item fit statistics, based on the standardized mean squared residuals (Wright & Masters, 1982, p. 100).

To further investigate the quality of the respective item estimates under a continuing rise in percentage of missing data, the baseline dataset was *artificially* charged with missing data points. Thus, starting with the complete baseline data, missing data points were added in a step-wise procedure using 5 % steps ranging from 5 % to 40 %. The four procedures of item parameter recovery were applied to every level of missing data to the respective data matrix. The underlying process of simulated missing data followed a random distribution as sampled over the eight items of the N4 scale. The simulated data loss mechanism could therefore be assumed to be more or less MCAR.

Results

Serving as a first verification of the global model fit a graphical model check was performed. The baseline data was divided at the median of the raw score distribution ($M_d = 12$) into two subsamples. Then the item parameters were estimated for the two subsamples and were plotted on the `x` and `y` axis against each other. Unfortunately for item `N137` this resulted in substantial deviation from parameter equality of the two subsamples due to low cell frequencies in the fourth category (score 3) of this item (see table 3). Hence it was excluded from the global fit test. The graphical Model Check for the remaining seven items is depicted in figure 8.

As can be seen some of the remaining seven items show some minor deviations from the diagonal of the plot. This provides some evidence for a moderately overall fit of the estimated model. In order to get a more detailed insight into the model fit, a closer look at any existing local model violations is given by inspecting the item fit statistics as depicted in table 4. Following the recommendations in interpreting the mean square residual fit statistic as given in Bond and Fox (2006), the item fit can be interpreted as rather conform to the assumptions of the model estimated. Thus the present dataset may be a sufficient starting point for further analysis concerned with comparing the different methods of item parameter estimation at different stages of missing data.

To visualize any existing deviations of the parameter estimates resulting from different estimation approaches, the Thurstonian thresholds of the item answer categories were plotted. The resulting item parameters and their respective standard errors from the baseline estimation – no missing data points – comparing CML, MML and the PAIR algorithm are depicted in figure 9. The results show a rather good match of the estimates resulting from the four methods, plotted as Thurstonian threshold lines. The respective numerical results for the point estimates as well as their respective standard errors for the three estimation algorithms (at baseline) are outlined in table 5.

For further comparisons at different levels of missing data, the estimates resulting out of MML estimation using ConQuest were transformed into Thurstonian thresholds to be comparable with the parameter estimates resulting from WINMIRA and pairwise. Figure 10 shows the Thurstonian threshold lines for the three algorithms applied at a rate of 35 % of missing data and their respective deviations from baseline estimation.

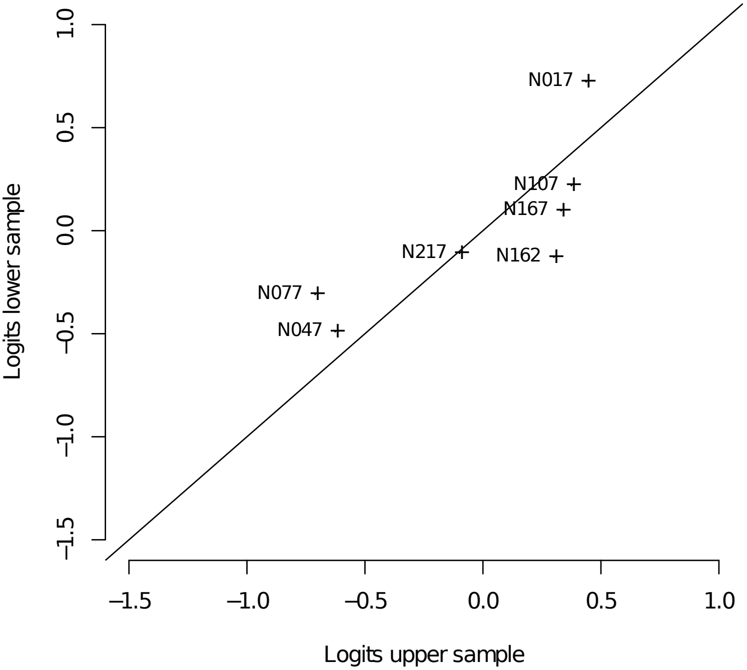

Figure 8: Graphical Model Check based on two subsamples (median split; $M_d = 12$) plotting the respective item parameters for seven items and complete data; $n = 620$.

