Journal of Person-Oriented Research
2016, 2(1–2)
Published by the *Scandinavian Society for Person-Oriented Research*
Freely available at http://www.person-research.org
DOI: 10.17505/jpor.2016.09

The Rasch-Model From an Individual's Perspective: The Item Rank Plot and the Compensation Test

Rainer W. Alexandrowicz[1]

[1] Alps-Adria-University Klagenfurt, Institute for Psychology, Applied Psychology and Methods Research Department

Contact
rainer.alexandrowicz@aau.at

How to cite this article
Alexandrowicz, R. W. (2016). The Rasch-Model From an Individual's Perspective: The Item Rank Plot and the Compensation Test. *Journal of Person-Oriented Research*, *2*(1–2), 87–101. DOI: 10.17505/jpor.2016.09

Abstract: The present study takes a closer look at the principles of estimating person parameters in the Rasch-Model and how they can be utilized for assessing model fit. After working out how the item parameters correspond to the person parameters and their standard errors, an order criterion is proposed, allowing for a further model check taking the person-oriented point of view into consideration. A simulation study established a means for an inferential check extending the assessment of model fit to the person side of the model. This method sets out to add to the existing methods of model checking and to allow for a deepened understanding of how our data correspond with the assumptions of the Rasch-Model.

Keywords: Rasch-Model, model fit, parameter estimation, Likelihood Ratio Test, Compensation Test

Introduction

The Rasch-Model is a widely used tool for (but not limited to) psychological and educational measurement. It allows for statements regarding a latent trait based on dichotomous responses. One of its major advantages is that we can reject its admissibility for a data set for empirical reasons and thus formulate a statement regarding the instrument (a psychological test, for example) used therein. Hence, the assessment of fit plays a major role and much effort has been put into the development of sophisticated methods for that purpose. The focus of these methods is on the item parameters, as will be detailled below.

In contrast, the person parameters are less frequently taken into consideration, even though the application of a psychological test aims in many cases at describing an individual. One domain, in which the person side of the model is taken into account, is the assessment of person fit, i.e., quantizing in a standardized manner the plausibility of a specific response vector given the estimated model parameters. A general embedding of the Rasch-Model into person-oriented research give von Eye, Bergmann, and Hsieh (2015, esp. pp. 825–827).

The present article approaches the question of model fit paying particular attention to the person parameter estimates. It starts with an introduction to the basics of the Rasch-Model with a special focus on how the item parameters affect the person parameter estimates and their standard errors. A few ad-hoc simulations and illustrations enhance this section and underline some important but rarely discussed details, resulting in recommendations for practitioners and test constructers. Next, the assessment of model fit is taken into consideration, focussing on the conditional Likelihood Ratio Test. Finally, a new criterion is proposed, which takes the person-oriented point of view into account. It will be shown that such an approach may improve the assessment of fit of the Rasch-Model.

The Model

The dichotomous logistic model according to Rasch (1960), henceforth denoted Rasch-Model (RM), is a discrete probability model for a binary response $X_{vi} \in \{0, 1\}$ of an in-

dividual v ($v = 1 \ldots n$) to an item i ($i = 1 \ldots k$). Let the realization $x_{vi} = 1$ denote the individual solving the task or endorsing a statement, and 0 the opposite.

The RM provides two real-valued parameters, θ_v describing the individual (in the context of an assessment frequently termed "person ability parameter") and β_i signifying the item (frequently termed "item difficulty parameter"). Using the logistic function the response probability is

$$P\big(X_{vi} = 1 \mid \theta_v, \beta_i\big) = \frac{e^{\theta_v - \beta_i}}{1 + e^{\theta_v - \beta_i}} =: p_{vi}. \tag{1}$$

Accordingly, the probability of a negative response is $1 - p_{vi} = (1 + \exp(\theta_v - \beta_i))^{-1}$. The inverse function of (1) is the logit function

$$\operatorname{logit}\big(p_{vi}\big) = \log\left(\frac{p_{vi}}{1 - p_{vi}}\right) = \theta_v - \beta_i. \tag{2}$$

The RM is a member of the exponential family, hence sufficient statistics exist and maximum likelihood theory is applicable.

The statistics $R_v = \sum_{i=1}^{k} X_{vi}$ and $S_i = \sum_{v=1}^{n} X_{vi}$ are sufficient for θ_v and β_i, respectively. Hence, all individuals with the same score r_v are assigned the same person parameter estimate $\hat{\theta}_v = \hat{\theta}_{r_v}$ (or, shorter, $\hat{\theta}_r$), and all items with the same sum s_i will be assigned the same item parameter estimate $\hat{\beta}_i$. We can therefore express equation (1) also as $p_{vi} =: p_{ri}$ (with $r = r_v$).

This feature allows for establishing a connection to the person-oriented perspective: The so-called "fifth tenet of person-oriented research" states that although there is theoretically an infinite number of possible patterns (here in a more general meaning than with the dichotomous responses considered in the Rasch-Model), "*the number of meaningful patterns is finite*" (von Eye et al., 2015, p. 799). In that sense, the Rasch-Model could be considered as a very radical translation of Tenet V. We will take up this point later.

Parameter Estimation

To obtain parameter estimates, we set the partial derivatives of the likelihood function

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{X}) = \prod_v \prod_i \frac{e^{x_{vi}(\theta_v - \beta_i)}}{1 + e^{\theta_v - \beta_i}} \tag{3}$$

equal to zero and solve for the unknown parameters. Taking the sufficient statistics into consideration, we can rewrite (3) without the individual responses x_{vi},

$$L(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{r}, \mathbf{s}) = \frac{e^{\sum_v r_v \theta_v - \sum_i s_i \beta_i}}{\prod_v \prod_i (1 + e^{\theta_v - \beta_i})}. \tag{4}$$

This formulation shows that all response matrices $\mathbf{X}$ yielding the same marginals are equally probable under the RM. However, rather than using Equation (4), we gain further from taking the natural logarithm, $\mathscr{L}(\cdot) = \log L(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{r}, \mathbf{s})$, yielding the following *support function* (cf. Edwards, 1972/1992)

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta}, \boldsymbol{\beta}; \mathbf{r}, \mathbf{s}) = &\sum_{v=1}^{n} r_v \theta_v - \sum_{i=1}^{k} s_i \beta_i \\ &- \sum_{v=1}^{n} \sum_{i=1}^{k} \log\big(1 + e^{\theta_v - \beta_i}\big).\end{aligned} \tag{5}$$

To identify the location of maximum support, we use the *(Fisher) scoring function*, i.e., the first partial derivatives of (5) with respect to the model parameters. Thus, we obtain the expressions

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \theta_v} &= r_v - \sum_i \frac{1}{1 + e^{\theta_v - \beta_i}} \cdot e^{\theta_v - \beta_i} \\ &= r_v - \sum_i p_{vi}\end{aligned} \tag{6a}$$

for the person parameters and

$$\begin{aligned}\frac{\partial \mathscr{L}}{\partial \beta_i} &= -s_i - \sum_v \frac{1}{1 + e^{\theta_v - \beta_i}} \cdot e^{\theta_v - \beta_i} \cdot (-1) \\ &= -s_i + \sum_v p_{vi}.\end{aligned} \tag{6b}$$

for the item parameters. The score is zero at the location of maximum support, hence we set Equations (6) equal to zero. By rearranging terms we obtain the equation systems

$$r_v = \sum_i p_{vi} \tag{7a}$$

$$s_i = \sum_v p_{vi}, \tag{7b}$$

i.e., to obtain parameter estimates, we set the sufficient statistics equal to their expected values—a feature, which is distinctive for the exponential family of models.

These two equation systems can be solved iteratively, and one obtains new estimates at step t by alternately applying

$$\hat{\theta}_v^{(t)} = \log(r_v) - \log \sum_i \frac{e^{-\beta_i}}{1 + e^{\theta_v^{(t-1)} - \beta_i}} \tag{8a}$$

$$\hat{\beta}_i^{(t)} = -\log(s_i) + \log \sum_v \frac{e^{\theta_v}}{1 + e^{\theta_v - \beta_i^{(t-1)}}}. \tag{8b}$$

The likelihood function of the RM is convex over the entire parameter space, hence we can take zero as starting value for all parameters. From model Equation (1) follows that each additive transformation of one parameter can be compensated for by the respective transformation of the other one, hence the parameter estimates are unique but for an additive constant. In order to fix the scale, one item must be assigned a reference value or the mean of the item parameters is set to zero.

The item parameters are regarded as structural parameters, because, usually—or, hopefully?—much effort has been invested into constructing the items under investigation. Hence, the item set cannot be arbitrarily increased.

In contrast, the person parameters are considered incidental, as we draw respondents at random. The simultaneous estimation of structural and incidental parameters gives rise to the incidental parameter problem as formulated by Neyman and Scott (1948). While corrective procedures are available (cf. Molenaar, 1995; Wright & Douglas, 1977), issues were raised regarding their effect (cf. Baker & Kim, 2004, ch. 5.6.2). Two methods of resolution have gained popularity, *marginalization* and *conditioning*. (cf. Pawitan, 2001).

In the *Marginal* Maximum Likelihood estimation approach (MML; cf. Baker & Kim, 2004; Molenaar, 1995), we replace the incidental parameters θ_v by an appropriately chosen marginal distribution $G(\theta)$, which can be integrated out. Rather than estimating the θ_v themselves, we now only have to estimate the (meta-)parameters $\boldsymbol{\tau}$ of $G(\cdot)$, which are no longer incidental. For example, in the case of the (frequently chosen) normal distribution, we estimate the mean μ_θ $(=\tau_1)$ and the variance σ^2_θ $(=\tau_2)$ of $G(\theta)$. We thus arrive at the *marginal* likelihood function

$$L_m(\boldsymbol{\tau},\boldsymbol{\beta};\mathbf{X}) = \prod_{v=1}^{n}\int_{-\infty}^{\infty}\prod_{i=1}^{k}\frac{\mathrm{e}^{x_{vi}(\theta_v-\beta_i)}}{1+\mathrm{e}^{\theta_v-\beta_i}}\,dG(\theta). \tag{9}$$

This method implies choosing a proper distribution $G(\cdot)$, the effects of failing to do so have been analyzed by Zwinderman and van den Wollenberg (1990).

