Ability Measures Based on Response Probabilities Representing Knowledge

G. Rohwer (July 2015)

Contents

1 Introduction

I refer to a competence test, T_m, consisting of m binary items. The items are represented by variables, $X_1, \ldots, X_m$, having values 1 (if there is a correct answer) or 0 (otherwise). Values of these variables for the members of a population (or sample) $\mathcal{P}$ are given by vectors $x_i = (x_{i1}, \ldots, x_{im})$, the sum score is denoted $s_i := \sum_j x_{ij}$; i identifies members of $\mathcal{P}$.

A standard approach to the estimation of individual abilities w.r.t. T_m uses the Rasch model. This model postulates item parameters $\delta = (\delta_1, \ldots, \delta_m)$, and for each person i a parameter θ_i, which together determine probabilities

$$\pi_{ij}^R := \Pr(X_j = 1 \,|\, \theta_i, \delta_j) := L(\theta_i - \delta_j) \tag{1}$$

where $L(x) := \exp(x)/(1 + \exp(x))$, for person i's correctly answering to item j. A problem with this approach concerns the interpretation of these probabilities. How to understand, for example, that a person can correctly solve a mathematical task with a probability 0.2, or 0.4, or 0.6?

In this paper I consider an alternative approach which defines response probabilities π_{ij} by a reference to a distinction between ‘knowing’ and ‘not knowing’ (and possibly guessing) the correct answer to an item. By introducing interval-valued response probabilities, this approach also allows one to express the idea that a person's ability to correctly solving items is, to some degree, a vague notion.

In Section 2 I introduce the approach for tests containing items which cannot be solved by guessing. In Section 3 I discuss multiple-choice (MC) items, and in Section 4 I compare the approach with the Rasch model.

2 Ability measures representing ‘knowledge’

2.1 Abilities and probabilities. I start from assuming that the interest concerns the degree to which a person can correctly solve the items of a test. The following approach is based on presupposing fixed values

$$c_{ij} = \begin{cases} 1 & \text{if person } i \text{ is able to solve item } j \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

However, the ordinary meaning of ‘a person is able to solve an item’ is to some extent vague and does not entail that the person can do so regardless of the circumstances. I therefore translate $c_{ij} = 1$ into ‘person i can solve item j with a high probability’, formally:

$$\pi_{ij} = \Pr(X_j = 1 \,|\, c_{ij} = 1) \geq 1 - \alpha \tag{3}$$

where α specifies the degree of indeterminacy, e.g., $\alpha = 0.1$.

On the other hand, the specification of a person's not being able to correctly solve an item depends on the kind of item. If the item cannot be solved by guessing, then

$$\pi_{ij} = \Pr(X_j = 1 \,|\, c_{ij} = 0) = 0 \tag{4}$$

If it is a multiple-choice (MC) item with a_j alternatives, different assumptions about how a person can guess a correct answer are possible. I distinguish between random guessing and informed guessing. If random guessing, then

$$\pi_{ij} = \Pr(X_j = 1 \,|\, c_{ij} = 0) = 1/a_j \tag{5}$$

Informed guessing is understood as random guessing based on a reduced number of alternatives (due to previously setting apart alternatives believed to be wrong). The guessing probability can then be assumed to be in the interval

$$\pi_{ij} = \Pr(X_j = 1 \,|\, c_{ij} = 0) \in [\,1/a_j, 1/2\,] \tag{6}$$

To allow for a uniform notation,[1] I use $\pi_{ij} \in I_{ij}$ with

$$I_{ij} := [\,1{-}\alpha, 1\,] \tag{7}$$

if person i is able to solve item j; and otherwise

$$I_{ij} := \begin{cases} [\,0, 0\,] & \text{if a correct answer cannot be guessed} \\ [\,1/a_j, 1/a_j\,] & \text{MC item, random guessing} \\ [\,1/a_j, 1/2\,] & \text{MC item, informed guessing} \end{cases} \tag{8}$$

[1] $[\,a, b\,]$ where $a \leq b$ is used to denote the interval of all real numbers between a and b including the endpoints.

2.2 A summary measure of abilities. In order to characterize a person's abilities w.r.t. the test T_m, one can use

$$s_i^c := \sum\nolimits_j c_{ij} \tag{9}$$

that is, the number of items the person can correctly solve. The task then is to estimate s_i^c from an observation s_i. This requires to consider s_i^c as the value of a random variable, say S_i^c, allowing one to use Bayes' rule:

$$\Pr(S_i^c{=}s^c \,|\, S_i{=}s_i) = \frac{\Pr(S_i{=}s_i \,|\, S_i^c{=}s^c)\Pr(S_i^c{=}s^c)}{\sum_{s=0}^m \Pr(S_i{=}s_i \,|\, S_i^c{=}s)\Pr(S_i^c{=}s)} \tag{10}$$

for $s^c = 0, \ldots, m$. With prior probabilities $\Pr(S_i^c = s^c) = 1/(m+1)$, the expression simplifies to

$$\Pr(S_i^c \;=\; s^c \,|\, S_i{=}s_i) = \frac{\Pr(S_i{=}s_i \,|\, S_i^c \;=\; s^c)}{\sum_{s=0}^m \Pr(S_i{=}s_i \,|\, S_i^c \;=\; s)} \tag{11}$$

Consideration of MC items will be postponed to Section 3. Here I consider constructed response items which cannot be solved by guessing. Note that this entails $s_i \leq s^c$. Let $p \in [\,1{-}\alpha, 1\,]$. Then

$$\Pr(S_i = s \,|\, S_i^c = s^c; p) = \binom{s^c}{s} p^s \, (1{-}p)^{s^c - s} \tag{12}$$

and

$$\Pr(S_i^c \;=\; s^c \,|\, S_i{=}s_i; p) = \frac{\binom{s^c}{s_i} p^{s_i} \, (1{-}p)^{s^c - s_i}}{\sum\limits_{s=s_i}^{m} \binom{s}{s_i} p^{s_i} \, (1{-}p)^{s - s_i}} \tag{13}$$

The expectation is

$$\mathrm{E}(S_i^c \,|\, S_i{=}s_i; p) = \frac{\sum\limits_{s=s_i}^{m} s \binom{s}{s_i} p^{s_i} \, (1{-}p)^{s - s_i}}{\sum\limits_{s=s_i}^{m} \binom{s}{s_i} p^{s_i} \, (1{-}p)^{s - s_i}} \tag{14}$$

For $p \in [\,1-\alpha, 1\,]$, this is a monotonically decreasing function of p, and therefore[2]

$$\mathrm{E}(S_i^c \mid S_i = s_i) = \bigcup_{p \in [1-\alpha,1]} \mathrm{E}(S_i^c \mid S_i = s_i; p) = \left[\, s_i \,,\; \frac{\sum_{s=s_i}^{m} s \binom{s}{s_i} (1-\alpha)^{s_i}\, \alpha^{s-s_i}}{\sum_{s=s_i}^{m} \binom{s}{s_i} (1-\alpha)^{s_i}\, \alpha^{s-s_i}} \,\right] \tag{15}$$

For example, if $m = 11$ and $\alpha = 0.1$, one finds the following intervals:

s_i	$\mathrm{E}(S_i^c \mid S_i = s_i)$	s_i	$\mathrm{E}(S_i^c \mid S_i = s_i)$
0	[0.0, 0.11]	6	[6.0, 6.77]
1	[1.0, 1.22]	7	[7.0, 7.87]
2	[2.0, 2.33]	8	[8.0, 8.91]
3	[3.0, 3.44]	9	[9.0, 9.82]
4	[4.0, 4.56]	10	[10.0, 10.52]
5	[5.0, 5.67]	11	[11.0, 11.00]

(16)

These intervals reflect both an estimation error and the indeterminacy due to the definition of the abilities to solve items.

2.3 Illustration with artificial data. To illustrate, I consider a test with $m = 11$ non-MC items. The number of persons who can solve s^c items is set to $100\,(|\,6.5 - |\,5.5 - s^c\,|\,|)$, $s^c = 1, \ldots, 10$, altogether $n = 4000$ cases. Then, given s_i^c, it is assumed that $c_{ij} = 1$ for $j \leq s_i^c$ and $c_{ij} = 0$ otherwise. Data are generated as follows: If $c_{ij} = 1$, π_{ij} is drawn from a uniform distribution in the interval $[\,1-\alpha,\, 1\,]$; if $c_{ij} = 0$, π_{ij} is set to 0. Then, using random numbers r_{ij} which are uniformly distributed in $[\,0, 1\,]$, $x_{ij} = 1$ if $r_{ij} \leq \pi_{ij}$, and $x_{ij} = 0$ otherwise. Finally, one can calculate $s_i = \sum_j x_{ij}$ and the intervals defined in (15).

In Figure 1, the solid line shows the cumulative distribution function (CDF) of the presupposed values s_i^c. Also shown is the interval-valued CDF of the estimated expectations.[3]

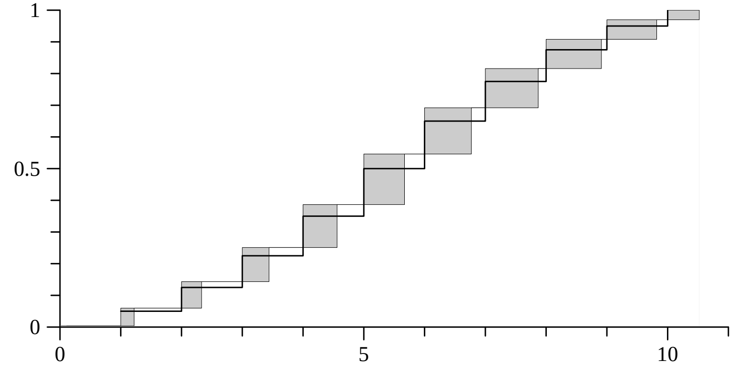


Fig. 1 CDF of the values s_i^c (solid line) and an interval-valued CDF of the estimated expectations based on 11 non-MC items.