Table 4: Item Fit statistics for complete data; $n = 620$.

	Chi	*df*	*p*	OUTFIT.MSQ	OUTFIT.ZSTD	INFIT.MSQ	INFIT.ZSTD
N017	528.34	619	1.00	0.85	-3.00	0.84	-3.39
N047	543.18	619	.99	0.88	-2.37	0.88	-2.28
N077	555.90	619	.97	0.90	-1.93	0.90	-1.97
N107	495.54	619	1.00	0.80	-3.94	0.80	-3.99
N137	506.54	619	1.00	0.82	-3.75	0.82	-3.64
N162	596.50	619	.74	0.96	-0.61	0.95	-0.95
N167	535.79	619	.99	0.86	-2.67	0.86	-2.71
N217	576.55	619	.89	0.93	-1.26	0.94	-1.18

Table 5: Thurstonian thresholds (1,2,3) point estimates (est.) and standard errors (SE) for pairwise (PAIR) and WINMIRA (CML) and ConQuest (MML) at baseline (no missing data); $n = 620$.

		pairwise			ConQuest			WINMIRA		
		1	2	3	1	2	3	1	2	3
N017	est.	-1.41	−0.04	2.73	−1.40	−0.02	2.49	−1.39	−0.05	2.51
	SE	0.11	0.12	0.23	0.11	0.10	0.14	0.06	0.08	0.09
N047	est.	-2.24	−1.05	0.98	−2.21	−1.09	1.06	−2.22	−1.09	1.03
	SE	0.15	0.10	0.12	0.10	0.09	0.13	0.06	0.10	0.08
N077	est.	-2.40	−0.94	1.52	−2.43	−0.95	1.60	−2.44	−0.96	1.58
	SE	0.17	0.12	0.14	0.12	0.09	0.14	0.06	0.09	0.09
N107	est.	-2.31	0.25	2.73	−2.17	0.18	3.04	−2.18	0.15	3.08
	SE	0.13	0.10	0.31	0.13	0.10	0.16	0.07	0.08	0.09
N137	est.	-1.21	0.92	3.70	−1.17	0.90	3.33	−1.18	0.87	3.44
	SE	0.09	0.12	1.35	0.10	0.12	0.15	0.07	0.08	0.11
N162	est.	-2.35	−1.00	2.74	−2.64	−0.98	2.81	−2.65	−0.99	2.83
	SE	0.21	0.09	0.18	0.15	0.10	0.17	0.07	0.09	0.09
N167	est.	-2.06	−0.22	2.37	−2.11	−0.21	2.47	−2.11	−0.23	2.49
	SE	0.14	0.11	0.21	0.12	0.10	0.15	0.07	0.09	0.09
N217	est.	-2.21	−0.77	2.29	−2.29	−0.80	2.59	−2.30	−0.81	2.60
	SE	0.15	0.11	0.18	0.17	0.14	0.19	0.07	0.09	0.09

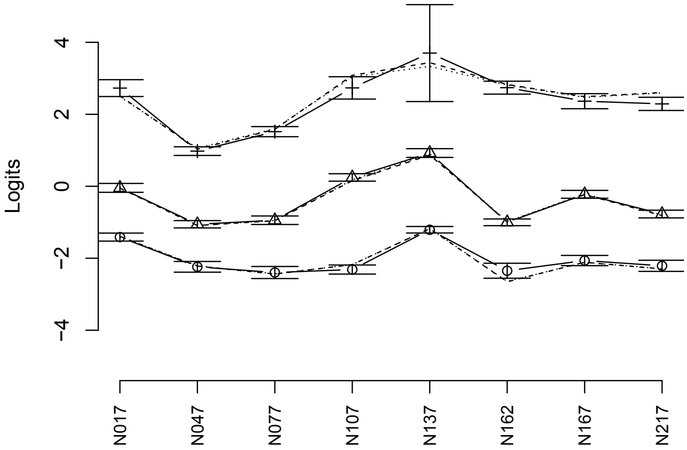

Figure 9: Thurstonian threshold point estimates and SE for PAIR (solid lines) and CML (dashed lines) and MML (dotted lines) at baseline (no missing data); $n = 620$.