Alternatively, we can apply the *Conditional* Maximum Likelihood estimation method (CML; cf. Baker & Kim, 2004; Molenaar, 1995), which has been adapted to the RM by Andersen (1970). This approach resorts to the existence of sufficient statistics and delivers item parameter estimates by conditioning on the observed values of the R_v to estimate the item parameters. This is achieved by maximizing the conditional likelihood function

$$L_c(\boldsymbol{\epsilon};\mathbf{s}|\mathbf{r}) = \frac{\prod_i \epsilon_i^{s_i}}{\prod_{r=1}^{k-1}\gamma_r^{n_r}}, \tag{10}$$

using the substitution $\epsilon_i = \mathrm{e}^{-\beta_i}$ for ease of notation. The γ_r denotes the elementary symmetric function of order r and n_r is the number of observations realizing a score $R_v = r$ (cf. Alexandrowicz, 2012; Gustafsson, 1980; MacDonald, 1995; Verhelst, Glas, & van der Sluis, 1984; Formann, 1986). To obtain item parameter estimates, we set the first partial derivatives of the log of the conditional likelihood function (10)

$$\frac{\partial \mathscr{L}_c}{\partial \epsilon_i} = \frac{s_i}{\epsilon_i} - \sum_v \frac{\gamma^{[i]}_{r_v-1}}{\gamma_{r_v}} \tag{11}$$

equal to zero ($\gamma^{[i]}_{r-1}$ denotes the first derivative with respect to item i of the elementary symmetric function of order $r-1$). After rearranging we obtain the equation system

$$s_i = \sum_v \frac{\epsilon_i \gamma^{[i]}_{r_v-1}}{\gamma_{r_v}}, \tag{12}$$

which can be solved for the item parameters by means of the Newton-Raphson algorithm. Thereafter, we use the item parameter estimates in place of the true values and obtain person parameter estimates by applying (6a). In the remainder of this text, we rely on the conditional approach.

Standard Errors of Parameters

The second derivative of a function denotes its curvature. A stronger curvature around the maximum of the support function makes identification of this maximum easier. Hence we may take the inverse of the curvature as a measure of preciseness of the estimates, establishing the ground for the estimates' standard errors. Due to a function's right curvature at a maximum, its second derivative is negative at that location. Hence we take the negative of the second derivatives of the support function with respect to each parameter, which is the *(Fisher) information function*. In our case, these are the second derivatives of (5), i.e.

$$I(\theta_v) = -\frac{\partial^2 \mathscr{L}}{\partial \theta_v^2} = -\sum_i \frac{\mathrm{e}^{\theta_v-\beta_i}}{(1+\mathrm{e}^{\theta_v-\beta_i})^2}, \tag{13a}$$

for the person parameters and

$$I(\beta_i) = -\frac{\partial^2 \mathscr{L}}{\partial \beta_i^2} = -\sum_v \frac{\mathrm{e}^{\theta_v-\beta_i}}{(1+\mathrm{e}^{\theta_v-\beta_i})^2}, \tag{13b}$$

for the item parameters. Evaluating Equations (13) at the location of maximum support (i.e. using the maximum likelihood estimates) yields the *observed information* $I(\hat{\theta})$ and $I(\hat{\beta})$. The variance is the inverse of the information, hence the standard errors of the estimates are

$$S.E.(\hat{\theta}_r) = \frac{1}{I(\hat{\theta}_r)} \tag{14a}$$

and

$$S.E.(\hat{\beta}_i) = \frac{1}{I(\hat{\beta})}. \tag{14b}$$

At this point, we observe an asymmetry. In contrast to the S_i, the R_v can only realize a very limited number of different values, namely $1\ldots k-1$ (the limits 0 and k are not of interest in the Rasch context, as they provide no information regarding the comparison of individuals; cf. Hoijtink & Boomsma, 1995; Warm, 1989). As we see from Equations (13), the observed information regarding an item parameter is a sum involving n terms, while that of an item parameter only totals k terms. Therefore, the standard errors of the person parameters are considerably larger than those of the item parameters. From a person-oriented point of view it is interesting, how the $S.E.(\hat{\theta}_r)$ are related to the length of an instrument. Let us therefore consider an instrument, in which all items are equally difficult, i.e. $\forall i : \beta_i = 0$, after centering. The number of items k varies from 5 to 300 in steps of 5 and then in steps of 25 up to 600 items. Figure 1 shows the resulting standard errors. The horizontal axis depicts the relative score r/k. Each line represents one k, with the red lines emphasizing test lengths of $k = 10, 25, 50, 80, 100, 150, 200,$ and 300 (from top to bottom).

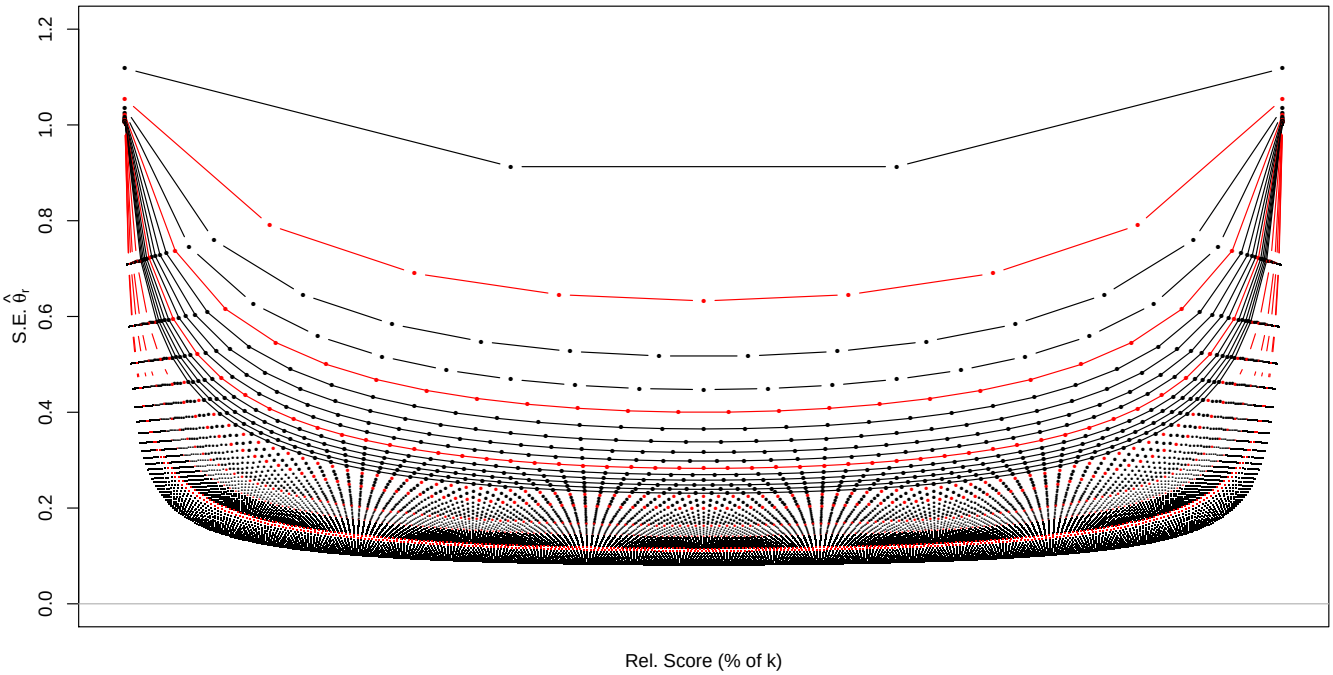

Figure 1. Standard errors of the estimated person parameters $\hat{\theta}_r$ (vertical axis) for all possible scores $r = 1 \ldots k-1$ (horizontal axis) for varying numbers of items (lines).

We see bath tub-shaped lines representing the standard errors for all levels of k. The top-most line represents a short instrument (test) of 5 items, in which the standard errors are comparably similar for all scores. An increase in the number of items results in a considerable drop in the standard errors of medium scores, while those for values of r close to 1 and $k-1$ remain high.

The red lines show that the largest gain in terms of reduction of standard error for medium scores is achieved by extending the number of items from 5 to 10 or maybe 25. But beyond 50 items, no appreciable reduction of standard errors can be achieved any more. This might serve test constructors as an orientation towards the required number of items for achieving a desired precision when assessing a testee's trait.

Note that virtually the same plots appear if we choose the β_i equidistantly from a given interval (e.g. $-5 \ldots +5$) or draw them even at random from such an interval[1]. Hence, conclusions drawn so far are not restricted to an admittedly artificial case, in which all items exhibit the identical difficulty, but generalize, in principle, to any realistic set of item parameters (one possible exception is described in the next section).

From a practical point of view, we may conclude that extremely long scales result in little gain as regards standard error of the person parameter estimates, but short scales will profit from any extension. About 15 to 25 items seem to be a reasonable choice.

[1]Interested readers can obtain the respective plots from the author upon request.

Linking Item and Person Parameters

This section illustrates how the item parameters affect the resulting person parameter estimates. For that purpose, we consider some prototypical cases, starting with an instrument comprising $k = 10$ items. Let us assume, first, that all item parameters are zero (because the $\hat{\beta}_i$ are used as if they were the true parameters in the CML context, we will omit the hat in the following; in applications, we use the CML-estimates). Solving Equation (6a) for the θ_r, we obtain a curve as shown in Figure 2 (left diagram). It displays the typical inversely S-shaped strictly monotone increasing curve, bending slightly outwards in the regions of low and high scores and running almost linear in the middle (i.e., in the region close to $k/2$). Furthermore, Figure 2 depicts increasing standard errors (reflected by larger confidence limits in the plot) for low and high scores, as less information is available for these scores (cf. Equations (13a) and (14a)). A shape similar to the one considered so far can frequently be observed in applications, because in many cases the majority of the items is of medium difficulty.