Since the expectations $\mathrm{E}(S_i^c \mid S_i = s_i)$ are intended to provide information about the theoretically postulated values s_i^c, one can consider measurement errors

$$\max \{ |s_i^c - \mathrm{E}(S_i^c \mid S_i = s_i)| \} \tag{17}$$

that is, the maximal difference between s_i^c and a value in the estimated interval. For example, if a person has $s_i^c = 5$ and the observed sum score is $s_i = 4$, the estimated expectation is $[\,4.0, 4.56\,]$ and the measurement error is 1. A CDF of these measurement errors is shown in Figure 2. Note that one can make use of the condition $s_i \leq s_i^c$ when interpreting the measurement errors.

2.4 Expectations of sum scores. As an alternative, or complement, to the measure s_i^c, one can refer to mean values (expectations) of sum scores in hypothetical replications of the test. Two somewhat different understandings of such replications are possible.

(a) In the first understanding there are fixed probabilities $\pi_{ij} \in I_{ij}$ which do not change over the replications. Then, for each possible choice of these

[2] The expression (12) will be taken to be 1 if $p = 1$ and $s = s^c$.

[3] Given intervals $[\,a_i, b_i\,]$ for $i = 1, \ldots, n$, the interval-valued CDF is defined as the function $F(x) = [\,a_x, b_x\,]$ where $a_x := \sum_i I[\,b_i \leq x\,]/n$ and $b_x := \sum_i I[\,a_i \leq x\,]/n$. $I[\ldots]$ denotes the indicator function.

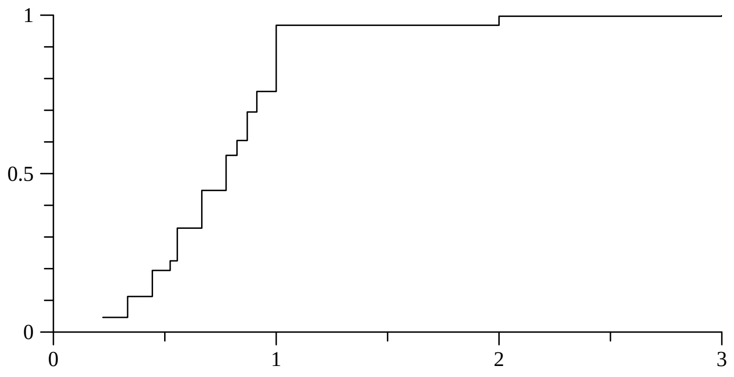


Fig. 2 CDF of the measurement errors defined in (17).

probabilities, there is a well-defined random variable, S_i, representing the sum scores, and this variable has a generalized binomial distribution

$$\Pr(S_i = s) = \sum_{x \in D_s} \prod_{j=1}^{m} \pi_{ij}^{x_j} (1 - \pi_{ij})^{1-x_j} \tag{18}$$

where D_s denotes the set of response patterns $x = (x_1, \ldots, x_m)$ with $\sum_j x_j = s$. So one can refer to the expectation of S_i,

$$\mathrm{E}(S_i) = \sum_{j=1}^{m} \pi_{ij} \tag{19}$$

and use this as a summary measure of a person's ability for solving the items of the test T_m.

(b) In the second understanding, the probabilities π_{ij} can change over replications. While one cannot immediately refer to expectations, one can well think that the interest concerns the mean value of the sum scores in a large number, say K, of hypothetical replications of the test. Let

$$S_i^{(k)} := \sum\nolimits_j X_{ij}^{(k)} \tag{20}$$

denote person i's sum score in the kth replication. For each item j, $X_{ij}^{(k)}$ has a binomial distribution with $\pi_{ij}^{(k)} \in I_{ij}$. So one can assume that the number of correct responses to item j is contained in the interval K times I_{ij} with a very high probability. Consequently,

$$\mathrm{E}_K(S_i) := \frac{1}{K} \sum\nolimits_k S_i^{(k)} = \sum\nolimits_j \frac{1}{K} \sum\nolimits_k X_{ij}^{(k)} \in_a I_i := \sum\nolimits_j I_{ij} \tag{21}$$

where $\in_a$ means 'contained in (with a very high probability, depending on K)'. For example, if $K = 1000$, the probability is almost zero that $\mathrm{E}_K(S_i)$ is not in the interval I_i.[4]

In order to represent the idea that a person's ability is to some extent a vague notion, the second understanding seems preferable. In any case, if a test only contains non-MC items, the postulated interval for the expectation of sum scores is $s_i^c\,[\,1{-}\alpha, 1\,]$. This suggests to use the estimates

$$\mathrm{E}(S_i^c \mid S_i = s_i)\,[\,1{-}\alpha, 1\,] \tag{22}$$

For example, if $m = 11$ and the observed sum score is $s_i = 4$, the estimated expectation of S_i^c is $[\,4.0, 4.56\,]$ and, with $\alpha = 0.1$, the estimated interval for the expectation of the sum score is

$$[\,4.0, 4.56\,]\,[\,0.9, 1\,] = [\,3.6, 4.56\,] \tag{23}$$

3 Tests with multiple-choice items

In this section I consider the approach introduced in the previous section for tests consisting of MC items.

3.1 Estimation of the summary measure. I first consider random guessing. To ease the application of the estimation approach described in

[4] Consider the following experiment where $m = 20$, $\alpha = 0.1$ and $K = 1000$. Let s_i^c be given. If $c_{ij} = 1$, randomly draw $\pi_{ij}^{(k)}$ from $[\,1{-}\alpha, 1\,]$ based on a uniform distribution. Then draw a random number $r_{ij}^{(k)}$ uniformly distributed in $[\,0, 1\,]$ and set $x_{ij}^{(k)} = 1$ if $r_{ij}^{(k)} \leq \pi_{ij}^{(k)}$, and zero otherwise. Set $s_i^{(k)} = \sum_j x_{ij}^{(k)}$ and check whether

$$\sum\nolimits_k s_i^{(k)}/K \in I_i$$

Even when replicating this experiment 1000 times for each possible value of s_i^c, I did not find one case where this condition was not true.

Subsection 2.2, I assume that the guessing probability is the same for all MC items, say $\gamma = 1/a_j$. The condition $s_i \leq s_i^c$ no longer holds. However, without loss of generality one can assume that person i is able to solve items $j = 1, \ldots, s_i^c$. Then, in order to find the probability of $S_i = s$ conditional on $S_i^c = s^c$, one can consider all response patterns with a sum score $s = s_a + s_b$ where s_a (that is, the number of items the person can solve) is contained in the set

$$D_{s,s^c} := \{ s_a \mid \max\{0, s^c + s - m\} \leq s_a \leq \min\{s, s^c\} \} \tag{24}$$

Let $p \in [\,1 - \alpha, 1\,]$. Then

$$\Pr(S_i = s \mid S_i^c = s^c; p) = \sum_{s_a \in D_{s,s^c}} \binom{s^c}{s_a} p^{s_a} (1-p)^{s^c - s_a} \binom{m - s^c}{s - s_a} \gamma^{s - s_a} (1-\gamma)^{m - s^c - s + s_a} \tag{25}$$

So one can apply formula (11) and derive the expectation

$$\mathrm{E}(S_i^c \mid S_i = s_i; p) = \frac{\sum_{s=0}^{m} s \Pr(S_i \; = \; s_i \mid S_i^c = s; p)}{\sum_{s=0}^{m} \Pr(S_i \; = \; s_i \mid S_i^c = s; p)} \tag{26}$$

For $p \in [\,1 - \alpha, 1\,]$, this is a monotone function of p, and one can find interval-valued expectations

$$\mathrm{E}(S_i^c \mid S_i = s_i) = \bigcup_{p \in [1-\alpha, 1]} \mathrm{E}(S_i^c \mid S_i = s_i; p) \tag{27}$$

For example, if $m = 11$, $\gamma = 0.25$ and $\alpha = 0.1$, one finds the following intervals:

s_i	$\mathrm{E}(S_i^c \mid S_i = s_i)$	s_i	$\mathrm{E}(S_i^c \mid S_i = s_i)$
0	[0.00, 0.15]	6	[4.08, 4.86]
1	[0.27, 0.46]	7	[5.35, 6.30]
2	[0.65, 0.90]	8	[6.67, 7.73]
3	[1.19, 1.53]	9	[8.00, 8.99]
4	[1.95, 2.40]	10	[9.33, 9.96]
5	[2.92, 3.52]	11	[10.62, 10.67]

(28)

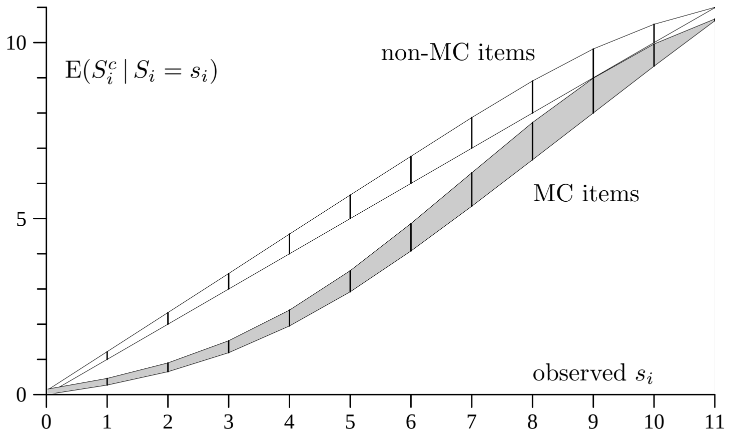


Fig. 3 Illustration of the intervals shown in (16) for non-MC items and in (28) for MC items.