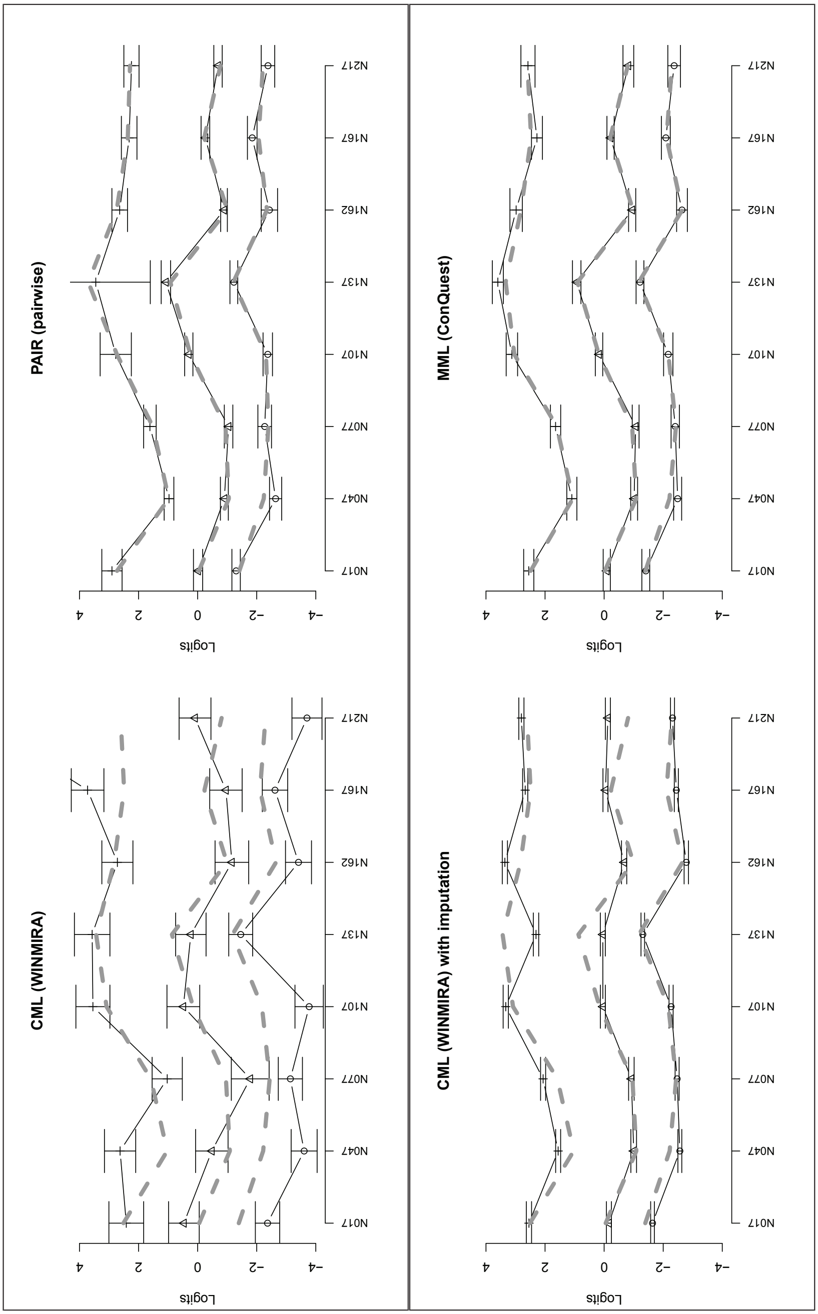

Figure 10: Thurstonian threshold point estimates and size of standard errors for CLM, imputed CML, PAIR and MML item parameters at 35 % of missing data (solid lines) in comparison to their respective estimates at baseline (no missing data – dashed lines); $n = 620$.

To serve as a summary the gradient of the range of deviations for item parameter estimates at several stages of percentage of missing data for the three algorithms is illustrated in figure 11. In that graph, the positive and negative maximum of deviations of the point estimates for the item parameters from the respective baseline measurement, estimated at increasing proportions of missing values, are plotted along x-axis for the four methods of item parameter recovery.

