NEPS
National Educational Panel Study

NEPS Working Papers

Götz Rohwer

Making Sense of Missing Answers in Competence Tests

NEPS Working Paper No. 30

Bamberg, October 2013

SPONSORED BY THE
Federal Ministry
of Education
and Research

Contact: German National Educational Panel Study (NEPS) − University of Bamberg − 96045 Bamberg − Germany − contact.neps@uni-bamberg.de

Making Sense of
Missing Answers in Competence Tests

Götz Rohwer, Ruhr-Universität Bochum

October 2013

Email address of the author:
goetz.rohwer@rub.de

Bibliographic data:

Rohwer, G. (2013). *Making Sense of Missing Answers in Competence Tests.* (NEPS Working Paper No. 30). Bamberg: University of Bamberg, National Educational Panel Study.

Making Sense of Missing Answers in Competence Tests

Abstract

The paper discusses how to understand, and cope with, missing answers in competence tests. Starting point is the insight that missing answers in competence tests are not missing values in a normal sense (which hide sensible 'true values') but need an interpretation and evaluation. The paper then distinguishes two types of missing answers: A person is not able or motivated to produce a correct answer (type 1), or there was not enough time to produce a correct answer (type 2).

It is argued that, if items are not of a multiple-choice type, missing answers of type 1 should be evaluated as wrong answers. It is shown that this does not contradict using a Rasch model for the construction of ability values and, in particular, does not lead to 'biased estimates'.

How to cope with missing answers of type 2 depends on which quantities one intends to estimate. Presupposing a standard Rasch model, missing answers of type 2 can be ignored when estimating item parameters. It is shown how this can be done with a conditional likelihood.

With respect to ability values, it depends on whether one is interested in the ability to produce correct answers in the given time limit, or one intends to estimate ability values for situations without time restrictions. In the former case, missing answers of type 2 should be evaluated in the same way as missing answers of type 1.

Given the latter intention, one needs to estimate a counterfactual score distribution for situations without time restrictions. The paper considers one possible estimation method that assumes observed numbers of correct and wrong responses being possibly censored observations, allowing one to use a two-dimensional Kaplan-Meier procedure. The estimated score distribution can then be used for multiple imputations.

Finally, the paper illustrates the discussed methods with NEPS data on math competencies of 5th grade pupils.

Keywords

Competence tests, Rasch models, missing answers, imputation methods

Contents

This paper aims to contribute to the discussion about missing answers in competence tests. The discussion is restricted to tests consisting of binary items providing three possible responses: a correct answer, a wrong answer, or no answer. I assume throughout that missing answers are not missing by design (due to not administered items).

1. How to think of missing answers?

It is important to realize that missing answers in a competence test are not missing values in a usual sense. In the usual sense, missing values "hide true values that are meaningful for analysis" (Little & Rubin 2002: 8). In case of a missing answer in a competence test, one cannot assume that there is an unknown 'true answer'. Missing answers belong to a separate category, different from wrong and correct answers. Missing answers must be interpreted and evaluated in order to establish a relationship with abilities.

How to interpret missing answers depends on the kind of item (question, task). As mentioned, I presuppose items allowing three responses: a correct answer, a wrong answer, or no answer. One can think of three causes leading a person to omit an answer:

a) The person does not answer a question because she has not enough time to consider the question and produce an answer.

b) The person is not able to provide a presumably correct answer, and therefore prefers to give no answer.

c) The person is not sufficiently motivated to produce an answer.

If the questions in a test should be answered in a prescribed sequential order, one can identify, with some confidence, missing answers of the first kind as missing answers to "not reached" items (Lord 1974: 248). Missing answers which are not of this kind can be interpreted in the second or third sense. A clear distinction is not possible. In order to make sense of a competence test, one has to think of abilities that participants of the test are willing to show in their responses. Their motivation, therefore, must be viewed as an integral part of the ability that can be assessed by the test.