Extreme Item Parameters

Let us now change one β_i to a very extreme value, say, 7. A difference of 7 units between the easiest and the most difficult item will rarely occur, hence we may consider this a borderline case. Estimating the $\hat{\theta}_r$ yields the middle plot of Figure 2. There is a clear buckling in the sequence of the $\hat{\theta}_r$ when changing from $r = k-2$ to $r = k-1$ (i.e., from the second last to the last r). This buckling is easy to explain: Although the score r is a (minimal) sufficient statistic for $\hat{\theta}_r$ and, therefore, results in exactly the same estimate, the actual response vector $\mathbf{x}'_v = (x_{v1}, x_{v2}, \ldots, x_{vi}, \ldots x_{vk})$ is not entirely irrelevant. Rather, different response vectors

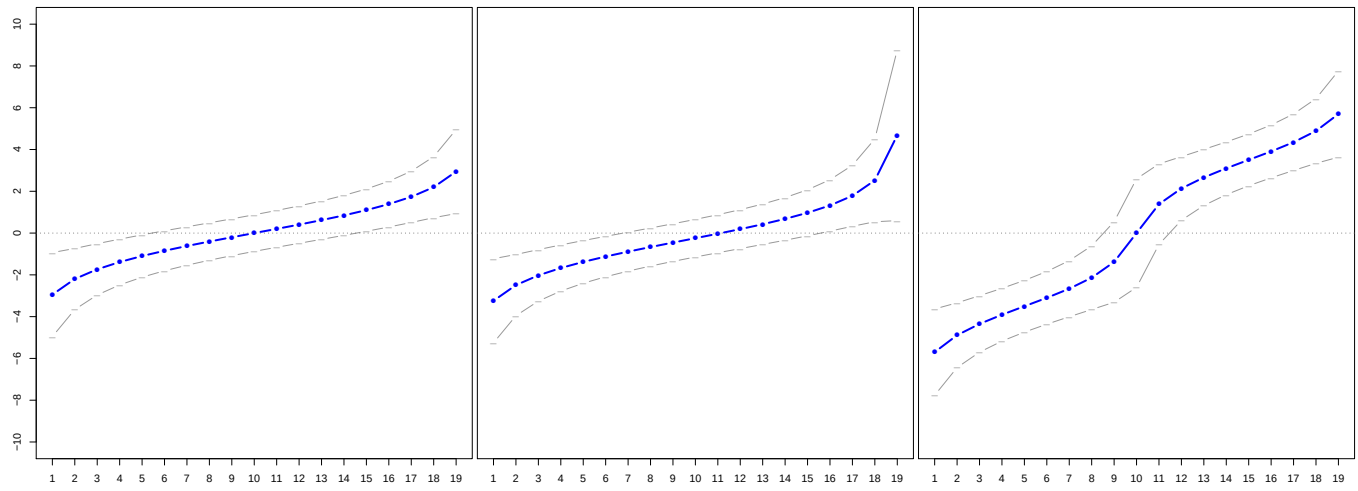

Figure 2. Estimated person parameters $\hat{\theta}_r$ (vertical axis) for all possible scores $1 \dots r$ (horizontal axis). Left diagram: all $\beta_i = 0$; middle: $\beta_1 = 7$, all others 0; right: half of the $\beta_i = 7$, all others 0. The grey lines indicate the 95% confidence limits. Note: Item parameters were centered prior to estimating the person parameters.

(yielding the same r) vary with respect to their likelihood. From all possible patterns resulting in a score r, the one with exactly the r easiest items solved has the highest likelihood. Hence we may say that the Rasch-Model "assumes" that a respondent who attains a score r has solved the r easiest items (which is also intuitively plausible). A numerical illustration is given in Appendix A.1.

From the person-oriented perspective, this feature is especially interesting for it corresponds to the so-called "fourth tenet of person-oriented research", termed *principle of pattern summary*, or, as proposed by von Eye et al. (2015), *principle of pattern as units of analysis* (p. 799).

Therefore, only an individual reaching the maximum score has—from the "model's point of view"—a chance of solving this extra difficult item. Thus, such individuals are considered extraordinarily capable and are therefore "rewarded" with an extra large parameter estimate.

To further illustrate the point, the rightmost plot in Figure 2 depicts a case, in which half of the items (i.e. 5) show a β_i of zero and the other half a value of 7. In this case, the model assumes that only respondents achieving a score of at least 6 solved the difficult ones and hence recompenses them with higher estimates. Therefore, we find the buckling in the middle of the sequence.

Such bucklings—especially when they appear at the margins—pose a possible problem for algorithms aiming at the estimation of $\hat{\theta}_{r=0}$ and $\hat{\theta}_{r=k}$: For example, the R package `eRm` (Mair, Hatzinger, & Maier, 2012) applies a spline-extrapolation from the estimated parameters for scores $r = 1 \dots k-1$ to the two extreme scores 0 and k. Such an extrapolation could fail for some of the extreme cases considered here, especially when one item differs considerably from the majority of items. A spline would not anticipate the buckling. However, this would only be the case in rare situations.

The rightmost plot in Figure 2 uncovers another important detail: The standard error in the vicinity of the buckling is larger compared to the other areas, which is a logical consequence of the item configuration: We have many items in the lowest region of the latent continuum and many in the highest region. Hence there are few (in fact: no) items at the buckling's location, which conforms to little information in the sense of equation (13), and therefore, the standard error is larger here. The same applies to the previous case, in which one item differs exceedingly from the remaining ones. Again, there is only "little" information for respondents solving all but one items, because only one item measures in this vicinity, and hence the standard error of person parameters located here must be larger.

Item Parameter Variation

The leftmost diagram in Figure 3 extends this last scenario by adding one extremely easy and one extremely difficult item while leaving the remaining items at a value of zero. Such a situation may arise in cases, in which test constructors realize that their items do not vary to a sufficient extent and therefore deliberately add extremely easy or difficult items.

As a consequence, we obtain an extremely inverse-S-shaped sequence of $\hat{\theta}_r$, resulting in a notably "flat" section in the middle, which differenciates little across most of the range, but assigns heavily deviating values for $\hat{\theta}_{r=1}$ and $\hat{\theta}_{r=k-1}$. At first sight, the situation might not be considered overly harmful, but taking the standard errors (and the corresponding confidence limits as depicted in the plots) into consideration shows that, for example, the 95%-confidence interval for $\hat{\theta}_7$ covers also the estimates $\hat{\theta}_8$, $\hat{\theta}_9$, $\hat{\theta}_{10}$, and $\hat{\theta}_{11}$. Hence, we may discriminate poorly between individuals realizing medium scores, which is slightly disadvantageous as exactly these scores occur most often. Moreover, the isolated items do not provide much information on the latent continuum, hence these extreme estimates are associated with an enormous standard error and thus also of limited value.

As a practical recommendation we can therefore conclude that extreme variation of item parameters, especially if caused by outliers (in the sense of single item difficulty

parameters far away from the majority of the items) should be considered with care. It generally impedes the interpretability of person parameter estimates.

Special Case: Implicitly Assuming Linearity

Another case seems interesting to explore: One might disregard the RM and use the scores directly for further evaluation. In this case, one not only assumes the model to hold (unexaminedly), but further assumes that the person parameters exhibit a perfect 1 : 1 relation with the score, i.e., $\forall r : \hat{\theta}_{r+1} = \hat{\theta}_r + c$. Such a relation holds, when the item parameters themselves are equidistantly spaced, i.e., $\forall i : \beta_{i+1} = \beta_i + d$. However, Figure 3 (middle and right plot) shows that even this assumption would not yet suffice to obtain a perfectly linear relationship.

The plot in the middle shows the person parameter estimates when the (in this case $k = 20$) items have equal distances ranging from -5 to $+5$. The sequence of the $\hat{\theta}_r$ is almost linear, as we see from the comparison with the superimposed regression line in red. Only the outmost values θ_1 and θ_{k-1} indicate a slight outward deviation. If we extend the values of the β_i to the interval of $-20 \ldots +20$, the linearity is even more pronounced (right diagram; mind the different scaling of the vertical axis). A perfectly linear relationship would be realized if the item parameters ranged from $-\infty$ to $+\infty$, which is impossible to realize. However, from a practical point of view, sufficient linearity might be achieved, but this would require a rather particular arrangement of the item parameters.

Practical Considerations

Let us now consider some more realistic cases and draw repeatedly item parameters randomly from a uniform distribution with limits $-b$ to b and oppose these with the resulting person parameter estimates. The number of items varies from k =10 to $k = 50$ (in steps of 1) and the betas are drawn in turn from $U(-1,1)$, $U(-2,2)$, $U(-3,3)$, and $U(-4,4)$. For each k one sample of betas was drawn.

Figure 4 superimposes the *sorted* β_i (red lines) and the resulting $\hat{\theta}_r$ (blue lines) for each draw of $\boldsymbol{\beta}$. To make the sequences of the $\hat{\theta}_r$ comparable, the horizontal axis again shows the relative score r/k. The grey lines indicate the limits of the uniform distributions the betas were drawn from.

Interestingly, the sequences of the $\hat{\theta}_r$ differ hardly ever if the betas are similar in value (i.e. drawn from a $U(-1,+1)$; top left plot of Figure 4). With an increasing range of item parameters, the sequences become somewhat more varied, but only to a limited extent. By and large, no substantial change of person parameters appears even if the items cover the typical range of -4 to $+4$ (bottom right plot of Figure 4).

We may therefore conclude that in cases, in which no blatant particularity of the item parameters (like those described in the previous sections) appears, the person parameters are more or less predictable from the score, no matter the item parameters. For example, an individual solving (or responding positively to) about 80% of the items will obtain a parameter of approximately 1.5 if the item parameters lie in the interval $[-1,+1]$, a value of approximately 1.8 for the interval $[-2,+2]$, a value of approximately 2.2 for the interval $[-3,+3]$, and a value of approximately 2.6 for the interval $[-4,+4]$, irrespective of the number of items. Moreover, taking also k into account, we can even derive a rough estimate of the standard error from Figure 1.