It has to be mentioned that, depending on the method of item parameter recovery, different approaches of handling the missing data, as implemented in the respective software package, were applied. For the CML approach without prior imputation task (see also top left panel in figure 10), as implemented in `WINMRA` for example, the handling of missing values was simply *list wise deletion*. Thus it should be noted that the percentage values that are plotted on the x-axis, always refer to the complete baseline data set with $n = 620$ participants.

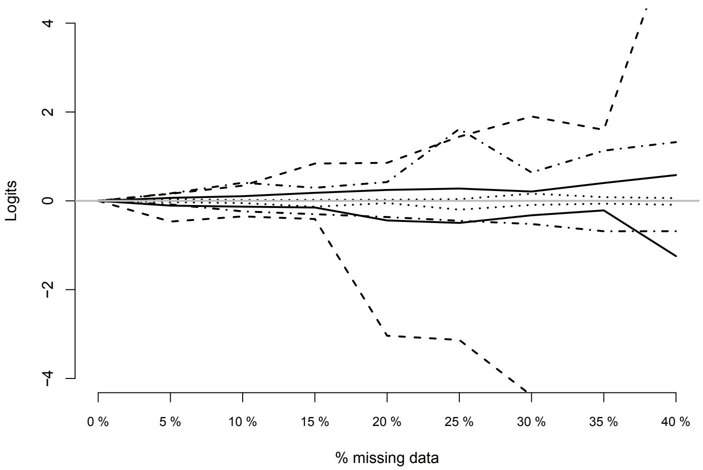

Figure 11: Gradient of range of deviations of item parameter point estimates at several percentages of missing data from baseline estimates for PAIR (solid lines), CML (dashed lines), CML with imputation (dotted-dashed lines) and MML (dotted lines).

Conclusion

The present paper introduced an alternative approach to pairwise item parameter recovery in the framework of IRT. The main benefit of this approach lies in its easy application to incomplete data matrices, even at higher rates of missing data points. It was outlined that this approach has some analogies to applications in tournament rankings in a non-psychometric context. Furthermore it was argued that it shares some fundamental assumptions and characteristics of the Rasch model. Starting with the basic equation of the Rasch model the numerical technique of recovering the item parameters with the pairwise comparison approach was derived. By means of a simple example the method introduced here was illustrated.

In the empirical part of the present paper a real dataset was analyzed applying the pairwise approach and three other methods to the same data set with increasing proportions of missing responses. The graphical model check and the item fit indices performed on the baseline data set using the functionality of the package `pairwise` indicated in an acceptable model fit, thus the data were taken as basis for further analysis. With regard to the standard errors returned by applying a bootstrap approach as implemented in the `pairwise` package (see figure 9) the large standard error for the point estimate of the third threshold for item `N137` has to be mentioned. Inspecting the item category frequencies depicted in table 3 the increased standard error for the third threshold parameter seems to be fairly obvious, with respect to the low cell frequencies for the fourth category of item `N137`. More over having a closer look at the item formulations given in table 1 for that item and furthermore taking into account the somehow special sample comprised male military recruits, it seems rather clear that this item might not be well targeted for this sample. Or to put it the other way round, the sample in this case seems not to be very eligible for the purpose of calibration of that item. Thus considering the low cell frequencies or rather the skewed distribution of responses for that item, the increased standard error seems quite logically to reflect the uncertainty of measurement when calibrating that item and using this sample. Nevertheless the point estimator for that threshold parameter matches the results from the other applied methods reasonably well. Referring to figure 9 one can even claim that the threshold estimates resulting from PAIR perfectly match those resulting from MML and CML estimation at baseline. Results from the estimation under the condition of 35 % missing reveals a relatively good performance in parameter recovery for PAIR as well as for MML estimation (compared to their respective baseline estimate). Accordingly the pairwise method can determine the item parameters of the Rasch model even under the difficult conditions incorporating missing values or low response category frequencies.