So there only remains a rough distinction between two kinds of missing answers: Missing answers expressing that a person is not able or motivated to produce a possibly correct answer, and missing answers due to the fact that there was not enough time to consider the question and produce an answer. In the following, I speak, respectively, of missing answers of type 1 and missing answers of type 2.

Notation

m is the number of items, n is the number of persons. Variables M_j and X_j (for $j = 1, \ldots, m$) represent person's responses. $X_j = 1$ if the answer is correct, $X_j = 0$ if the answer is wrong. M_j can take the following values:

$$M_j = \begin{cases} 0 & \text{if } X_j = 0 \text{ or } X_j = 1 \\ 1 & \text{if there is a missing answer of type 1} \\ 2 & \text{if there is a missing answer of type 2} \end{cases}$$

Observed values of the variables will be denoted by x_{ij} and m_{ij}, respectively. Note: X_j is undefined if $M_j > 0$. Consequently, x_{ij} is undefined and cannot be used if $m_{ij} > 0$.

2. Reference to a scaling model

How to deal with missing answers also depends on the operationalization of the ability that the test is intended to assess. Such operationalizations are often defined by a scaling model. Here I refer to a Rasch model. Using variables $X_1, \ldots, X_m$, representing m binary items, having values $x_j \in \{0,1\}$,[1] the model has the form

$$\Pr(X_1 = x_1, \ldots, X_m = x_m \mid C = c; \delta) \approx \prod_{j=1}^{m} \frac{\exp(c - \delta_j)^{x_j}}{1 + \exp(c - \delta_j)} \tag{1}$$

$\delta := (\delta_1, \ldots, \delta_m)$ is the vector of item parameters, c is the value of a variable, C, that is to be constructed as a quantification of competencies (= abilities for correctly responding to the given items).

Scaling versus explanation

I stress that the Rasch model is a model for the *construction* of a variable, C, intended to represent abilities that can be assessed by the given test items (even if considered as a sample from a larger set of items). The word 'construction' is to be understood in contrast to 'description' – of facts existing independently of the test and the scaling model. Changing the test items, or changing the scaling model, will lead to different 'estimates' of abilities. In fact, the word 'estimation' is ambiguous in the present context. While one can sensibly speak of estimating the parameters of a scaling model, one cannot say that values of C are estimates of ability values which exist in some sense independently of the test and the scaling model.

Referring to the observable variables $X_1, \ldots, X_m$, one can also think of an explanatory model. That would be a model that considers $X_1, \ldots, X_m$ as dependent variables and makes their distribution dependent on values of other observable variables which can be interpreted as representing conditions of processes generating the abilities to be assessed (e.g., age, sex, family background, actual testing conditions). The Rasch model is not an explanatory model. Not just because C is not an observable variable, but because C is derived from the variables $X_1, \ldots, X_m$. Of course, calling C a 'latent variable' cannot change this logical relationship.[2]

That C is derived from $X_1, \ldots, X_m$ has further consequences. Referring to the Rasch model, this entails that there is a one-to-one relationship between values of C and values of a variable S representing a person's total number of correct answers ($s = \sum_j I[x_j = 1]$). Now consider what the Rasch model assumes for a single item:

$$\Pr(X_j = 1 \mid C = c; \delta_j) \approx \frac{\exp(c - \delta_j)}{1 + \exp(c - \delta_j)} \tag{2}$$

This is often interpreted in the following way: The probability of a correct answer to item j depends positively on the ability value c, and negatively on the item parameter δ_j. While this is true in a formal sense, the referenced probability has no operational meaning. In order to better understand the meaning of (2), one can make use of the correspondence between C and S. This shows that the probability defined in (2) relates to all persons belonging to a corresponding score group, say s. For example, the statement $\Pr(X_j = 1 \mid S = s; \delta_j) = 0.6$ would mean that about 60 % of the persons belonging to score group s will give a correct answer to item j. In a sense, one might also say: for a person who is able to correctly answer s of m items, the probability is 0.6 that item j belongs to these s items.