As illustrated in Figure 3, these intervals are quite different from the estimates based on non-MC items in (16).

3.2 Illustration with artificial data. To illustrate, I create artificial data as described in Subsection 2.3; the only difference is that the 11 items are know assumed to have an MC format with four alternatives ($\gamma = 0.25$).

As shown in Figure 4, the presupposed distribution of the s_i^c values (the same as in Figure 1) can well be estimated. However, the measurement errors, as defined in (17), are much greater when using MC items. This is illustrated in Figure 5.

3.3 Expectations of sum scores. As discussed in Subsection 2.4, one can also consider intervals for the expectation of sum scores, now defined by

$$s_i^c\,[\,1 - \alpha, 1\,] + (m - s_i^c)\,[\,\gamma, \gamma\,] \tag{29}$$

Obviously, expectations of sum scores no longer reflect the ability measured by the number of items a person is able to solve (s_i^c).

Corresponding to (29), in order to estimate an interval for expectations

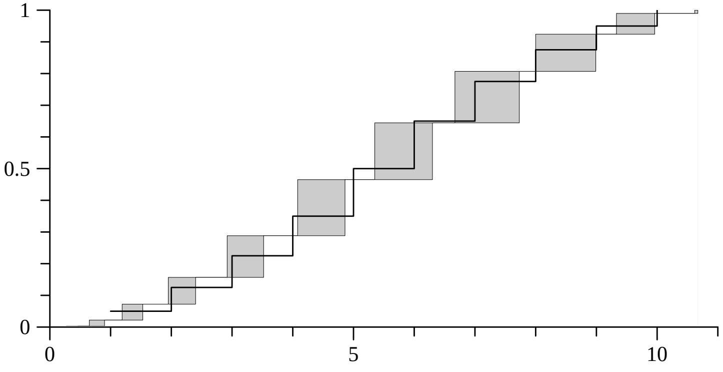


Fig. 4 CDF of the values s_i^c (solid line) and an interval-valued CDF of the estimated expectations based on 11 MC items.

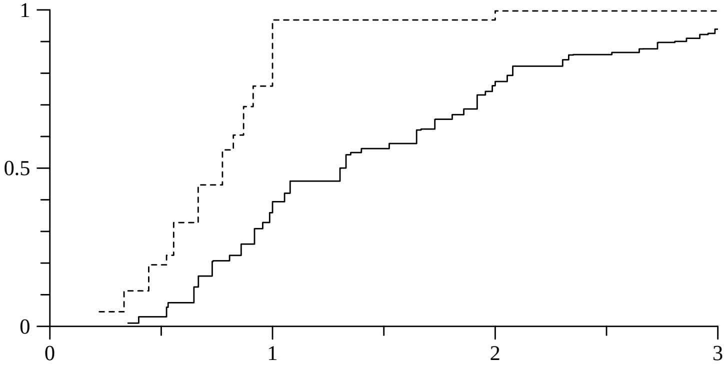


Fig. 5 CDF of the measurement errors defined in (17) for MC items (solid) and non-MC items (dashed, already shown in Figure 2).

of sum scores, one can use

$$\mathrm{E}(S_i^c \mid S_i = s_i)\,[\,1-\alpha, 1\,] + \gamma\,([\,m, m\,] - \mathrm{E}(S_i^c \mid S_i = s_i)\,) \tag{30}$$

For example, if a person has $s_i^c = 5$ and the observed sum score is $s_i = 6$, the estimated expectation is $[\,4.08, 4.86\,]$, and with $\alpha = 0.1$, the estimated interval for the expectation of the sum score is

$$[\,4.08, 4.86\,]\,[\,0.9, 1\,] + 0.25\,([\,11, 11\,] - [\,4.08, 4.86\,]) = [\,5.21, 6.59\,]$$

3.4 Informed guessing. I now consider informed guessing. Response probabilities can then take any value in the interval $[\,\gamma, 1/2\,]$. Instead of (25), one has to use

$$\Pr(S_i = s \mid S_i^c = s^c; p, q) = \tag{31}$$
$$\sum_{s_a \in D_{s,s^c}} \binom{s^c}{s_a} p^{s_a}\,(1-p)^{s^c - s_a} \binom{m - s^c}{s - s_a} q^{s - s_a}\,(1-q)^{m - s^c - s + s_a}$$

where $p \in [\,1-\alpha, 1\,]$ and $q \in [\,\gamma, 1/2\,]$. An interval-valued expectation is then defined by

$$\mathrm{E}(S_i^c \mid S_i = s_i) = \bigcup_{\substack{p \in [1-\alpha, 1] \\ q \in [\gamma, 1/2]}} \mathrm{E}(S_i^c \mid S_i = s_i; p, q) \tag{32}$$

For example, if $m = 11$, $\gamma = 0.25$ and $\alpha = 0.1$, one finds the following intervals:

s_i	$\mathrm{E}(S_i^c \mid S_i = s_i)$	s_i	$\mathrm{E}(S_i^c \mid S_i = s_i)$
0	[0.00, 0.25]	6	[2.21, 4.86]
1	[0.15, 0.46]	7	[3.20, 6.30]
2	[0.35, 0.90]	8	[4.52, 7.73]
3	[0.62, 1.53]	9	[6.16, 8.99]
4	[0.99, 2.40]	10	[8.03, 9.96]
5	[1.50, 3.52]	11	[9.76, 10.67]

(33)

Compared with (28), the intervals are much broader, reflecting the greater indeterminacy.

4 Comparison with the Rasch model

In this section I compare the approach introduced in the two previous sections with the Rasch model. To illustrate the discussion, I use data on math competencies from the first wave of cohort 3 (5th grade) of the NEPS.[5] The test consists of 23 binary items and one item having a partial

[5] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3–5th Grade, `doi:10.5157/NEPS:SC3:2.0.0`. From 2008 to 2013, NEPS data

credit format (see Duchhardt and Gerdes, 2012); 11 of the binary items require a short construction, 12 have a multiple-choice format. I use the following items for 5194 persons who have at least one valid answer.

j	non-MC	MC
1	Mag5q291_c	Mag5d041_c
2	Mag5q292_c	Mag5v271_c
3	Mag5q231_c	Mag5r171_c
4	Mag5q301_c	Mag5d051_c
5	Mag5q221_c	Mag5d052_c
6	Mag5q14s_c	Mag5q121_c
7	Mag5q131_c	Mag5r101_c
8	Mag5d02s_c	Mag5r201_c
9	Mag5d023_c	Mag5r251_c
10	Mag5v024_c	Mag5v071_c
11	Mag5v321_c	Mag5r191_c

(34)

All MC items have four alternatives so that the guessing probability is $\gamma = 0.25$.

One has to decide how to evaluate missing answers. While missing answers to non-MC items can sensibly be evaluated as wrong answers, this seems not appropriate for MC items which in any case could have been answered simply by guessing. I therefore use a random generator to substitute missing answers by correct answers with probability 1/4 and wrong answers with probability 3/4.[6]

were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. For a general description see Blossfeld, Roßbach and von Maurice (eds., 2011).

[6] This procedure is intended to eliminate, as far as it is possible, differences in the guessing behavior of the test takers. As remarked by Lord (1964), this procedure increases measurement errors in the sense of statistical variance. The procedure seems nevertheless appropriate when the goal is to assess abilities in the sense of knowledge, in contrast to modeling the behavior of the test takers in the given circumstances.

4.1 Item-specific response probabilities. The most relevant difference between the two approaches concerns the item-specific probabilities of correct answers, π_{ij}. The notions of these probabilities introduced in Subsection 2.1 can be interpreted as representing a person's ability to provide a correct answer to an item. In general, however, this interpretation is not sensible for the probabilities π^R_{ij} postulated by the Rasch model. ML estimation of the Rasch model entails the equation

$$\sum_{j=1}^{m} L(\hat{\theta}_i - \hat{\delta}_j) = s_i \tag{35}$$

All persons who belong to a score group $\mathcal{P}_s$ (that is, all persons with the sum score s) get the same value of $\hat{\theta}$, say $\hat{\theta}_s$.[7] Consequently, all members of a score group also get the same probabilities

$$\hat{\pi}^R_{sj} = L(\hat{\theta}_s - \hat{\delta}_j) \tag{37}$$

For the 11 non-MC items, based on item parameters estimated with a conditional maximum likelihood (CML) method, using the constraint $\sum_j \delta_j =$

[7] For $s = 0$ and $s = m$, the equation has no solution. One could use instead weighted maximum likelihood estimates (WLEs) proposed by Warm (1989). This proposal concerns the second step, after the item parameters have been calculated. For the estimation of person parameters Warm proposes to use the weighted likelihood function

$$\mathcal{L}^w := \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{\exp(\theta^w_i - \hat{\delta}_j)^{x_{ij}}}{1 + \exp(\theta^w_i - \hat{\delta}_j)}\, w(c_i)$$

where the weights are defined by

$$w(c_i) := \Big(\sum_{j=1}^{m} \frac{\exp(\theta^w_i - \hat{\delta}_j)}{(1 + \exp(\theta^w_i - \hat{\delta}_j))^2} \Big)^{1/2}$$

Maximizing this likelihood entails the equation

$$\sum_{j=1}^{m} e_{ij} - \frac{\sum_{j=1}^{m} e_{ij}\,(1 - e_{ij})\,(1 - 2\,e_{ij})}{2\,\sum_{j=1}^{m} e_{ij}\,(1 - e_{ij})} = \sum_{j=1}^{m} x_{ij} \tag{36}$$

where

$$e_{ij} := \frac{\exp(\theta^w_i - \hat{\delta}_j)}{1 + \exp(\theta^w_i - \hat{\delta}_j)}$$

WLEs $\hat{\theta}^w_i$ are found by solving (36) instead of (35).