In addition to that some limiting aspects have to be mentioned. With regard to the minimum simulation scenario no replications (at any stage of missing data) were calculated in this case. As a result the respective deviations of item parameter estimates reported here, may in a way only represent a first snapshot regarding the impact of missing data on different estimation approaches. Also, in this case, the simulated data loss mechanism charging the baseline data was assumed to be completely at random (MCAR). This is certainly the easiest and probably most unproblematic assumption in regard to item parameter recovery in the framework of IRT. Therefore, an important question could be whether the performance of the four methods of item parameter determination would develop under different assumptions regarding the data loss mechanism (additionally MAR and MNAR), depending on the proportion of missing values. As a consequence future research should concentrate on further simulation (1) comprising replications at different stages of missing data and (2) should include the investigation of several data loss mechanisms and their impact on different approaches of item parameter recovery, including the pairwise approach introduced here. However, the empirical example and the minimal simulation scenario conducted here, support some evidence for the performance of the pairwise method. Even with a relatively large number of missing data points, the item parameter of the Rasch model were determined quite well (1) compared to their respective baseline estimation using the complete dataset and (2) compared to other estimation approaches at baseline.

With regard to the numerical approach within the pairwise method of explicitly calculating the item parameters it has to be mentioned that this does per se not lead to standard errors derived from likelihood theory as in other approaches. Thus a somewhat more complex approach to solving this problem had to be selected, namely performing bootstrapped standard errors by repeatedly calculating the item parameter based on subsamples. Although this might at first seem to be the critical or somewhat laborious point of the method, it must be emphasized that particularly this method of calculating standard errors might offer some advantages. As regards stratified large-scale samples not following a simple random selection approach when sampling the population, bootstrapping the standard errors is a common practice (Rust & Rao, 1996), not least because in such complex sampling designs formulas for the explicit calculation of standard errors are not always appropriately derivable (Rust, 1985). Hence sampling and replication weights are often introduced in such sampling designs to account for unbiased estimation of item parameter point estimates and standard errors. However, not yet implemented in the current version of the package `pairwise`, the use of sampling and replication weights for estimating item parameters as well as their standard errors could easily be implemented in the PAIR algorithm due to its explicit numerical approach. The consideration of case weights for item parameter estimation within the pairwise approach would therefor follow a principle already expressed in literature about pairwise comparisons. Kendall (1955) showed that taking into account the differences in reliability in judges, rating several objects by means of pairwise comparisons, could be expressed by introducing weights for each rater. By analogy, the *item rating* of each person in the sample might be simply weighted using the respective case weight when calibrating items using a stratified sampling design. The application of the pairwise method could hence be particularly useful in such complex sample designs.

Acknowledgment

The authors wish to thank all anonymous reviewers for their remarks, which helped to improve this paper.

References

Ahrens, W. (1901). Zur relativen Bewertung von Turnierpartien. *Wiener Schachzeitung*, *IV*(1), 181–192. Retrieved from `http://anno.onb.ac.at/cgi-content/anno-plus?aid=sze&datum=1901`

Andrich, D., & Luo, G. (2003). Conditional Pairwise Estimation in the Rasch Model for Ordered Response Categories using Principal Components. *Journal of Applied Measurement*, *4*(3), 205–221.

Blasius, J. (2001). *Korrespondenzanalyse*. München: Oldenbourg Wissenschaftsverlag.

Bock, R. D., & Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm. *Psychometrika*, *46*(4), 443–459. doi: 10.1007/BF02293801

Bock, R. D., Mislevy, R., & Woodson, C. (1982). The Next Stage in Educational Assessment. *Educational Researcher*, *11*(3), 4–16. doi: 10.3102/0013189X011003004

Bond, T. G., & Fox, C. M. (2006). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates Inc.

Brennan, R. L. (2001). *Generalizability Theory*. New York: Springer.

Bühner, M. (2006). *Einführung in die Test- und Fragebogenkonstruktion* (2nd ed.). München: Pearson.

Burt, C. (1950). The Factorial Analysis of Qualitative Data. *British Journal of Statistical Psychology*, *3*(3), 166–185. doi: 10.1111/j.2044-8317.1950.tb00296.x

Choppin, B. (1968). Item Bank using Sample-free Calibration. *Nature*, *219*(5156), 870–872. doi: 10.1038/219870a0

Choppin, B. (1985). A fully Conditional Estimation Procedure for Rasch Model Parameters. *Evaluation in Education*, *9*(1), 29–42. doi: 10.1016/0191-765X(83)90005-8

Cohen, L. (1979). Approximate Expressions for Parameter Estimates in the Rasch Model. *British Journal of Mathematical and Statistical Psychology*, *32*(1), 113–120. doi: 10.1111/j.2044-8317.1979.tb00756.x

Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *Theory of Generalizability for Scores and Profiles*. New York: Wiley.