Specific Objectivity

Let

$$
\begin{aligned}
p_{vi} &=: P(X_{vi}=1|\theta_v,\beta_i) = \frac{e^{\theta_v-\beta_i}}{1+e^{\theta_v-\beta_i}} \\
p_{vj} &=: P(X_{vj}=1|\theta_v,\beta_j) = \frac{e^{\theta_v-\beta_j}}{1+e^{\theta_v-\beta_j}} \\
p_{wi} &=: P(X_{wi}=1|\theta_w,\beta_i) = \frac{e^{\theta_w-\beta_i}}{1+e^{\theta_w-\beta_i}}
\end{aligned}
$$

with $i \neq j$ and $v \neq w$. The logits of the respective probabilities are $\text{logit}(p_{vi}) = \theta_v - \beta_i$, $\text{logit}(p_{vj}) = \theta_v - \beta_j$, and $\text{logit}(p_{wi}) = \theta_w - \beta_i$. Taking the ratio of the logits of either two items and one person or two persons and one item shows that in the former case, the person parameter cancels out and in the latter case the item parameter. A graphical representation is given in Figure 5. In the left diagram we see that the distance of the two item curves equals the difference of the two logits, Δ_{ij}, irrespective of the location of the individual θ_v; the right diagram shows that the logit difference with respect to the two individuals remains constant at Δ_{vw}, irrespective of the item used for comparison.

Therefore, if the model holds, we can compare items (i.e. estimate item parameters unbiasedly) using (almost) any selection of individuals (properly: irrespective of the distribution of the person parameters); likewise, we can compare individuals using any proper set of items (cf. Rasch, 1966a; Rasch, 1966b). This is the algebraic foundation for several advantageous features of the RM, which include supporting adaptive testing, testlet building, or yielding unbiased item parameter estimates even from non-representative samples. Rasch has termed this feature "specific objectivity".

Model Tests

Numerous methods for assessing the fit of the RM have been proposed. Glas and Verhelst (1995), for example, give an overview of many. The present article focusses on a method, which is intrinsic to the RM. The specific objectivity property of the model allows for a rigid assessment of the model adequacy. Rasch (1960) pointed out that *"If a relationship between two or more statistical variables is to be considered really important, (...) the relationship should be found in several sets of data which differ materially in some relevant respects"* (p. 9). In terms of the RM this means that item parameter estimates will not differ across subsamples but for random variation. In the person-oriented

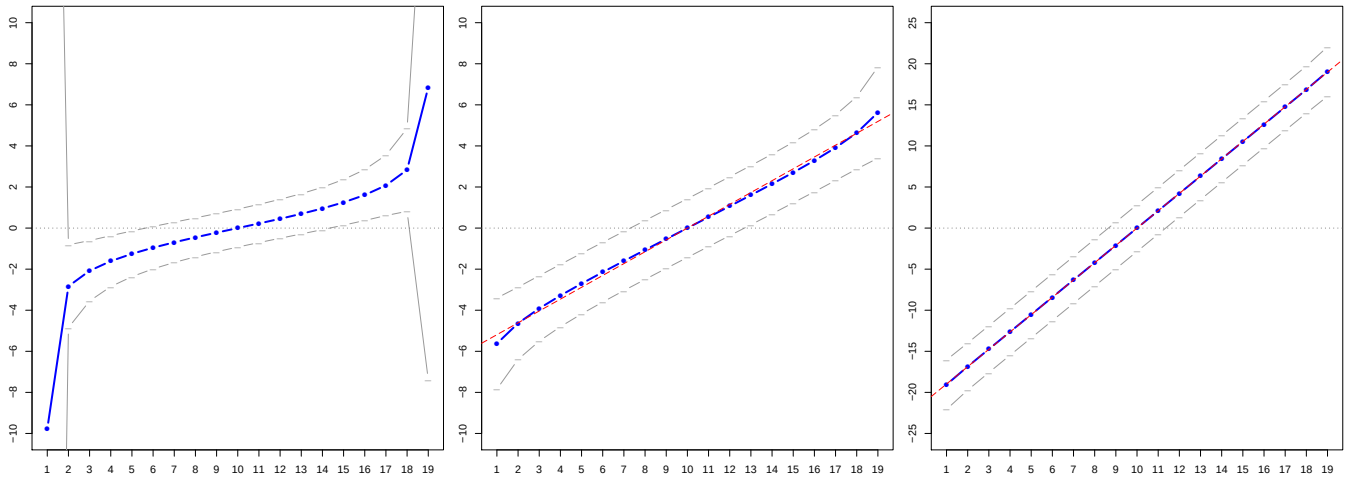

Figure 3. Estimated person parameters $\hat{\theta}_r$ (vertical axis) for all possible scores $1 \ldots r$ (horizontal axis). Left diagram: $\beta_1 = -20$, $\beta_k = +20$, all remaining $\beta_i = 0$; middle: $\beta_i = -5 \ldots +5$, equidistantly spaced; right: $\beta_i = -20 \ldots +20$, equidistantly spaced. Note: Item parameters were centered prior to estimating the person parameters

research tradition, this concept is known as *dimensional identity* ("seventh tenet"; cf. von Eye, 2010, p. 279; von Eye et al., 2015, p. 799).

Andersen (1973) developed from Rasch's conclusion a conditional Likelihood Ratio Test (cLRT) using the test statistic

$$\Lambda = -2 \log \frac{L_c\left(\widehat{\boldsymbol{\beta}} \mid \mathbf{r}\right)}{\prod_s L_c\left(\widehat{\boldsymbol{\beta}}^{(s)} \mid \mathbf{r}^{(s)}\right)} \tag{15}$$

with $\hat{\boldsymbol{\beta}}$ the vector of the item parameter estimates derived from the entire sample, $\mathbf{r}$ the vector of the sufficient statistics of the entire sample, and $\hat{\boldsymbol{\beta}}^{(s)}$ and $\mathbf{r}^{(s)}$ the respective estimates and statistics from subsamples $s = 1 \ldots S$. If the model holds, the test statistic is approximately distributed s χ^2 with $(k-1)(S-1)$ degrees of freedom. The subsamples may be obtained by splitting the sample by score (e.g. using the score median) or a substantial criterion like gender, treatment, or any other relevant criterion. While we may not prove the null-hypothesis of model fit, repeated failure to reject it (i.e., using several split criteria) increases its degree of corroboration (cf. Popper, 1959/2010, p. 67).

The Role of the Sample Size

It is typical for any inferential assessment that large samples may yield significant test results for trivial effects, whereas, with a small sample substantial effects may go undetected. In order to prevent both kinds of misleading decisions, we have to determine the optimal sample size allowing for the detection of an effect, which is considered meaningful from a substantial point of view with a given risk α for an error of the first kind and a given risk β for an error of the second kind. While such calculations are readily available for most tests (e.g. Cohen, 1988), no solution has been developed for the cLRT until recently. Draxler and Alexandrowicz (2015) have identified the non-central χ^2-distribution required for the power analysis of the cLRT, which allows for determining the probability of an error of the second kind for a given (or substantively interesting) model violation.

To determine the appropriate non-central distribution of the test statistic, we have to find a proper effect size measure, which allows one to identify the non-centrality parameter of the respective distribution. If the model holds, the probabilities of a correct response do not differ across subsamples, hence the null-hypothesis can be written as

$$H_0 : p_{ri}^{(1)} = p_{ri}^{(2)} = \ldots = p_{ri}^{(s)} = p_{ri}^{(0)} \tag{16}$$

with $p_{ri}^{(s)}$ denoting the probability of a correct response of subsample s using $\hat{\beta}_i^{(s)}$ and $p_{ri}^{(0)}$ denoting the probability of a correct response based on the item parameter estimates of the entire sample. We may then define a model violation as

$$\delta_{ri}^{(s)} = p_{ri}^{(s)} - p_{ri}^{(0)}. \tag{17}$$

This formulation is equivalent to the assumption of equal item parameter estimates across subsamples by extending Equations (7) to $r^{(s)} = \sum_i p_{ri}^{(s)}$. Fixing the deviation $\delta_{ri}^{(s)}$ a priori to a value of substantive interest allows one to determine the optimal sample size n^* required for detecting this violation with predefined risks of errors of the first and the second kind. This solution of Draxler and Alexandrowicz (2015) is a helpful contribution, dealing with the fundamental problem of over- or underpowered model tests in the CML context.

If only a small sample is available (e.g., because power analysis indicated it or only a limited number of respondents was available), we also have to consider the speed of approximation of the test statistic (15) to its limiting distribution. This aspect has been covered extensively in Alexandrowicz and Draxler (2016).