Table 1

s	$\hat{\theta}^w_s$	$\hat{\theta}_s$	$\hat{\pi}^R_{s,1}$	$\hat{\pi}^R_{s,2}$	$\hat{\pi}^R_{s,3}$	$\hat{\pi}^R_{s,4}$	$\hat{\pi}^R_{s,5}$	$\hat{\pi}^R_{s,6}$	$\hat{\pi}^R_{s,7}$	$\hat{\pi}^R_{s,8}$	$\hat{\pi}^R_{s,9}$	$\hat{\pi}^R_{s,10}$	$\hat{\pi}^R_{s,11}$
0	-3.57												
1	-2.31	-2.65	0.11	0.09	0.02	0.03	0.23	0.07	0.15	0.20	0.06	0.04	0.01
2	-1.63	-1.78	0.23	0.19	0.05	0.06	0.41	0.14	0.29	0.37	0.13	0.08	0.03
3	-1.10	-1.18	0.35	0.30	0.09	0.10	0.56	0.24	0.42	0.52	0.22	0.14	0.06
4	-0.64	-0.68	0.47	0.41	0.14	0.16	0.68	0.34	0.55	0.64	0.31	0.21	0.10
5	-0.22	-0.23	0.59	0.53	0.20	0.23	0.77	0.44	0.66	0.74	0.42	0.29	0.14
6	0.20	0.22	0.69	0.64	0.28	0.31	0.84	0.55	0.75	0.81	0.53	0.39	0.21
7	0.63	0.67	0.78	0.73	0.38	0.42	0.89	0.66	0.83	0.87	0.64	0.51	0.29
8	1.09	1.18	0.85	0.82	0.51	0.54	0.93	0.76	0.89	0.92	0.74	0.63	0.40
9	1.63	1.78	0.91	0.89	0.65	0.69	0.96	0.86	0.93	0.95	0.84	0.76	0.55
10	2.32	2.66	0.96	0.95	0.82	0.84	0.98	0.93	0.97	0.98	0.93	0.88	0.75
11	3.59												
$\hat{\delta}_j$			-0.58	-0.34	1.15	1.00	-1.42	0.00	-0.88	-1.26	0.11	0.65	1.57

0, these probabilities are shown in the columns labelled $j = 1, \dots, 11$ of Table 1. The table also shows the estimated item parameters, $\hat{\delta}_j$, and the estimates $\hat{\theta}_s$ and $\hat{\theta}^w_s$.

How are these probabilities to be interpreted? There are two reasons why $\hat{\pi}^R_{ij}$ should not be interpreted as person i's probability of a correct answer to item j. First, these estimates only depend on the person's sum score, s_i, and do not take into account how the person has answered to item j. Second, in most cases it would be difficult to understand the value of $\hat{\pi}^R_{ij}$ as representing a person's probability of solving a particular item. Consider, for example, a person in the score group $s = 4$. For item $j = 7$, the estimated probability is 0.55, but what could it mean that this person can solve item 7 with this probability?

The fact that the probabilities $\hat{\pi}^R_{sj}$ are identical for all members of the score group $\mathcal{P}_s$ suggests to interpret these estimates as reflecting the mean behavior of the members of the group; formally:

$$\text{for all } i \in \mathcal{P}_s: \; \hat{\pi}^R_{ij} = \hat{\pi}^R_{sj} = \frac{1}{n_s} \sum_{i \in \mathcal{P}_s} \hat{\pi}^R_{ij} \approx q_{sj} := \frac{1}{n_s} \sum_{i \in \mathcal{P}_s} x_{ij} \qquad (38)$$

where n_s is the number of persons in $\mathcal{P}_s$. This suggests to interpret $\hat{\pi}^R_{ij}$ as an estimate of the proportion of persons in the score group to which i belongs who can correctly solve item j; and this is conceptually different from person i's probability of correctly solving that item.

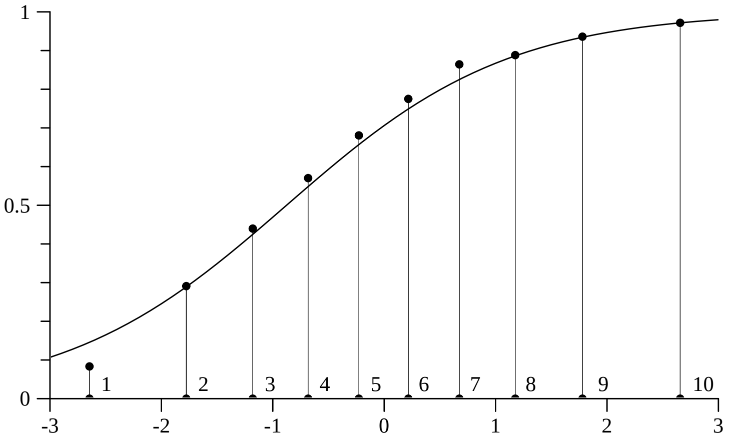


Fig. 6 Comparison of $\hat{\pi}^R_{sj}$ and q_{sj} for $j = 7$ and $s = 1, \dots, 10$.

The idea that (38) should be approximately valid is often considered as a starting point for goodness-of-fit tests of the Rasch model. For example, one can graphically compare $\hat{\pi}^R_{sj}$ and q_{sj}. This is illustrated in Figure 6 for $j = 7$ and $s = 1, \dots, 10$. However, such considerations do not relate to the response probabilities of individual persons which cannot be investigated with data from only a single test.

4.2 Processes generating responses. The approach introduced in Section 2 starts from presupposing parameters, c_{ij}, representing item-specific abilities. In contrast, the Rasch model presupposes parameters representing the difficulty of items and a single further parameter for each person. This allows defining the response probabilities π^R_{ij} which describe the supposed process generating responses.

However, as I have argued in the previous subsection, estimates of π^R_{ij} should not be interpreted as describing how individual persons generate responses. In this subsection, I use an example in order to show that the fit of a Rasch model can be compatible with the approach introduced in Section 2 and therefore is not a sufficient argument for interpreting π_{ij}^R as the response probability of person i.

I consider an example with $m = 9$ items and $n = 4000$ persons. In order to create a distribution of c_{ij} values, I start from values $d_j = 0.1\,j$. For each person i and item j, $c_{ij} = 1$ if $r_{ij} \geq d_j$, where r_{ij} is a random number uniformly distributed in $[\,0,1\,]$. The values c_{ij} are then used to calculate item difficulties $d_j^* := \frac{1}{n}\sum_{i=1}^{n} I[c_{ij} = 0]$ which represent the abilities of the persons in the population considered. Using values transformed to a logit scale, $\delta_j^* := \log(d_j^*/(1-d_j^*))$, one can consider the function $h^*(\theta) := \sum_j L(\theta - \delta_j^*)$. Then, for each value $s^c = 0, 1, \ldots, m$, one can define a corresponding value $\theta^*(s^c) := h^{*-1}(s^c)$. The following table shows the result.

j	d_j^*	δ_j^*	s^c	$\theta^*(s^c)$
1	0.11	−2.13	1	−2.64
2	0.20	−1.36	2	−1.65
3	0.29	−0.89	3	−0.92
4	0.40	−0.39	4	−0.29
5	0.50	0.01	5	0.32
6	0.61	0.44	6	0.95
7	0.71	0.90	7	1.68
8	0.79	1.35	8	2.68
9	0.90	2.21		

(39)

Based on these values, Figure 7 shows the ICCs (item characteristic curves) $L(\theta - \delta_j^*)$, and Figure 8 shows the TCC (test characteristic curve) $h^*(\theta)$.

I now use the values c_{ij} to generate responses x_{ij} in the following way ($\alpha = 0.1$): If $c_{ij} = 1$, π_{ij} is drawn from a uniform distribution in the interval $[\,1{-}\alpha,\,1\,]$; otherwise, if $c_{ij} = 0$, π_{ij} is set to 0. Then, using random numbers r_{ij} which are uniformly distributed in $[\,0,1\,]$, $x_{ij} = 1$ if $r_{ij} \leq \pi_{ij}$, and $x_{ij} = 0$ otherwise. Finally, the data x_{ij} are used to fit a Rasch model (CML estimation).