[1] For ease of notation, these are the same variable names as used above for representing valid responses; however, here the variables serve to formulate a model. The intended meaning will become clear from the context.

[2] For an understanding of the notion of 'latent variables' see Rohwer & Pötter 2002: 216-226.

3. Can missing answers be ignored?

Using the Rasch model (1), one has to estimate the item parameters, δ_j, and, for each score group s, a value of C. It seems evident that missing answers cannot be ignored when estimating values of C. Another question, to be discussed in the present section, concerns ignoring missing answers in the estimation of item parameters.

Consider a person i. Let $V_i := \{j \mid m_{ij} = 0\}$ and $U_i := \{j \mid m_{ij} > 0\}$. i's contribution to the likelihood of the data can be written as a product of two terms:

$$\prod_{j \in V_i} \Pr(X_j = x_{ij} \mid c_i, \delta_j) \prod_{j \in U_i} \Pr(M_j = m_{ij} \mid c_i, \delta_j, \phi)$$

Ignoring missing answers means that only the first term is used for parameter estimation. This could be justified if the second term does not depend on parameters occurring in the first term.

This cannot, in general, be assumed for missing answers of type 1. Whether a person omits an answer to an item that she has considered will normally depend both on the kind of item and on her ability (see also Lord 1974: 250-1).

The situation is different for missing answers of type 2. Given that items are presented in a random order, one can reasonably assume that missing answers of type 2 do not depend on item parameters. But do they depend on person's abilities? Here it is important how we understand 'ability'. There are two possibilities:

a) One is interested in the ability to produce correct answers under the given time restriction.

b) One is interested in the ability to produce correct answers in a situation without time restrictions.

In the first case, missing answers of type 2 obviously depend on ability and cannot be ignored. Whether the second understanding can be used to justify ignoring missing answers will be discussed in Section 7.

4. The scaling problem with missing answers

Given that missing answers cannot be ignored, the question is how such answers should be taken into account when estimating the parameters of the scaling model. Is there some way to impute valid responses instead of the missing answers? This would require to use probabilities having the form

$$\Pr(X_j = 1 \mid M_j > 0\,[\text{, further conditions}]) \tag{3}$$

However, for the time being these probabilities are undefined.[3] Therefore, before one can think about how to estimate such probabilities, they need a definition.

One might think that the Rasch model could be used to derive probabilities having the form (3) by using the following definition:

$$\Pr(X_j = 1 \mid M_j > 0,\, C = c;\, \delta_j) := \frac{\exp(c - \delta_j)}{1 + \exp(c - \delta_j)} \tag{4}$$

However, the presupposed ability level, c, is not available if there are missing answers. A person's

[3] This has to be stressed because in the literature it is often assumed that such probabilities have a clear meaning just from the beginning. This is assumed, in particular, by writers who proposed to apply Rubin's theory of missing values to the estimation of IRT models; see, e.g., Mislevy & Wu 1988, 1996.

ability level, as operationalized by the Rasch model, depends on her number of correct answers, but this number is not available until one has decided about the evaluation of her missing answers.

In fact, it is not evident how to make any sense of probabilities having the form (3). Consider the question: What might be a person's response if she would have answered to a question which she, in fact, left unanswered? Since this counterfactual question has no sensible meaning, probabilities having the form (3) cannot be estimated in any clear sense of 'estimation'; and there only remain two ways to deal with missing answers:

a) To base probabilities having the form (3) on an explicit evaluation of the missing answers.

b) To hypothetically construct a situation in which a person can sensibly be assumed to provide an answer to an item that she left unanswered.

In the following, I use the first approach in the discussion of missing answers of type 1 ('omitted'), and the second approach as one possibility to cope with missing answers of type 2 ('not reached').