Cronbach, L. J., Linn, R. L., Brennan, R. L., & Haertel, E. H. (1997). Generalizability Analysis for Performance Assessments of Student Achievement or School Effectiveness. *Educational and Psychological Measurement*, *57*(3), 373–399. doi: 10.1177/0013164497057003001

David, H. A. (1971). Ranking the Players in a Round Robin Tournament. *Review of the International Statistical Institute*, *39*(2), 137. doi: 10.2307/1402170

David, H. A. (1987). Ranking from Unbalanced Paired-Comparison Data. *Biometrika*, *74*(2), 432–436. doi: 10.2307/2336160

David, H. A. (1988). *The Method of Paired Comparisons*. London: Griffin.

De Ayala, R. J., Plake, B. S., & Impara, J. C. (2001). The Impact of Omitted Responses on the Accuracy of Ability Estimation in Item Response Theory. *Journal of Educational Measurement*, *38*(3), 213–234. doi: 10.1111/j.1745-3984.2001.tb01124.x

DeMars, C. (2002). Incomplete Data and Item Parameter Estimates Under JMLE and MML Estimation. *Applied Measurement in Education*, *15*, 15–31. doi: 10.1207/S15324818AME1501_02

Drobny, F. (1900). Zum Begrif Turnierstärke. *Wiener Schachzeitung*, *III*(1), 171–176. Retrieved from `http://anno.onb.ac.at/cgi-content/anno-plus?aid=sze&datum=1900`

Drobny, F. (1901). Ueber eine neue Art der Preisvertheilung. *Wiener Schachzeitung*, *IV*(1), 2–4. Retrieved from `http://anno.onb.ac.at/cgi-content/anno-plus?aid=sze&datum=1901`

Fechner, G. T. (1860). *Elemente der Psychophysik*. Leipzig: Breitkopf und Härtel.

Finch, H. (2008). Estimation of Item Response Theory Parameters in the Presence of Missing Data. *Journal of Educational Measurement*, *45*(3), 225–245. doi: 10.1111/j.1745-3984.2008.00062.x

Fischer, G. H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Fischer, G. H. (1981). On the Existence and Uniqueness of Maximum-Likelihood Estimates in the Rasch Model. *Psychometrika*, *46*(1), 59–77. doi: 10.1007/BF02293919

Fischer, G. H. (1988). Spezifische Objektivität: Eine wissenschaftstheoretische Grundlage des Rasch-Modells. In K. D. Kubinger (Ed.), *Moderne Testtheorie* (pp. 87–111). München; Weinheim: Psychologie Verlags Union.

Garner, M., & Engelhard, G. (2000). Rasch Measurement Teory, the Method of Paired Comparisons and Graph Theory. In M. Wilson & G. Engelhard (Eds.), *Objective Measurement: Theory Into Practice* (Vol. 5, pp. 259–286). Stamford, Conneticut: Ablex Publishing Corporation.

Greenacre, M. J. (Ed.). (1984). *Theory and Applications of Correspondence Analysis*. London ; Orlando, Fla: Academic Press.

Heene, M., Draxler, C., Ziegler, M., & Bühner, M. (2011). Performance of the Bootstrap Rasch Model Test Under Violations of Non-Intersecting Item Response Functions. *Psychological Test and Assessment Modeling*, *53*(3), 283–294. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/01_Heene.pdf

Heine, J.-H. (2014). *pairwise: Rasch Model Parameters by Pairwise Algorithm (V 0.2.5)*. Retrieved from http://cran.r-project.org/web/packages/pairwise/index.html (R package version 0.2.5)

Hohensinn, C., & Kubinger, K. D. (2011). On the Impact of Missing Values on Item Fit and the Model Validness of the Rasch Model. *Psychological Test and Assessment Modeling*, *53*(3), 380–393. Retrieved from http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/07_Hohensinn.pdf

Johnson, M. S. (2007). Marginal Maximum Likelihood Estimation of Item Response Models in R. *Journal of Statistical Software*, *20*(10), 1–24. Retrieved from http://www.jstatsoft.org/v20/i10

Kendall, M. G. (1955). Further Contributions to the Theory of Paired Comparisons. *Biometrics*, *11*(1), 43–62. doi: 10.2307/3001479

Kubinger, K. D. (1988). Aktueller Stand und kritsche Würdigung der Probabilistischen

Testtheorie. In K. D. Kubinger (Ed.), *Moderne Testtheorie* (pp. 19–83). München; Weinheim: Psychologie Verlags Union.