Figure 9 shows the estimated ICCs and proportions of correct answers in the score groups. Obviously, the Rasch model provides a good fit,

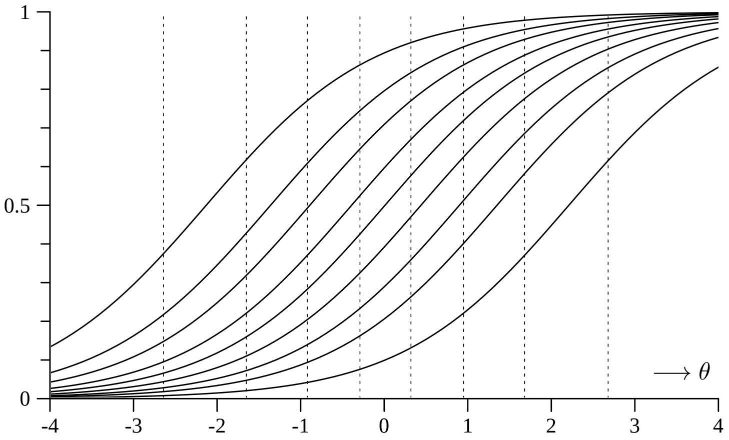


Fig. 7 The functions $L(\theta - \delta_j^*)$ and values of $\theta^*(s^c)$, based on Table (39).

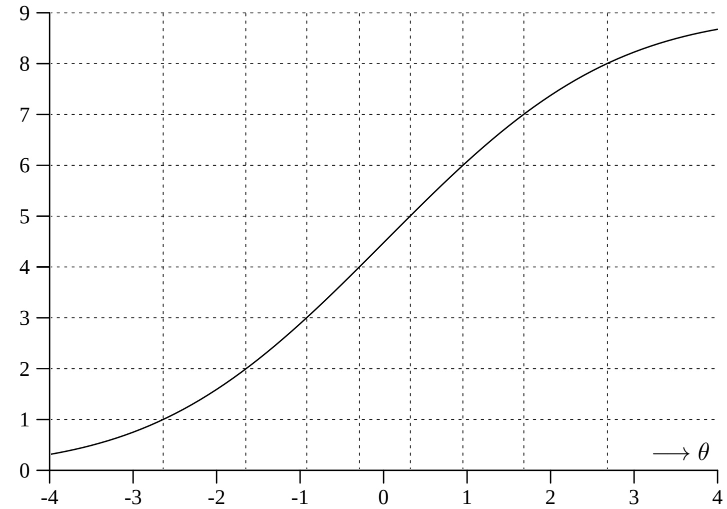


Fig. 8 The function $h^*(\theta)$ and values of $\theta^*(s^c)$, based on Table (39).

although the responses are generated according to

$$\Pr(X_j = 1 \,|\, c_{ij}) \begin{cases} \in [\,1-\alpha, 1\,] & \text{if } c_{ij} = 1 \\ = 0 & \text{if } c_{ij} = 0 \end{cases} \tag{40}$$

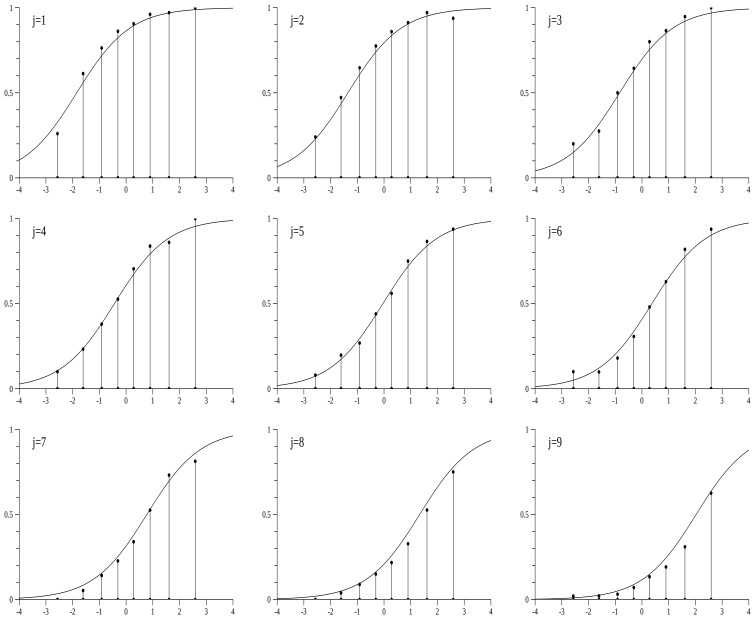


Fig. 9 Estimated ICCs ($\hat{\delta}_j$) and observed proportions $\mathrm{P}(X_j = 1 \mid S = s)$ for $s = 1, \ldots, 8$ and $j = 1, \ldots, 9$.

One can conclude that the fit of a Rasch model is not a sufficient basis for arguments about the processes which lead from persons' abilities to their responses to the items of a test (also see García-Pérez, 1999).

4.3 Variability of ability estimates. What can be said about the variability of ability estimates across successive tests? With data from only a single test an evidence-based answer is not possible. Instead, one has to refer to hypothetical replications which presuppose a particular model, including the assumptions about the processes generating the individual responses (see Subsection 4.2) which cannot be tested with data from a single test.

I begin with the Rasch model. Person i's ability as postulated by the Rasch model is θ_i. Given item parameters δ_j (which are defined by the reference to a particular population), postulating a value θ_i is equivalent with postulating probabilities π_{ij}^R as defined in (1). So one can consider a variable S_i^R, representing sum scores in hypothetical replications, whose distribution is defined by (18) where π_{ij} is substituted by π_{ij}^R. The expectation is

$$\mathrm{E}(S_i^R) = \sum_{j=1}^{m} \pi_{ij}^R = h(\theta_i) \tag{41}$$

where

$$h(\theta) := \sum_{j=1}^{m} L(\theta - \delta_j) \tag{42}$$

is the test characteristic curve (TCC). This shows that, if the item parameters are viewed as fixed by the reference to a population, θ_i and $\mathrm{E}(S_i^R) = h(\theta_i)$ are equivalent representations of a person's competence. In the following, I first consider estimates of $\mathrm{E}(S_i^R)$; variances of estimates of θ_i will be considered in the next Subsection.

As shown by (35), ML estimates of $\mathrm{E}(S_i^R)$ are given by s_i. Thus, the variance of these estimates is simply $\mathrm{V}(S_i^R)$. An estimate of this variance can be derived from estimated response probabilities:

$$\hat{\mathrm{V}}(S_i^R) = \sum_{j=1}^{m} \hat{\pi}_{ij}^R (1 - \hat{\pi}_{ij}^R) \tag{43}$$

To illustrate, I use the 11 non-MC items of the math test. The dashed line in Figure 10 shows the function $s_i \longrightarrow \hat{\mathrm{V}}(S_i^R)$.[8] Obviously, the variability of the ability estimates is highest in the middle region of the ability spectrum.

I now consider the approach introduced in Section 2. If values of S_i^c were known, one could refer to interval-valued variances. Depending on the kind of items, the intervals are

$$\mathrm{V}(S_i \mid S_i^c = s_i^c) = s_i^c \, [0, \alpha\,(1 - \alpha)] \tag{44}$$

[8] Given item parameters resulting from CML estimation, this function is calculated by the following steps: $s_i \longrightarrow \hat{\theta}_i \longrightarrow \hat{\pi}_{ij} \longrightarrow \hat{\mathrm{V}}(S_i^R)$.

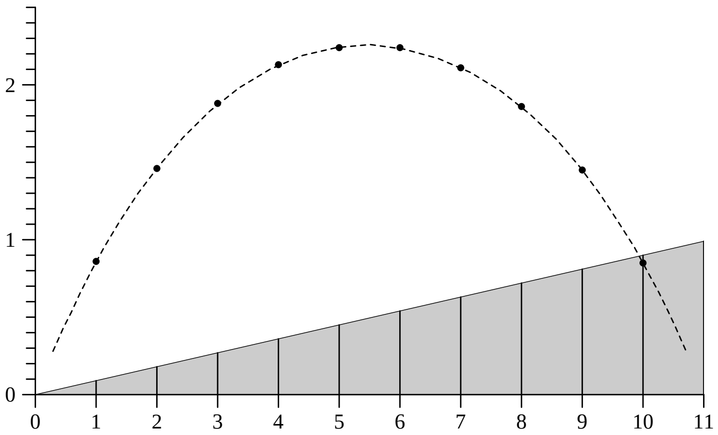


Fig. 10 Based on the 11 non-MC items of the math test, the dashed line shows the function $s_i \longrightarrow \hat{\mathrm{V}}(S_i^R)$ for $s_i = 1, \dots, 10$. The grey-shaded area shows how the intervals $\mathrm{V}(S_i \,|\, S_i^c = s_i^c)$, defined in (44), depend on $s_i^c = 0, \dots, 11$.

for non-MC items, and

$$\mathrm{V}(S_i \,|\, S_i^c = s_i^c) = s_i^c \,[\,0, \alpha\,(1-\alpha)\,] + (m - s_i^c)\,(\gamma\,(1-\gamma)) \tag{45}$$

for MC items. In Figure 10, the grey-shaded area illustrates how the intervals for non-MC items depend on $s_i^c = 0, \dots, 11$. Due to the definition of response probabilities representing knowledge, these variances are, in general, much smaller than those entailed by the Rasch model. Since the lower limits of the ability estimates $\mathrm{E}(S_i^c \,|\, S_i = s_i)$ equal s_i, also the variability of these estimates is relatively small (see also the argument about measurement errors in Subsection 2.3). On the other hand, as immediately seen by (45), MC items heavily increase the variances of sum scores, and consequently the variances of the derived ability estimates (see Subsection 3.1).