5. Missing answers of type 1 ('omitted')

I begin with the question how to deal with missing answers of type 1. This depends on the test item. There are basically two kinds depending on whether a correct answer can, or cannot, be produced by guessing.[4]

If correct answers cannot be guessed, a missing answer should be evaluated as a wrong answer. The argument simply is that we are interested in person's abilities to produce *correct* answers. Of course, the participants of the test should be informed about this treatment of their missing answers, and this information should be considered as being an essential part of the test conditions.

Note that evaluating a missing answer as a wrong answer does not contradict the Rasch model because this model does not entail anything about the meaning of missing answers. Consider a person i with a missing answer to item j. The Rasch model entails

$$\text{for all } s > 0: \quad \Pr(X_j = 1 \mid S = s; \delta_j) > 0 \tag{5}$$

This might suggest to think that also person i has a positive probability for correctly answering item j, contradicting the treatment of person i's missing answer as a wrong answer. However, as already noted in Section 2, the probability referred to in (5) does not relate to an individual person but to a score group. The statement (5) simply says that, in a sufficiently large sample, each score group (except $s = 0$) will contain at least one person who provides a correct answer to item j.

Multiple-choice items

The situation is different for multiple-choice items where there is some positive probability for guessing a correct answer.[5] Assume that, for item j, there is a positive probability, say π_j, for guessing a correct answer. It seems sensible, then, to substitute missing answers by wrong and

[4] Duchhardt & Gerdes (2012: 5) report that, of the 24 math items in the NEPS SC3 data, 13 have a multiple-choice form, 11 require a short constructed response. Unfortunately, the different types cannot be identified in the scientific use file.

[5] This is often presupposed by writers criticizing the evaluation of missing answers as wrong answers; e.g., Lord 1983, Mislevy & Wu 1996: 20.

correct answers according to these probabilities. As before, the participants of the test should be informed about this treatment of their missing answers; so they have the choice between guessing on their own or leaving this to the data user. And again, this treatment of missing answers should be considered as being an integral part of the test conditions. This can be justified with the following argument: Using multiple-choice questions in a competence test presupposes that being able to guess an answer is an essential part of the ability to be assessed.

A different approach proposed by Lord

Referring to the discussion in Section 4, the above proposal is based on the definition

$$\Pr(X_j = 1 \mid M_j = 1) := \pi_j \tag{6}$$

where $\pi_j > 0$ if an item is of a multiple-choice type, and $\pi_j = 0$ if it is not possible to produce a correct answer by guessing. Lord (1974: 249) has criticized this evaluation of missing answers and proposed a different approach. His critique presupposes that one should start from (4), so that the above proposal seems to entail

$$\Pr(X_j = 1 \mid M_j > 0,\, C = c;\, \delta_j) = \frac{\exp(c - \delta_j)}{1 + \exp(c - \delta_j)} = \pi_j$$

This obviously is not sensible because it would imply that all persons omitting the same item have the same ability level. However, the critique does not apply to definition (6) which does not presuppose that (4) is meaningful.

Based on the view that one should start from (4), Lord (1974: 251) proposed a different approach (see also Mislevy & Wu 1996: 19-20). He proposed to consider a missing answer as ‘fractionally correct’ and use the following pseudo likelihood:

$$\prod_{i=1}^{n} \prod_{j=1}^{m} \Pr(X_j = 1 \mid c_i, \delta_j)^{u_{ij}} \Pr(X_j = 1 \mid c_i, \delta_j)^{1-u_{ij}} \tag{7}$$

where $u_{ij} = x_{ij}$ if $m_{ij} = 0$, and $u_{ij} = \pi_j$ if $m_{ij} = 1$. Note that this is not a likelihood in the usual sense because it is not based on a model for the occurrence of missing answers. In fact, it is a proposal for the evaluation of missing answers.