Effect and Impact

It is a constitutive feature of the CML approach to focus on the items. But no consideration of the person parameters

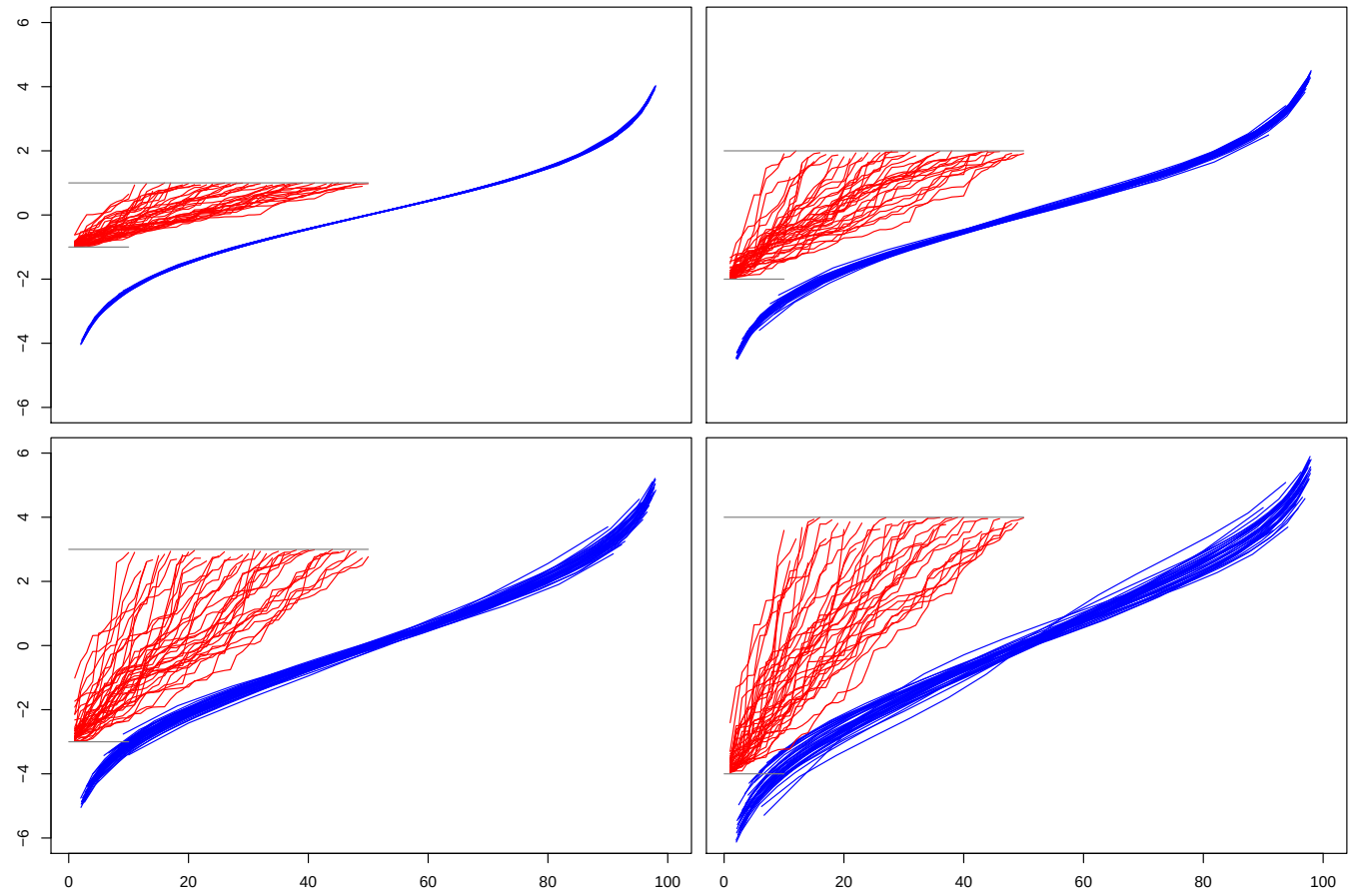

Figure 4. Item parameter sets (red; sorted by size) and the resulting person parameter sequences (blue) for $k = 10, \ldots, 50$. Horizontal axis: relative score $100 \cdot r/k$ (regarding the blue lines) and item number $i = 1 \ldots k$ (regarding the red lines), respectively; vertical axis: β_i and the resulting $\hat{\theta}_r$ (superimposed). Top left: $\beta_i \in U(-1, +1)$; top right: $\beta_i \in U(-2, +2)$; bottom left: $\beta_i \in U(-3, +3)$; bottom right: $\beta_i \in U(-4, +4)$;

(or, more precisely, their estimates) takes places when assessing model fit. We will, therefore, extend the inferential assessment of model fit by taking the person-oriented view into consideration. When assessing model fit by means of ascertaining item parameters' equivalence across subsamples obtained by splitting along criteria of substantive interest, we have to consider the equivalence of the person parameters' estimates as well.

Let us, therefore, term item parameter differences across subsamples as the *effect* that is to be detected with a desired power $1 - \beta$, and *impact* as the resulting difference of the resulting person parameters' estimates. As has been shown before, the item configuration will affect the sequence of the person parameters' estimates with regard to the score r. We will extend our considerations to the comparison of the $\hat{\theta}_r^{(s)}$ after splitting the sample into S subsamples. We consider the two group split ($S = 2$), first, because it allows for a clearly arranged presentation, and second, because it constitutes the most frequently applied split in applications.

Figures 2 and 3 above illustrated, how the item parameters' configuration affects the sequence of the $\hat{\theta}_r$. When we turn to the assessment of model fit, we have to ascertain, whether and how these sequences change across subgroups. Figure 4 indicates that the item parameter estimates seem to be only marginally affected by the actual item parameters, hence little is to be expected from an inspection of the $\hat{\theta}_r^{(s)}$. But this is in fact not necessarily the case, as will be shown in the following section.

Effect versus Impact

One might come up with the idea of directly comparing an ad-hoc measure of effect and impact as defined before. This could—in the two-group-split—be accomplished by the root-mean-square deviation (RMSD)

$$\Delta_{\boldsymbol{\beta}} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\beta_i^{(1)} - \beta_i^{(2)}\right)^2} \tag{18a}$$

$$\Delta_{\hat{\boldsymbol{\theta}}} = \sqrt{\frac{1}{k-1}\sum_{r=1}^{k-1}\left(\hat{\theta}_r^{(1)} - \hat{\theta}_r^{(2)}\right)^2}. \tag{18b}$$

A little simulation reveals that such an approach is only of limited value: Draw $k = 10$, 30, and 50 item parameters randomly from a $U(-3, 3)$ representing the $\boldsymbol{\beta}^{(1)}$ and add an error to each item, $e_i \sim U(-2, 2)$ yielding the $\boldsymbol{\beta}^{(2)}$. Estimate the respective $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(2)}$ and determine the RMSD according to Equations (18a) and (18b). Repeat this procedure 10,000 times.

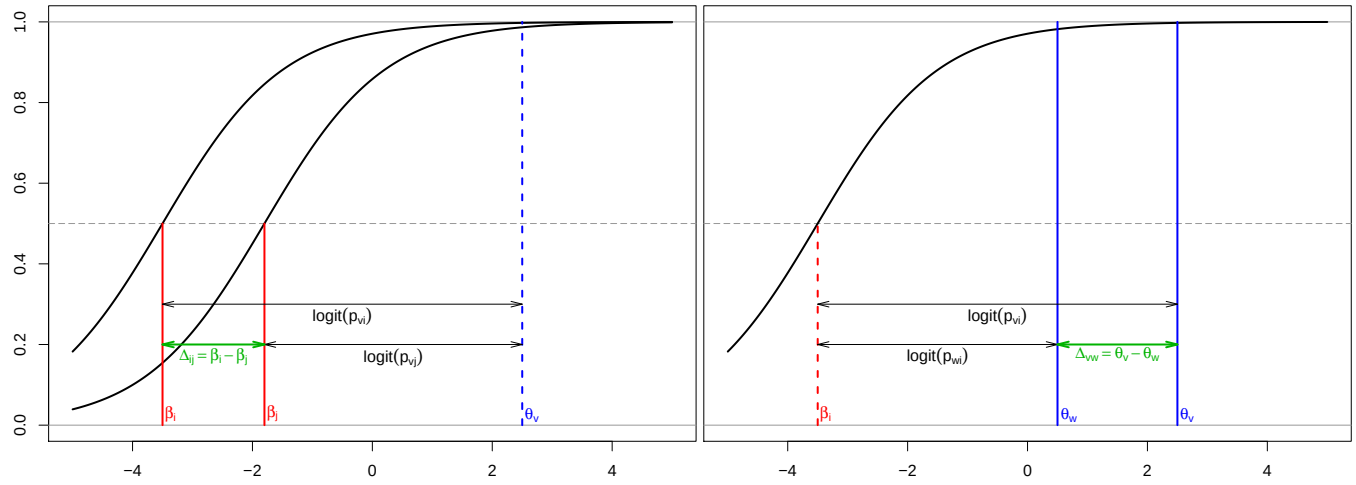

Figure 5. Illustration of the constant logit differences regarding two items (left plot) and two persons (right plot). Because the locations of θ_v in the left plot and the β_i in the right plot can be modified without affecting the respective differences Δ_{ij} and Δ_{vw}, hence their indicators (the blue line in the left diagram and the red line in the right diagram) are drawn as dashed lines.

Figure 6 (left diagram) opposes the $\Delta_{\boldsymbol{\beta}}$ and the $\Delta_{\hat{\boldsymbol{\theta}}}$ with colors indicating the scale length k. Clearly, there is no linear relationship between the two measures, but rather a triangle-shaped one. Large differences on the item side can be associated with both large and small differences on the person side. The corresponding correlation coefficients are $r = 0.301$ ($k = 10$), $r = 0.315$ ($k = 30$), and $r = 0.297$ ($k = 50$). All dots appear beneath the identity line, hence subsample differences of the person parameter estimates are generally smaller in value than those of the item parameters.

This effect can easily be explained by the characteristics described above: The person parameter estimate relies on the configuration of the item parameters, but not on the response vector itself. It is therefore irrelevant, which items a person has responded positively to, only the score matters. If, for example, one item has subgroup parameters $\beta_i^{(1)} = -1$ and $\beta_i^{(2)} = +1$ (i.e. differs considerably), and another item has parameters $\beta_{i'}^{(1)} = +1$ and $\beta_{i'}^{(2)} = -1$ (i.e. differs considerably as well), their combined appearance causes the person parameter estimates to remain entirely unaffected. The two items have compensated their role in the two subsets. If such compensation phenomena occur frequently between the subsets, $\Delta_{\boldsymbol{\beta}_i}$ and $\Delta_{\hat{\boldsymbol{\theta}}}$ correspond to entirely different items and thus lack comparability. This caused the low correlation observed in Figure 6 and hampers conclusions from effect upon impact. We therefore will, if such compensations occur, not be able to evaluate the consequences of item parameter differences between the subsamples with respect to differences in the resulting person parameter estimates $\hat{\boldsymbol{\theta}}_r^{(s)}$.