4.4 Transformation to a logit scale. As shown by (41), θ_i can be considered as a transformation of the expectation of S_i^R to a logit scale. The standard approach to assess the variability of estimates of θ_i uses the framework of ML estimation. When using CML estimation, one can

proceed in two steps. In a first step, one estimates item parameters $\hat{\delta}_j$. Then, with these parameters fixed, one can consider for each person i a likelihood

$$\mathcal{L}_i(\theta_i) = \prod_{j=1}^{m} \frac{\exp(\theta_i - \hat{\delta}_j)^{x_{ij}}}{1 + \exp(\theta_i - \hat{\delta}_j)} \tag{46}$$

It seems possible, then, to calculate an estimate of the variance of the MLE $\hat{\theta}_i$ by

$$\hat{\mathrm{V}}(\hat{\theta}_i) = \Big(-\frac{\partial^2 \log(\mathcal{L}_i(\theta_i))}{\partial\, \theta_i^2} \Big)^{-1}_{\theta_i = \hat{\theta}_i} = \frac{1}{\sum_j \hat{\pi}_{ij}^R\,(1 - \hat{\pi}_{ij}^R)} \tag{47}$$

However, the usual interpretation which requires asymptotic considerations is not applicable because the number of random variables X_{ij} is fixed by the test T_m.

As an alternative, one can make use of the fact that, given item parameters $\hat{\delta}_j$, there is a deterministic relationship $\hat{\theta}_i = \hat{h}^{-1}(s_i)$ with $\hat{h}(\theta) := \sum_j L(\theta - \hat{\delta}_j)$.[9] So one can derive an estimate of the variance of $\hat{\theta}_i$ from an estimate of the variance of S_i^R. The delta method provides a simple approximation. If $\hat{\theta}_i = g(S_i^R)$, then

$$\mathrm{V}(\hat{\theta}_i) \approx \mathrm{V}(S_i^R)\;[g'(\mathrm{E}(S_i^R))]^2 \tag{48}$$

where g' denotes the derivative of g. Using $g(s) = \hat{h}^{-1}(\theta)$, one finds

$$g'(s) = \big(\hat{h}'(\theta)\big)^{-1} = \big(\sum\nolimits_j L(\theta - \hat{\delta}_j)\,[1 - L(\theta - \hat{\delta}_j)]\,\big)^{-1} \tag{49}$$

Since $\mathrm{E}(S_i^R)$ is estimated by s_i, $g(\mathrm{E}(S_i^R)) = g(s_i) = \hat{h}^{-1}(\hat{\theta}_i)$, and therefore

$$g'(s) = \big(\sum\nolimits_j \hat{\pi}_{ij}^R\,(1 - \hat{\pi}_{ij}^R)\,\big)^{-1} \tag{50}$$

Finally, inserting this into (48) and using (43), one finds

$$\hat{\mathrm{V}}(\hat{\theta}_i) \approx \big(\sum\nolimits_j \hat{\pi}_{ij}^R\,(1 - \hat{\pi}_{ij}^R)\,\big)^{-1} \tag{51}$$

which equals the ML estimate of the variance of $\hat{\theta}_i$.

[9] As before, I refer to ML estimates of θ_i. Similar considerations are possible when referring to weighted ML estimates as proposed by Warm (1989).

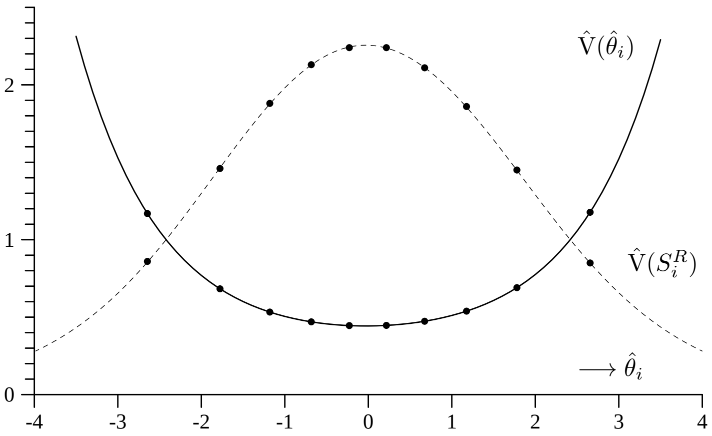


Fig. 11 Based on the 11 non-MC items of the math test, for $s_i = 1, \dots, 10$ and $\hat{\theta}_i = \hat{h}^{-1}(s_i)$, the solid line shows the estimated variance $\hat{V}(\hat{\theta}_i)$ and the dashed line shows the estimated variance of the corresponding sum score variable.

To illustrate I use again the 11 non-MC items of the math test. The solid line in Figure 11 shows how the estimated variance $\hat{V}(\hat{\theta}_i)$ depends on $\hat{\theta}_i = \hat{h}^{-1}(s_i)$, for $s_i = 1, \dots, 10$. The dashed line shows the estimated variance of the corresponding sum score variable S_i^R. Obviously, compared with $\hat{V}(S_i^R)$, the dependence of $\hat{V}(\hat{\theta}_i)$ is reversed. However, this is simply a consequence of the logit transformation which is used for defining θ_i.

4.5 The Rasch model and MC items. A measure of competence should inform about a person's ability to correctly solve items, and this is different from being able to guess correct answers. Therefore, when a test contains MC items, one needs an approach to distinguishing 'knowing' and 'guessing'. Since the approach introduced in Section 2 starts from quantities c_{ij} defined by a reference to 'knowing', the summary measure s_i^c can well be interpreted as a measure of a person's 'knowing'. As shown in Subsection 3.3, when a test contains MC items, there clearly is a difference between this measure of competence and a reference to expectations of sum scores. In contrast, the Rasch model does not distinguish between non-MC and MC items. This has several consequences.

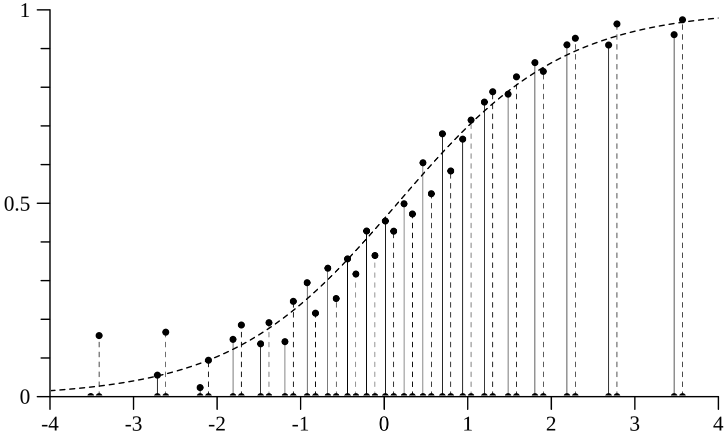


Fig. 12 An ICC centered at 0.16, and proportions of persons who have correctly answered the non-MC item $j = 9$ (solid) and the MC item $j = 11$ (dashed).

(a) Ability measures estimated with a Rasch model must be interpreted as reflecting abilities to provide correct answers irrespective of whether these answers result from 'knowing' or from 'guessing'.[10]

(b) In regions of low capabilities, item characteristic curves (ICC) will often provide only a poor fit. To illustrate, I estimate a Rasch model with all 22 items of the math test (again CML, $\sum_j \hat{\delta}_j = 0$). Estimation results indicate that the non-MC item $j = 11$ and the MC item $j = 12$ have almost the same difficulty (estimated item parameters are 0.13 and 0.19, respectively). Figure 12 shows an ICC centered at 0.16, and proportions of persons who have correctly answered to these items in the score groups $s = 1, \dots, 21$. The solid lines relate to the non-MC item, the dashed lines relate to the MC item.

(c) A further consequence concerns the interpretation of the estimated

[10] It does not follow that the Rasch model is not compatible with guessing as it is sometimes claimed (e.g., Andrich and Marais, 2014). It simply means that the assessed competence is a mixture of knowing and guessing.

Rasch probabilities $\hat{\pi}^R_{ij}$. As argued above, these relate to score groups (SG), not to individual persons. These score groups result from the actual outcomes of the test. If the test contains MC items, membership in a score group also depends on whether a person was lucky or not in guessing a correct answer (and by the random guessing that was used to substitute missing answers). So there is a further problem for the interpretation of the estimated Rasch probabilities $\hat{\pi}^R_{ij}$. These estimates not only do not characterize individual persons, they also do not relate to groups of persons defined by similar abilities.

(d) It is questionable whether non-MC and MC items fit a common Rasch model. One can check this with a test proposed by Martin-Löf (Verhelst, 2001; Bartolucci, 2007). For the math test, one has to estimate three Rasch models, one for the $m_1 = 11$ non-MC items, one for the $m_2 = 11$ MC items, and one common model for the set of $m = m_1 + m_2 = 22$ items. CML estimation provides the following log-likelihoods:

11 non-MC items	$\log(L_1) = -20648.7$
11 MC items	$\log(L_2) = -21293.6$
joint model	$\log(L_0) = -49468.1$

The test statistic is

$$-2\,\log\left(\frac{L_0\,\prod_{s=0}^{m}(n_s/n)^{n_s}}{L_1\,L_2\,\prod_{r=0}^{m_1}\prod_{s=0}^{m_2}(n_{rs}/n)^{n_{rs}}}\right) \qquad (52)$$

where n_s is the number of persons with a sum score s in the joint test, n_{rs} is the number of persons with sum score r in the first test and sum score s in the second test (shown in Table 2), and $n = 5194$ is the total number of persons. If the joint model fits both sets of items, the test statistic (52) asymptotically follows a χ^2 distribution with $m_1\,m_2 - 1$ degrees of freedom.[11] In our example, the test statistics has the value 303.1, and with 120 degrees of freedom this is highly improbable. So one should conclude that a Rasch model does not fit both sets of items.