Landau, E. (1914). Über Preisverteilung bei Spielturnieren. *Zeitschrift für Mathematik und Physik*, *63*, 192–202.

Linacre, J. M. (1998). Detecting Multidimensionality: Which Residual Data-type Works Best? *Journal of Outcome Measurement*, *2*(3), 266–283. Retrieved from http://www.jampress.org/JOM_V2N3.pdf

Linacre, J. M. (1999). Understanding Rasch Measurement: Estimation Methods for Rasch Measures. *Journal of Outcome Measurement*, *3*(4), 382–405. Retrieved from http://www.jampress.org/JOM_V3N4.pdf

Linacre, J. M. (2004). Rasch Model Estimation: Further Topics. *Journal of Applied Measurement*, *5*(1), 95–110.

Little, R. J. A., & Rubin, D. B. (1983). On Jointly Estimating Parameters and Missing Data by Maximizing the Complete-Data Likelihood. *The American Statistician*, *37*(3), 218–220. doi: 10.2307/2683374

Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. New York: Wiley.

Mair, P., & Hatzinger, R. (2007). Extended Rasch Modeling: The eRm Package for the Application of IRT Models in R. *Journal of Statistical Software*, *20*(9), 1–20. Retrieved from http://www.jstatsoft.org/v20/i09

Masters, G. (1982). A Rasch Model for Partial Credit Scoring. *Psychometrika*, *47*(2), 149–174. doi: 10.1007/BF02296272

Mislevy, R. J. (1988). Exploiting Auxiliary Information About Items in the Estimation of Rasch Item Difficulty Parameters. *Applied Psychological Measurement*, *12*(3), 281–296. doi: 10.1177/014662168801200306

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (2005). Estimating Population Characteristics from Sparse Matrix Samples of Item Responses. *Journal of Educational Measurement*, *29*(2), 133–161. doi: 10.1111/j.1745-3984.1992.tb00371.x

Mislevy, R. J., & Wu, P.-K. (1988). *Inferring Examinee Ability when some Item Responses are Missing* (Tech. Rep.). Princeton, N.J.: DTIC Document.

Mosteller, F. (1951). Remarks on the Method of Paired Comparisons: I. The Least Squares Solution Assuming Equal Standard Deviations and Equal Correlations. *Psychometrika*, *16*(1), 3–9. doi: 10.1007/BF02313422

Neyman, J., & Scott, E. L. (1948). Consistent Estimates Based on Partially Consistent Observations. *Econometrica*, *16*(1), 1–32. doi: 10.2307/1914288

Ostendorf, F. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae* (Rev. ed.). Göttingen [u.a.]: Hogrefe, Verl. für Psychologie.

R Development Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danmarks pædagogiske Institut.

Rasch, G. (1966). An Informal Report on the Present State of a Theory of Objectivity in Comparisons. In *Proceedings of the NUFFIC International Summer Session in Science at "Het Oude Hof".* The Hague. Retrieved from `http://www.rasch.org/memo1966.pdf`

Rasch, G. (1977). *On Specific Objectivity: An Attempt at Formalizing the Request for Generality and Validity of Scientific Statements.* Retrieved from `http://www.rasch.org/memo18.htm`

Rasch, G., & Wright, B. D. (1979). *Fragments of Letters between Georg Rasch and Ben Wright regarding Development of the Rasch Model.* Retrieved from `www.rasch.org/memo19792.pdf`

Rost, J. (2004). *Lehrbuch Testtheorie - Testkonstruktion* (2nd ed.). Bern; Göttingen: Hans Huber.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika*, *63*(3), 581–592. doi: 10.1093/biomet/63.3.581

Rust, K. F. (1985). Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*, *1*(4), 381–397.