We must rather consider the *ordered* sequence of the item parameters, which shall be denoted $\beta_{[i]}$, i.e., $\beta_{[1]}$ is the item with the smallest parameter (easiest item), $\beta_{[2]}$ the one with the second smallest parameter, and so on, up to $\beta_{[k]}$ the item with the largest parameter (most difficult item). We can therefore extend Equations (18) and add the respective RMSD for the ordered item parameters

$$\Delta_{\boldsymbol{\beta}^*} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\beta_{[i]}^{(1)} - \beta_{[i]}^{(2)}\right)^2}. \tag{18c}$$

Using $\Delta_{\boldsymbol{\beta}^*}$ rather than $\Delta_{\boldsymbol{\beta}}$ in the simulation, we obtain the plot shown in Figure 6 (right diagram). Clearly, the strength of the relationship of the two measures is greater than before, with $r = 0.714$ ($k = 10$), $r = 0.722$ ($k = 30$), and $r = 0.727$ ($k = 50$).

An Order Criterion

Obviously, Δ_{β^*} captures more of what constitutes a deviation from a person oriented point of view. Remember that the model "assumes" the r easiest items have been solved, hence the item ordering gains importance. If an item changes its position across subgroups, the person parameter estimates refer to a different set of items. If the model holds, we can consider the entire set of items unidimensional, hence it makes no difference. But—and this is, what the cLRT is after—if the items change their location, the assumption becomes increasingly questionable. The item-based approach takes the numerical differences of the $\hat{\beta}_i^{(s)}$ across the subsamples into consideration, which constitutes a purely quantitative measure. In contrast, the person-oriented perspective has to consider the item ordering as well, i.e., we introduce a qualitative aspect: The occurrence of (relevant) changes in the location of items contradicts the assumption of model fit.

Figure 7 proposes a simple graphical means to recognize such exchanges by juxtaposing the subgroup estimates on two separate lines in a stripchart-like style. Solid lines connect the individual items, while the dotted red lines connect the items according to their ranks. Hence, the latter may in case of rank exchanges connect different items (in our example: items 3 and 7 and, to a lesser extent, items 5 and 8). An R script for the plot shown in Figure 7 along with an example call is given in Listing 1.1 in Appendix A.2.

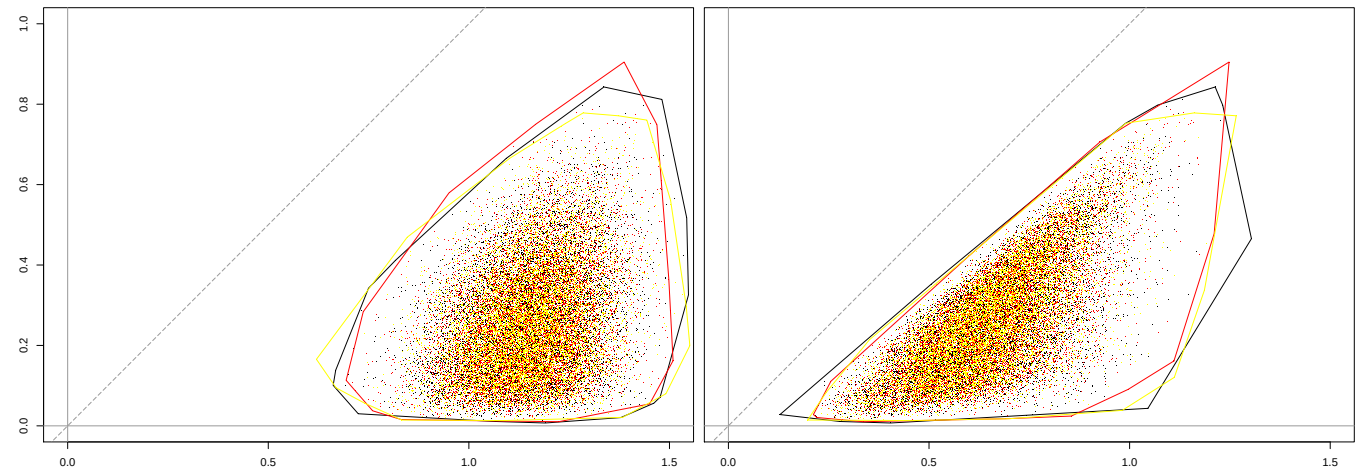

Figure 6. Left diagram: Contrasting Δ_{β_i} (horizontal axis) and $\Delta_{\hat{\theta}}$ (vertical axis); right diagram: Contrasting Δ_{beta^*} (horizontal axis) and $\Delta_{\hat{\theta}}$ (vertical axis). The colors indicate the number of items (black: $k = 10$; red: $k = 30$; yellow: $k = 50$). The convex hulls according to each k are superimposed in the respective colors.

In Figure 7, we can differentiate three prototypical cases: (*i*) an item retains its position (in our example, these are items 6, 1, 2, and 4); (*ii*) items change position, but the difference is small (items 8 and 5); and (*iii*) items change their position with a large shift (items 3 and 7). While case (*i*) represents the ideal situation, case (*ii*) may as well occur, but could be considered more or less harmless. In contrast, case (*iii*) is what we are looking for—or even a more serious case (*iv*), in which items switch several positions (not appearing in our example).

When items are of similar difficulty, switching may appear more frequently than when they cover a broad range of values. When there are many items, switching is even more likely to happen, because the range of parameter estimates will not grow considerably and, therefore, items lie necessarily closer to each other. Hence, it is unlikely that no switching appears at all, even if a data set conforms very well to the model. We have to find out, what can be expected under a valid null-hypothesis and what should raise our concerns.

Approaching the H_0-distribution

A simple means for summarizing the mis-ordering of items across subsamples is to count the number of inversions appearing between subsamples in cases, in which the model holds. For that purpose, a simulation study was undertaken. It determines the distibution of rank exchanges for sample sizes of $n/2 = 100$, 250, and 500 and for test length of $k = 5$, 10, 20, and 30 items. Item parameters were drawn randomly from a $U(-2, 2)$. Two subsamples of size $n/2$ were generated in line with the RM using one item parameter set, and then merged. The person parameter estimates were determined for each subsample and the rank differences were calculated. This procedure was repeated 10,000 times per sample. Figure 8 shows a histogram of the distributions of the rank differences. Further, a normal curve (orange) and a Poisson curve (blue) were superimposed, using the observed mean and (for the normal) the standard deviation of the observed values were used.

First of all, Figure 8 shows that, with an increasing number of items, a certain number of rank differences are likely to appear. Only the short instrument with 5 items shows a considerable number of zero rank differences. All distributions are skewed to the right but to a lesser extent, the more items we have. Regarding the shape of the distributions, the Poisson seems a sensible candidate, especially for small k. With increasing length of the instrument, the Poisson and the normal curve become more similar, which is in line with theory.

One could use this distribution for testing the null-hypothesis that the observed number of rank differences is compatible with the number of rank differences occurring when the model holds. Hence, if the observed number of item rank differences is a member of the $100 \times \alpha$ percent most extreme values of the bootstrap generated distribution, it could be considered significant. We might therefore expand our decision on model fit to examining the invariance of the item parameters (via the well-established cLRT) on the one hand and the existence of compensation effects on the other hand. The proposed test could therefore be termed *Compensation Test*.

Let us expedite the supposition that this distribution resembles a Poisson distribution by comparing quantiles of the bootstrap distributions with the limiting ones. A frequently used decision criterion is the 95%-quantile, which is used in Table 1. We see that in 4 of the considered distributions (5/100, 20/100, 20/500, and 30/500), the quantiles differ by 1, while the remaining ones are equal. Hence, the Poisson seems to allow for a useful approximation. However, further examination is required to evaluate this conjecture.

Alternatively, we could also compare the rank differences in the two-sample-case with the Wilcoxon-Mann-Whitney-test (or U-Test; Mann & Whitney, 1947; Wilcoxon, 1945; for a more recent treatment see Wiedermann & Alexandrowicz, 2007). However, considering the fact that usually a rather limited number of items is analyzed, we

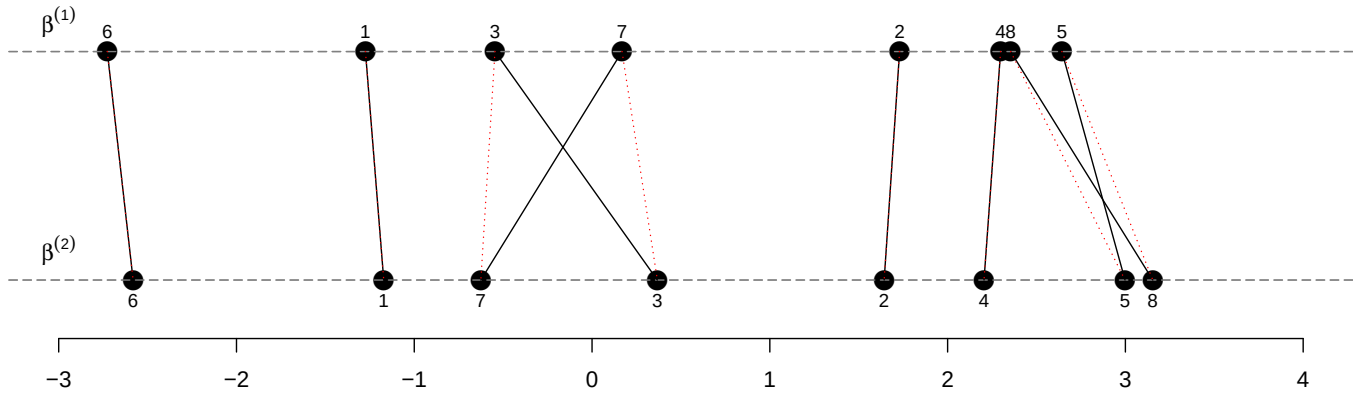

Figure 7. An example Item Rank Plot for 8 items.

Table 1. Comparison of the observed 95%-quantiles of the bootstrap distributions and of a Poisson distribution with $\lambda = \bar{x}_{\text{bootstrap}}$

k	n	obs.	theor.	diff.
5	100	2	3	−1
5	250	2	2	0
5	500	2	2	0
10	100	7	7	0
10	250	5	5	0
10	500	4	4	0
20	100	20	21	−1
20	250	15	15	0
20	500	11	12	−1
30	100	41	41	0
30	250	29	29	0
30	500	22	23	−1

should not expect too much from this test, as its power will be considerably small.