Table 2 Number of persons with r correct non-MC items and s correct MC items.

	s = 0	1	2	3	4	5	6	7	8	9	10	11
r = 0	2	16	20	22	17	7	4	2	0	0	0	0
1	4	13	36	46	45	25	17	4	2	0	0	0
2	2	27	31	74	63	49	33	15	5	0	0	0
3	3	13	42	83	69	64	52	30	10	6	1	0
4	2	14	30	71	90	92	91	56	27	10	1	0
5	0	6	26	54	99	112	125	88	63	19	4	2
6	0	4	18	41	82	120	124	134	87	49	9	3
7	0	3	6	25	53	106	133	175	121	85	24	6
8	0	0	5	16	26	69	105	110	165	111	45	10
9	0	0	1	2	13	48	69	118	135	123	61	23
10	0	0	0	3	2	11	29	55	77	86	56	34
11	0	0	0	0	1	2	7	13	20	31	44	24

4.6 Using a modified Rasch model? Can a distinction between 'knowing' and 'guessing' also be made with the Rasch model? In the following, I consider a proposal made by Keats (1974) and White (1976), also discussed by Weitzman (1996). The proposal consists of two parts.

a) Knowing the correct answer to an item is defined as 'one can give the correct answer with probability 1'; and not knowing the correct answer is assumed to entail guessing which is defined as giving the correct answer with probability $\gamma_j := 1/a_j$ (a_j being the number of

[11] The null hypothesis is that a Rasch model fits the data. Since this model can be estimated with CML, the likelihood is

$$\prod_{i=1}^{n}\Pr(X_1 = x_{i1},\ldots,X_m = x_{im}\,|\,S = s_i)\;\prod_{i=1}^{n}\Pr(S = s_i)$$

The first product is equal to the conditional likelihood L_0; the second product can be estimated by $\prod_{s=0}^{m}(n_s/n)^{n_s}$, so there are $2\,m-1$ degrees of freedom. The likelihood for the alternative hypothesis is

$$\prod_{i=1}^{n}\Pr(X_1 = x_{i1},\ldots,X_m = x_{im}\,|\,S_1 = s_{i1},S_2 = s_{i2})\;\prod_{i=1}^{m}\Pr(S_1 = s_{i1},S_2 = s_{i2})$$

The first product is equal to the product of the conditional likelihoods $L_1\,L_2$; the second product can be estimated by $\prod_{r=0}^{m_1}\prod_{s=0}^{m_2}(n_{rs}/n)^{n_{rs}}$, so there are $m_1\,m_2 + 2\,m - 2$ degrees of freedom.

alternatives). Obviously, this equals the proposal made in Subsection 2.1 with $\alpha = 0$, and I therefore use again the variables C_{ij} to represent 'knowledge'.

b) It is assumed that one can sensibly think of the probability that a person knows the correct answer to an item, and it is proposed that this probability can be modeled by a Rasch model:

$$\Pr(C_{ij} = 1) = L(\theta_i - \delta_j) \tag{53}$$

The proposal entails:

$$\begin{aligned} &\Pr(X_j = 1 \mid \theta_i, \delta_j) = \\ &\quad \Pr(X_j = 1 \mid C_{ij} = 1) \Pr(C_{ij} = 1 \mid \theta_i, \delta_j) + \\ &\quad \Pr(X_j = 1 \mid C_{ij} = 0) \Pr(C_{ij} = 0 \mid \theta_i, \delta_j) = \\ &\quad L(\theta_i - \delta_j) + \gamma_j \, (1 - L(\theta_i - \delta_j)) = \frac{\exp(\theta_i - \delta_j) + \gamma_j}{1 + \exp(\theta_i - \delta_j)} \end{aligned} \tag{54}$$

If one further assumes local independence, one arrives at a 'generalized' Rasch model that allows one to estimate θ_i and δ_j, and consequently the probabilities $\Pr(C_{ij} = 1 \mid \theta_i, \delta_j)$.

To illustrate this approach I use the 11 MC items of the math test. For this application, the likelihood function of the model is

$$\mathcal{L}(\theta_i, \delta_j) = \prod_{i=1}^{n} \prod_{j=1}^{m} \frac{(\exp(\theta_i - \delta_j) + \gamma)^{x_{ij}} \, (1 - \gamma)^{1 - x_{ij}}}{1 + \exp(\theta_i - \delta_j)} \tag{55}$$

where $\gamma = 0.25$. Conditional ML estimation is no longer possible; I therefore use a marginal likelihood as proposed by Bock and Aitkin (1981). In parallel to (55), a marginal likelihood function can be written as

$$\mathcal{L}^*(\delta_j) = \prod_{i=1}^{n} \int_u \prod_{j=1}^{m} \frac{(\exp(u - \delta_j) + \gamma)^{x_{ij}} \, (1 - \gamma)^{1 - x_{ij}}}{1 + \exp(u - \delta_j)} \, f(u) \, \mathrm{d}\, u \tag{56}$$

where $f(u)$ is a presupposed density function of the quantities θ_i in a given population. To ease calculations, I specify $f(u)$ as a standard normal density function.

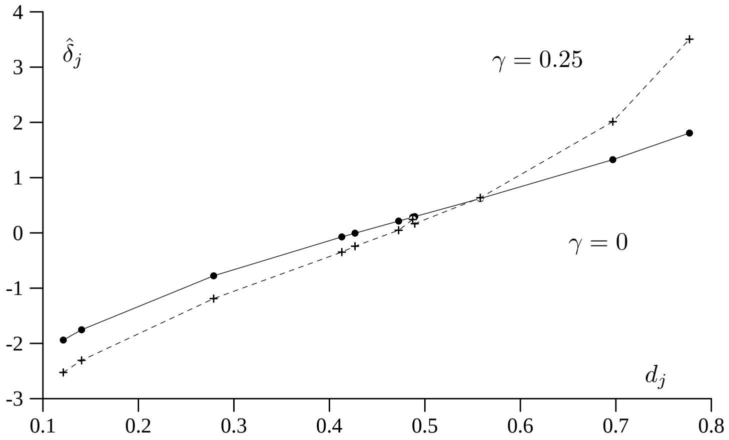


Fig. 13 Comparison of item parameters of 11 MC items of the math test and item difficulties d_j.

Item parameters are estimated with the constraint $\sum_j \hat{\delta}_j = 0$. Figure 13 shows how the relationship between item parameters and item difficulties, defined by

$$d_j := \frac{1}{n} \sum_{i=1}^{n} I[x_{ij} = 0] \tag{57}$$

has changed between a Rasch model with $\gamma = 0$ and the modified model with $\gamma = 0.25$.

In order to find estimates of the parameters θ_i, one can start from the likelihood function (55). Maximization of this function entails the first-order condition

$$s_i^* := \sum\nolimits_j x_{ij} \frac{\exp(\theta_i - \delta_j)}{\exp(\theta_i - \delta_j) + \gamma} = \sum\nolimits_j \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)} \tag{58}$$

This is similar to (35), but instead of the observed sum score s_i one now uses the score s_i^* which is downscaled by the guessing probability γ.

Employing this equation, using estimates of δ_j resulting from the MML approach (56), one can try to find solutions $\hat{\theta}_i$. However, as a consequence of the downscaling, also for persons with a sum score $0 < s < 11$ there is not always a solution. The following table shows the number of cases where the equation cannot be solved.

observed sum score	number of cases	cases with no solution	observed sum score	number of cases	cases with no solution
0	13	13	6	789	5
1	96	38	7	800	2
2	215	46	8	712	0
3	437	33	9	520	0
4	560	20	10	245	0
5	705	10	11	102	102

For 269 of the 5194 persons one cannot find a value of $\hat{\theta}_i$.[12,13]

For the remaining 4925 persons, due to the additional parameter γ, the modified model will provide a better overall fit. However, the fit of the modified ICCs will not always be better. To illustrate, I consider MC item $j = 2$. Figure 14 shows the ICC from the Rasch model with $\gamma = 0$. The vertical lines indicate the proportion of correct answers to this item for score groups $s = 1, \dots, 10$. Figure 15 shows the ICC for the same item from the modified model with $\gamma = 0.25$. There is no longer a one-to-one correspondence between score groups and values of $\hat{\theta}_i$. I therefore partitioned the abscissa into 14 intervals (cut points are: $-4.25, -3.75, \dots, 1.25, 1.75$), and then calculated the proportion of correct answers in each of these intervals.

Looking at these figures, one can well imagine models with more flexible ICCs providing a better fit.[14] However, the main reason for introducing the modified model was not to achieve a better fit, but to provide measures of competence which in a sense correct for guessing. Does the model achieve this goal? In order to interpret the parameters θ_i, one has to understand the probabilities $\Pr(C_{ij} = 1 \,|\, \theta_i, \delta_j)$. These probabilities cannot be understood as relating to individual persons. Consider the following example where $\Pr(C_{ij} = 1 \,|\, \theta = 0.41, \delta_j = 0) = 0.6$ so that $\Pr(X_j\!=\!1 \,|\, \theta\!=\!0.41, \delta_j\!=\!0)\!=\!0.70$.