Rust, K. F., & Rao, J. (1996). Variance Estimation for Complex Surveys using Replication Techniques. *Statistical Methods in Medical Research*, *5*(3), 283–310. doi: 10.1177/096228029600500305

Saaty, T. (2008). Relative Measurement and its Generalization in Decision Making: Why Pairwise Comparisons are Central in Mathematics for the Measurement of Intangible Factors – The Analytic Hierarchy/Network Process. *Revista de la Real Academia de Ciencias Exactas, Fisicas y Naturales. Serie A. Matematicas*, *102*(2), 251–318. doi: 10.1007/bf03191825

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our View of the State of the Art.

Psychological Methods, *7*(2), 147–177. doi: 10.1037/1082-989X.7.2.147

Schnell, R. (1997). *Nonresponse in Bevölkerungsumfragen: Ausmass, Entwicklung und Ursachen.* Opladen: Leske + Budrich.

SPSS, I. C. (2010). *IBM SPSS Statistics for Windows.* Armonk, NY: IBM Corp.

Tietz, V. (1900a). Ueber eine neu Art der Preisvertheilung. *Wiener Schachzeitung*, *III*(1), 223–230. Retrieved from `http://anno.onb.ac.at/cgi-content/anno-plus?aid=sze&datum=1900`

Tietz, V. (1900b). Zum Capitel: Turnierstärken. *Wiener Schachzeitung*, *III*(1), 1–4. Retrieved from `http://anno.onb.ac.at/cgi-content/anno-plus?aid=sze&datum=1900`

von Davier, M. (1997). Bootstrapping Goodness of Fit Statistics for Sparse Categorical Data - Results of a Monte Carlo Study. *Methods of Psychological Research*, *2*(2), 29–48. Retrieved from `http://www.dgps.de/fachgruppen/methoden/mpr-online/`

von Davier, M. (2001). *WINMIRA 2001.* Groningen, The Netherlands: ASC-Assessment Systems Corporation USA and Science Plus Group.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika*, *54*(3), 427–450. doi: 10.1007/BF02294627

Watanabe, M., & Yamaguchi, K. (2004). *The Em Algorithm and Related Statistical Models.* New York; Basel: CRC Press.

Willse, J. T. (2011). Mixture Rasch Models With Joint Maximum Likelihood Estimation. *Educational and Psychological Measurement*, *71*(1), 5–19. doi: 10.1177/0013164410387335

Wright, B. D. (1988). The Efficacy of Unconditional Maximum Likelihood Bias Correction: Comment on Jansen, van den Wollenberg, and Wierda. *Applied Psychological Measurement*, *12*(3), 315–318. doi: 10.1177/014662168801200309

Wright, B. D. (1996). Comparing Rasch Measurement and Factor Analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 3–24. doi: 10.1080/10705519609540026

Wright, B. D., & Masters, G. N. (1982). *Rating Scale Analysis.* Chicago: MESA Press.

Wu, M., Adams, R. J., Wilson, M., & Halda. (2012). *ACER ConQuest: Generalised Item Response Modeling Software.* Melbourne: ACER.

Zermelo, E. (1929). Die Berechnung der Turnier-Ergebnisse als ein Maximumproblem der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, *29*(1), 436–460. Retrieved from http://gdz.sub.uni-goettingen.de/dms/load/img/?PPN=GDZPPN002370808&IDDOC=82727

Zwinderman, A. H. (1995). Pairwise Parameter Estimation in Rasch Models. *Applied Psychological Measurement*, *19*(4), 369–375. doi: 10.1177/014662169501900406

The R-Package pairwise

The package pairwise used in this paper is available from the website of the comprehensive R Archive Network at http://CRAN.R-project.org/package=pairwise. Within the package pairwise the item parameters are explicitly calculated with the function pair() using an non-iterative pairwise comparison approach (Choppin, 1968, 1985). This method provides least squares estimates of the item parameters (Garner & Engelhard, 2000). Person parameters can be estimated within the package under the assumption of fixed item parameters following a weighted likelihood approach introduced by Warm (1989), using the function pers(). For the resulting objects out of these two main functions several plotting and summary S3–Methods are available. Item- and person fit statistics are provided by the functions pairwise.item.fit() and pairwise.person.fit() respectively. These fit statistics are calculated based on the squared and standardized residuals (Wright & Masters, 1982, p. 100), based on the observed and expected person-item matrix. To detect multidimensionality within a set of Items a rasch–residual–factor–analysis proposed by (Wright, 1996) and further discussed by Linacre (1998) can be performed using the function rfa().