Worked Example

To exemplify the proposed procedure, let us consider a data set, which has been used in Alexandrowicz, Fritzsche, and Keller (2014). In brief, the study analyzed compared a clinical and a non-clinical population with respect to the applicability of the Beck Depression Inventory Version II (BDI-II; Beck, Steer, & Brown, 1996; german version Hautzinger, Keller, & Kühner, 2009). From that study, only the students' data shall be analyzed ($n = 468$) and responses were dichotomized (0 vs. 1+) to fit the present frame of reference. One respondent anwsered only questions 1 to 10 and was therefore omitted from analysis; the remaining 27 missing values (0.28% of all responses) were scatterd across the data set and replaced by zero.

The LRT using the score median split resulted in a χ^2 of 38.5 ($df = 20$, $p = 0.008$) indicating that some discrepancies exist between the two split groups. A logical next step would involve identifying possibly deviating items, however this is not the focus of the present study. Rather, we will continue with the person-oriented analysis and consider the *effect* and the *impact* as defined above. The raw RMSD according to equation (18a) was $\Delta_{\beta} = 0.441$ and the corrected one following equation (18c) was $\Delta_{\beta^*} = 0.306$. The *impact* according to equation (18b) was $\Delta_{\theta} = 0.081$. Considering the descriptive results as shown in Figure 6, the values could be considered small—however, such evaluations are only tentative at the moment.

A total of 20 rank exchanges occured and Figure 9 (in Appendix A.3) shows the Item Rank Plot for this split. Let us pick out items number 2 and number 16 to illustrate the message: Item 16 shows a comparably large shift of its difficulty estimate, but remains the easiest in both samples. In contrast, item 2 shows a similar difference but, moreover, it changes its position by 3 ranks (from 4$^{\text{th}}$ most difficult to 7$^{\text{th}}$ most difficult). Both example items indicate subsample differences not in line with the parameter invariance assumption. But item 2 will also affect the person parameter estimate in the sense that the model assumes (for example) that an individual realizing a score of 15 is likely to have solved this item in subgroup 2 but not in subgroup 1.

The Wilcoxon-Mann-Whitney-U-Test resulted in a test statistic of 216 ($p = 0.92$). Hence, this test would not indicate any appreciable shift of item parameters across subsamples. But as has been argued before, this test could be underpowered. For that reason, also the proposed compensation test has been applied using a parametric bootstrap. Scores 0 and k were handled according to a method discussed in Alexandrowicz and Draxler (2016), namely using the WLE estimates $\hat{\theta}_0$ and $\hat{\theta}_k$ for the two extreme scores and the ML estimates for the remaining scores 1 to $k-1$ when simulating the bootstrap data sets. This analysis yielded a p-value of 0.13. Again, the result is not statistically significant, but the remarkably lower p-value can be taken as an indicator that this procedure is more powerful. We therefore retain the null hypothesis that the items' rank positions are in line with the assumptions of the Rasch-Model. Reconsidering that the model assigns the largest likelihood to solving the r easiest items when determining the person parameter estimates, we have no indication to reject the assumption that the $\hat{\theta}_r$ rely on fairly the same items in both sample subsets and thus no compensation as described above has occured.

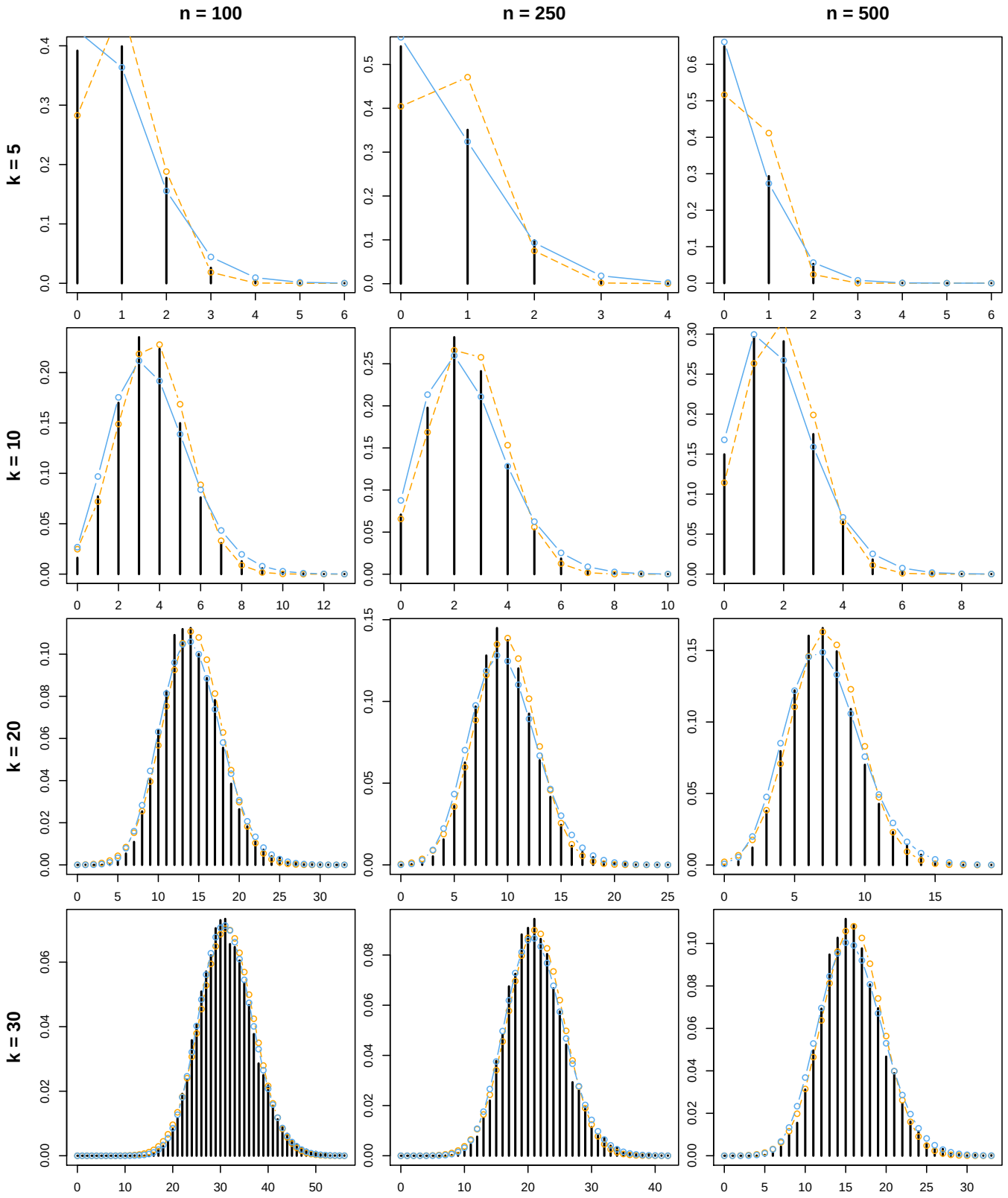

Figure 8. Distributions of the rank differences for various combinations of number of items and sample size. Orange dashed lines signify normal distributions with the same mean and standard deviation and the blue lines indicate Poisson distributions with the same mean as the simulated distributions.

Discussion

The present article takes a closer look at the genesis of the Rasch person parameters in the CML-context. First, we see that in many cases the $\hat{\theta}_r$ would not vary considerably even if the item parameters do so. Some exceptional cases have been introduced, pointing us towards possible bucklings the sequence of the $\hat{\theta}_r$ might show when item outliers exist. This is of interest when one considers to extend an existing set of items, which has been recognized to cover too small a range of the latent dimension, with a few items deliberately crafted as easy or difficult. Especially the standard errors of the person parameters will grow considerably in the vicinity of such bucklings. But in the general case (i.e. when the β_i are scattered more or less evenly across the latent dimension), we might even predict the $\hat{\theta}_r$ from the score r and the test length k without knowing the β_i.

Further considerations regarding the standard errors of the person parameter estimates have revealed that these also cover a limited range of values depending on the score r and the number of items k. Hence, the person side of the RM is to a certain extent predictable—if the model holds. But model violations may invalidate any conclusion regarding (groups of) individuals. We dispose of a number of tools to check model adequacy, among the strongest of which, the cLRT, has been considered here. However, this method focuses exclusively on the items. The present study extends this view by analyzing consequences of model violations upon the person parameter estimates. By taking into account that when estimating the $\hat{\theta}_r$ we implicitly assume that a score r has most likely arisen from positive responses to the r easiest items, a leverage has been identified, which allows for further examining the consequences of item misfit. Basically, the method checks, whether scores of split groups rely on the same items or not. For that purpose, an order criterion has been established, which allows for identifying cases, in which entirely different items are assumed to be responsible for an individual's score.

A simple graphical means, the Itemk Rank Plot, has been proposed, allowing for a rough assessment of the incidence of item position exchanges across subgroups of respondents. While the item parameter estimates' differences enter the rationale of the cLRT, their ordering is relevant for the person parameters' estimates. Hence, item position permutations also indicate model misfit as regards the person parameters. The proposed Compensation Test allows for an inferential assessment, whether item ranks differ significantly between subsamples.

The Item Rank Plot shown in Figure 7 can straightforwardly be extended to both a multi-group split and to the polytomous case (and a combination of both). For a multi-group split, one just needs to insert the respective lines. Attention should be paid then to an optimal sequence of groups in the plot, e.g. following a natural ordering of subgroups or considering an intelligible sequence of item exchanges from top to bottom. The polytomous case is equally easy to achieve by using the threshold parameters in the same way as we have done with the item parameters here—in fact, item difficulty parameters (in the sense of the RM) and threshold parameters (in the sense of a PCM) do not differ substantially in what they represent. Hence, the distribution of the rank differences of the threshold parameters of a PCM under the H_0 is likely to follow the same principles as those identified for the RM—however, this remains to be shown.