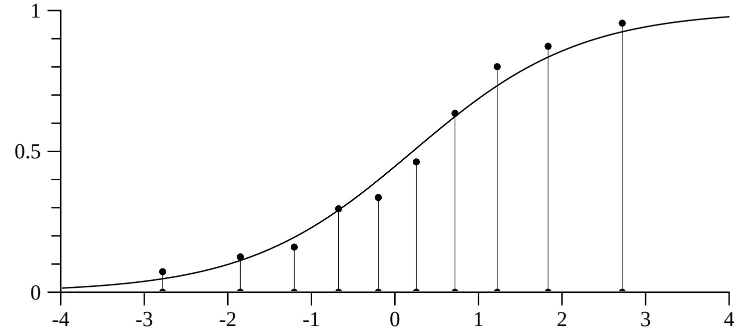


Fig. 14 ICC for MC item $j = 2$ from the model with $\gamma = 0$. The vertical lines indicate the proportion of correct answers to this item for score groups $s = 1, \dots, 10$.

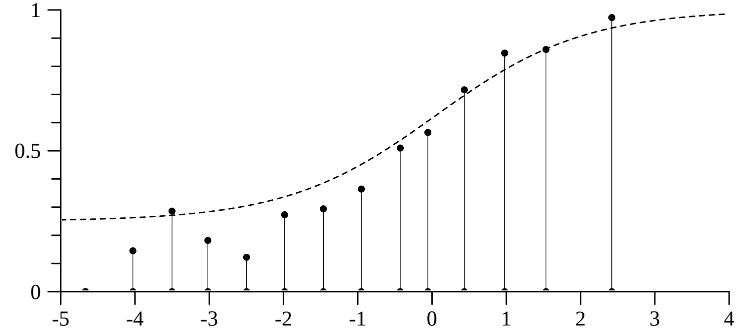
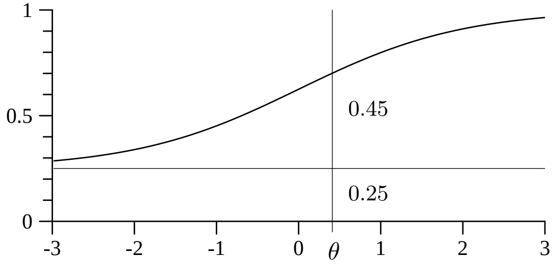


Fig. 15 ICC for MC item $j = 2$ from the model with $\gamma = 0.25$. The vertical lines indicate the proportion of correct answers to this item in 14 intervals on the abscissa.

[12] Instead of the parameters θ_i, researchers starting from estimating item parameters with a marginal likelihood function often use so-called EAP values (mean values of the distribution $f(u)$ conditional on a person's observed test results), see e.g. Kubinger and Draxler (2007). For a critical discussion see Rohwer (2015).

[13] When estimating the modified model for all 22 math items, it turns out that at least one correct answer to a non-MC item is sufficient for there being a solution of (58) for persons with a sum score $0 < s < 22$.

[14] This could be achieved with a 3PL model, or a reduced version of the 3PL proposed by Kubinger and Draxler (2007), or a generalized version proposed by Andrich et al. (2012). For a discussion of problems of estimation and interpretation of the 3PL model see Maris (2002), Han (2012), García-Pérez (1999), San Martin et al. (2015).



This cannot be true for any particular person because the model presupposes that

$$\Pr(X_j = 1) \;=\; \begin{cases} 1 & \text{if } c_{ij} = 1 \\ \gamma & \text{otherwise} \end{cases} \tag{59}$$

How to deal with this incoherence?

(a) In order to avoid the problem one can drop the assumption (53) and consider values of C_{ij} as parameters whose sum, $s_i^c = \sum_j c_{ij}$, is to be estimated. Probabilities for values of C_{ij} are then understood in an epistemic sense, expressing our only partial knowledge about a person's ability. In this way, as was done in Section 2, one can understand probabilistic statements about values of $\Sigma_j C_{ij}$ as quantifying an epistemic expectation about the number of items a person can correctly solve. This approach also allows one to avoid the implausible assumption $\alpha = 0$.

(b) A second way out of the problem tries to reconcile (53) with considering values of C_{ij} as parameters which cannot change during hypothetical replications of the test. The basic idea is to assume that the population, $\mathcal{P}$, consists of subpopulations $\mathcal{P}_\theta$ so that $\theta_i = \theta$ for all members $i \in \mathcal{P}_\theta$. One can then think of a process generating responses in the following way:

(1) In an initial step, values of C_{ij} are generated according to (53).

(2) Given values of C_{ij}, values of X_j are generated according to (59).

In this set-up, $\Pr(X_j = 1 \,|\, \theta_i, \delta_j)$ relates to the subpopulation to which person i belongs. For example, $\Pr(X_j = 1 \,|\, \theta_i, \delta_j) = 0.7$ means that approximately 70 % of the members of $\mathcal{P}_{\theta_i}$ correctly answer to item j. In the same way, one has to interpret $\Pr(C_{ij} = 1 \,|\, \theta_i, \delta_j)$ as relating to $\mathcal{P}_{\theta_i}$. For example, $\Pr(C_{ij} = 1 \,|\, \theta_i, \delta_j) = 0.6$ means that approximately 60 % of the members of $\mathcal{P}_{\theta_i}$ are able (have the knowledge) to correctly answer to item j. Consequently, the variable $\tilde{S}_i^c := \sum_j C_{ij}$ is the same for all members of $\mathcal{P}_{\theta_i}$ and has a generalized binomial distribution with expectation

$$\mathrm{E}(\tilde{S}_i^c \,|\, \theta_i, \delta) = \sum\nolimits_j \Pr(C_{ij} = 1 \,|\, \theta_i, \delta_j) \tag{60}$$

This expectation must not be confused with the expectation $\mathrm{E}(S_i^c \,|\, S_i = s_i)$ considered in Sections 2 and 3 which is an epistemic expectation for a particular parameter s_i^c postulated for a particular person i. In contrast, the expectation (60) concerns a random variable $\tilde{S}_i^c$ whose values are assumed to be generated by a random process modeled by (53).

References

Andrich, D., Marais, I., Humphry, S. (2012). Using a Theorem by Andersen and the Dichotomous Rasch Model to Assess the Presence of Random Guessing in Multiple Choice Items. *Journal of Educational and Behavioral Statistics* 37, 417–442.

Andrich, D., Marais, I. (2014). Person Proficiency Estimates in the Dichotomous Rasch Model when Random Guessing is Removed from Difficulty Estimates of Multiple Choice Items. *Applied Psychological Measurement* 38, 432–449.

Bartolucci, F. (2007). A Class of Multidimensional IRT Models for Testing Unidimensionality and Clustering Items. *Psychometrika* 72, 141–157.

Blossfeld, H.-P., Roßbach, H.-G., von Maurice, J. (eds.) (2011). Education as a Lifelong Process. The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, Special Issue 14.

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

Duchhardt, C., Gerdes, A. (2012). NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade. *NEPS Working Paper* No. 17. Bamberg: NEPS.

García-Pérez, M. A. (1999). Fitting Logistic IRT Models: Small Wonder, *The Spanish Journal of Psychology* 2, 74–94.

Han, K. T. (2012). Fixing the c Parameter in the Three-Parameter Logistic Model. *Practical Assessment, Research & Evaluation* 17, No. 1.

Keats, J. A. (1974). Applications of Projective Transformations to Test Theory. *Psychometrika* 39, 359–360.

Kubinger, K. D., Draxler, C. (2007). A Comparison of the Rasch Model and Constrained Item Response Theory Models for Pertinent Psychological Test Data. In: M. von Davier, C.H. Carstensen (eds.), *Multivariate and Mixture Distribution Rasch Models*, 293–309, New York: Springer Science.

Li, Y., Jiao, H., Lissitz, R. W. (2012): Applying Multidimensional Item Response Theory Models in Validating Test Dimensionality. *Journal of Applied Testing Technology* 13, Issue #2.

Lord, F. M. (1964). The Effect of Random Guessing on Test Validity. *Educational and Psychological Measurement* 24, 745–747.

Lord, F. M., Novick, M. R., (1968). *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Maris, G. (2002). Concerning the Identification of the 3PL Model. *Measurement and Research Department Reports* 2002-3. Arnheim: Cito National Institute for Educational Measurement.

Neumann, I., Duchhardt, C., Grüßing, M., Heine, A, Knopp, E., Ehmke, T. (2013). Modeling and Assessing Mathematical Competence Over the Lifespan. *Journal for Educational Research Online*, 5, 80–109.

Pohl, S., Haberkorn, K., Hardt, K., Wiegand, E. (2012). NEPS Technical Report for Reading – Scaling Results of Starting Cohort 3 in Fifth Grade. *NEPS Working Paper*, No. 15. Bamberg: Leibniz Institute for Educational Trajectories.

Rohwer, G. (2015). Competence Distributions, Latent Regression Models and Plausible Values. *NEPS Working Paper* No. 55. Bamberg: Leibniz Institute for Educational Trajectories.

San Martin, E., Gonzalez, J., Tuerlinckx, F. (2015). On the Unidentifiability of the Fixed-Effects 3PL Model. *Psychometrika* 80, 450–467.

Verhelst, N. (2001). Testing the Unidimensionality Assumption of the Rasch Model. *Methods of Psychological Research Online*, 6(3), 231–271.

Wang, W.-C. (1999). Direct Estimation of Correlations Among Latent Traits within IRT Framework. *Methods of Psychological Research Online* 4 (2), 47–68.

Warm, T. A. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika* 54, 427–450.

Weitzman, R. A. (1996). The Rasch Model plus Guessing. *Educational and Psychological Measurement* 56, 779–790.

White, P. O. (1976). A Note on Keats' Generalization of the Rasch Model. *Psychometrika* 41, 405–407.