The methods presented here are based on the CML approach. One might consider this a disadvantage or, at least, a limitation. But remember that the specific objectivity (which lays the foundation of the CML approach) constitutes a distinguishing feature of the RM. It allows for the unbiased estimation of item parameters without having to determine a (possibly fallible) latent distribution of the person parameters (like in the MML approach) or venturing biased estimates (like in the JML approach).

After all, a psychological test is used for assessing individuals' characteristics regarding certain traits or abilities. While much research focuses on the items' characteristics, we should not lose sight of the person parameter estimates and their standard errors. The present contribution re-emphasizes the person-oriented perspective by illustrating some rarely discussed features regarding the person parameters on the one hand and by proposing a new fit measure and a model test on the other hand. Interestingly, already Rasch (1960) anticipated the idea of person-oriented research by citing the following statement of Joseph Zubin (1955) in the preface of his infamous and pioneering book: "*Recourse must be had to individual statistics, treating each patient as a separate universe. Unfortunately, present day statistical methods are entirely group-centered so that there is a real need for developing individual-centered statistics*" (p. VII). The present article shall contribute to this idea.

Acknowledgements

The author is indebted to Wolfgang Wiedermann for helpful comments on an earlier version of this article.

References

Alexandrowicz, R. W. (2012). GANZ RASCH: A free software for categorical data analysis. *Social Science Computer Review*, *30*, 369–379.

Alexandrowicz, R. W., & Draxler, C. (2016). Testing the Rasch model with the conditional likelihood ratio test: sample size requirements and bootstrap algorithms. *Journal of Statistical Distributions and Applications*, *3:2*, 1–25.

Alexandrowicz, R. W., Fritzsche, S., & Keller, F. (2014). Die Anwendbarkeit des BDI-II in klinischen und nichtklinischen Populationen aus psychometrischer Sicht. Eine vergleichende Analyse mit dem Rasch-Modell [Applicability of the BDI-II in clinical and not-clinial populations from a psychometric perspective: A comparative analysis with the Rasch model]. *Neuropsychiatrie*, *28*, 63–73.

Andersen, E. B. (1970). Asymptotic Properties of Conditional Maximum Likelihood Estimators. *Journal of the Royal Statistical Society, Series B*, *32*, 283–301.

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

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory. Parameter Estimation Techniques.* (2^{nd} revised and expanded ed.). NY: Marcel Dekker.

Beck, A., Steer, R., & Brown, G. (1996). Beck Depression Inventory (2nd ed.) [Computer software manual]. San Antonio: Psychological Corporation.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Draxler, C., & Alexandrowicz, R. W. (2015). Sample Size Determination within the Scope of Conditional Maximum Likelihood Estimation with Special Focus on Testing the Rasch Model. *Psychometrika*, *80*, 897–919.

Edwards, A. W. F. (1972/1992). *Likelihood (Expanded Edition)*. Cambridge: University Press.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch Models. Foundations, Recent Developments, and Applications*. NY: Springer.

Formann, A. K. (1986). A note on the computation of the second-order derivatives of the elementary symmetric functions in the Rasch Model. *Psychometrika*, 335–339.

Glas, C. A. W., & Verhelst, N. D. (1995). Testing the Rasch Model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models. Foundations, Recent Developments, and Applications* (pp. 69–95). NY: Springer.

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.

Hautzinger, M., Keller, F., & Kühner, C. (2009). *BDI-II Beck-Depressions-Inventar. Revision. 2. Auflage* (2nd ed.). Frankfurt: Pearson Assessment.

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). NY: Springer.

MacDonald, I. G. (1995). *Symmetric Functions and Hall Polynomials*. Oxford: Clarendon.

Mair, P., Hatzinger, R., & Maier, M. J. (2012). eRm: Extended Rasch Modeling [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=eRm (R package version 0.15-1)

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*, 50–60.

Molenaar, I. W. (1995). Estimation of Item Parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch Models. Foundations, Recent Developments, and Applications* (pp. 39–51). NY: Springer.

Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford: Clarendon.

Popper, K. (1959/2010). *The Logic of Scientific Discovery*. London and NY: Routledge.

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

Rasch, G. (1966a). An Individualistic Approach to Item Analysis. In P. F. Lazarsfeld & N. W. Henry (Eds.), *Readings in Mathematical Social Science* (pp. 89–107). Cambridge, MA: The M.I.T. Press.

Rasch, G. (1966b). *An informal report on the present state of a theory of objectivity in comparisons.* Retrieved 1.12.2015, from http://www.rasch.org/memo1966.pdf

Verhelst, N. D., Glas, C. A. W., & van der Sluis, A. (1984). Estimation Problems in the Rasch-Model: The Basic Symmetric Functions. *Computational Statistics Quarterly*, *1*, 245–262.

von Eye, A. (2010). Developing the person-oriented approach: Theory and methods of analysis. *Development and Psychopathology*, *22*, 277–285.

von Eye, A., Bergmann, L. R., & Hsieh, C.-A. (2015). Person-Oriented Methodological Approaches. In W. Damon & R. M. Lerner (Eds.), *Handbook of Child Psychology and Developmental Science (Volume I)* (pp. 789–841). Hoboken, NJ: Wiley.

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

Wiedermann, W., & Alexandrowicz, R. W. (2007). A plea for the Wilcoxon-Mann-Whitney-test: Further considerations on Rasch and Guiard's 'The robustness of parametric statistical methods'. *Psychological Test and Assessment Modelling (former: Psychology Science)*, *49*, 2–12.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*, 80–83.

Wright, B. D., & Douglas, G. A. (1977). Conditional versus unconditional procedures for sample-free item analysis. *Educational and Psychological Measurement*, *37*, 573–586.

Zwinderman, A., & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement*, *14*, 73–81.

A Appendix

A.1 Likelihood of a response vector

Let us consider five items with difficulty parameters $\beta_i = (-2, -1, 0, 1, 2)$ and a person with $\theta_v = 0$. Consider exemplarily the response vectors $\boldsymbol{x}_v^{(1)} = (1, 1, 0, 0, 0)$, $\boldsymbol{x}_v^{(2)} = (0, 1, 1, 0, 0)$, $\boldsymbol{x}_v^{(3)} = (0, 0, 1, 1, 0)$, and $\boldsymbol{x}_v^{(4)} = (0, 0, 0, 1, 1)$. The likelihood of each response vector is given by

$$L(\theta_v, \beta_i; \boldsymbol{x}_v) = \prod_{i=1}^{5} \frac{e^{x_{vi}(\theta_v - \beta_i)}}{1 + e^{\theta_v - \beta_i}} \tag{19}$$

Evaluating equation (19) for the four vectors yields $L(\theta_v, \beta_i; \boldsymbol{x}_v^{(1)}) = 0.2073$, $L(\theta_v, \beta_i; \boldsymbol{x}_v^{(2)}) = 0.0281$, $L(\theta_v, \beta_i; \boldsymbol{x}_v^{(3)}) = 0.0038$, and $L(\theta_v, \beta_i; \boldsymbol{x}_v^{(4)}) = 0.0005$. The vector with positive responses to the two easiest items, $\boldsymbol{x}_v^{(1)}$, attains the largest likelihood. This would also apply if we considered the remaining 16 possible vectors resulting in a score of 2.

A.2 Drawing the Item Rank Plot with R

Listing 1.1: R script for the Item Rank Plot

```
1
2  # --- function definition:
3
4    itemrankplot = function(b1,b2) {
5       b = c(b1,b2)
6       k = length(b1)
7       mi = floor(min(b))
8       ma = ceiling(max(b))
9       par(mar=c(2,0,0,0)+0.1)
10      plot(0:1,0:1,type="n",xlim=c(mi,ma),ylim=c(2.2,0.8),axes=F,xlab="",ylab="")
11      abline(h=c(1,2),col=grey(.6),lty=5)
12      axis(1)
13      points(cbind(b1,1),pch=16,cex=2)
14      points(cbind(b2,2),pch=16,cex=2)
15      text(b1,1,adj=c(0.5,-1),cex=0.8)
16      text(b2,2,adj=c(0.5, 2),cex=0.8)
17      text(mi,1,expression(beta^(1)),adj=c(0.5,-0.5))
18      text(mi,2,expression(beta^(2)),adj=c(0.5,-0.5))
19      abline(h=c(1,2),col=grey(.5),lty=5)
20      segments(b1,1,b2,2)
21      segments(sort(b1),1,sort(b2),2,col="red",lty=3)
22   }
23
24 # --- example call:
25
26   set.seed(123)  # reproduce values of Figure 7
27   k = 8
28   beta1 = runif(k,-3,3)
29   beta2 = beta1 + runif(k,-1,1)
30   itemrankplot(beta1,beta2)
```

A.3 Item Rank Plot for the Worked Example Data

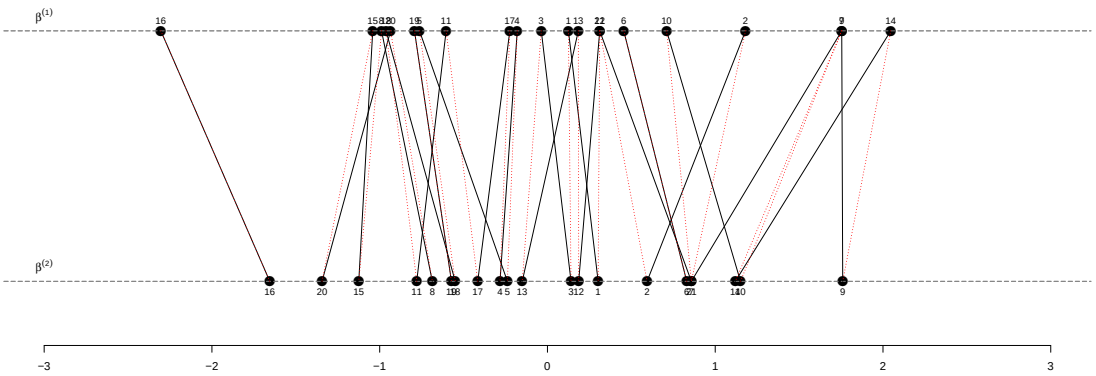

Figure 9. Item plot for the worked example data.