This also shows up in the calculation of ability values. Given estimates of the item parameters, $\hat{\delta}_j$, such values can be calculated by solving the equation

$$\sum_{j=1}^{m} \frac{\exp(c_i - \hat{\delta}_j)}{1 + \exp(c_i - \hat{\delta}_j)} = \sum_{j \in V_i} x_{ij} + \sum_{j \in U_i} \pi_j \tag{8}$$

which is implied by the first-order conditions for the maximization of (7). The right-hand side is the total number of correctly answered items plus the value of ‘fractionally correct’ items. In this respect, Lord’s proposal is similar to the above mentioned approach where missing answers are substituted by correct and wrong answers, randomly drawn with probabilities π_j and $1-\pi_j$, respectively. The expectation of the total-right score then equals the right-hand side of (8).

6. Simulation studies

The methods for dealing with missing answers of type 1 discussed in the previous section are based on evaluations to be considered as being part of the test conditions. In other words, these evaluations belong to the operationalization of the ability to be assessed by the test. It would not make sense, therefore, to think of the resulting ability values as being possibly biased estimates.

In order to stress this point, I briefly consider simulation studies that have been proposed to investigate strategies for coping with missing answers. Most of these simulation studies (e.g., De Ayala, Plake & Impara 2001, Finch 2008, Culbertson 2011) proceed as follows:

a) One presupposes a scaling model with known item parameters for, say, m items. This can be a Rasch model as specified in (1), but also used are logistic models with two or three parameters.

b) One assumes a set of individuals with a known distribution of the variable C. This allows generating individual values c_i for $i = 1, \ldots, n$.

c) One generates for each individual a response pattern $(x_{i1}, \ldots, x_{im})$ with $x_{ij} \in \{0, 1\}$. There are different ways to do this. For example, De Ayala, Plake & Impara (2001) begin with the calculation of probabilities

$$p_{ij} := \Pr(X_j = 1 \mid C = c_i; \text{parameters for item } j) \tag{9}$$

They then compare these probabilities with random numbers equally distributed in the interval $[0, 1]$. "If the random number was less than or equal to the probability of a correct response, the response was coded as "1" for correct, "0" otherwise." (p. 218)

d) One specifies a mechanism for the generation of missing answers. This mechanism can be made dependent on the item parameters and on values of C or the sum score of the already generated x_{ij} values.

e) One uses the mechanism to mark some of the previously generated x_{ij} values as missing answers.

The finally generated data set can be used, in combination with different strategies for the treatment of missing answers, to estimate values of the item parameters and/or the distribution of the variable C, and these estimates can be compared with their presupposed values.

What can be learned from such simulation studies? In particular, can such studies show that certain strategies to take into account missing answers produce more or less biased estimates? As an example, I consider the strategy to treat missing answers as wrong answers. Several authors have claimed that their simulations showed that this strategy, more than other ones, produced biased estimates.[6] However, one can easily see that this result is simply a consequence of a specific way of generating missing answers in the simulation.

Remember how the simulated data are generated. One first uses (9) to create a data set without missing answers. Then, a selected subset of both correct and wrong responses is marked as missing. Obviously, if one subsequently treats all missing answers as wrong answers, estimates will be biased.

However, other ways to generate missing answers will lead to different results. For example, one could mark only wrong answers as missing, and then all strategies which do *not* treat missing answers as wrong answers would produce biased results. In general, when a simulation study reveals a bias, this only reflects a discrepancy between the method of generating missing answers and the treatment of such answers when estimating model parameters or calculating ability values.

There is, however, a more fundamental shortcoming. By assuming a given distribution of ability values, in accord with the presupposed scaling model, the studies neglect the primarily important

[6] E.g., De Ayala, Plake & Impara 2001, Finch 2008, Culbertson 2011; see also Pohl and Carstensen 2012: 8-9.

question: How to establish a relationship between missing answers and abilities? Instead of addressing this question as part of the operationalization of the abilities to be assessed by the test, the studies simply assume that missing answers are irrelevant for the definition of abilities.

7. Missing answers of type 2 (‘not reached’)

I refer to the two understandings of ‘ability’ distinguished in Section 3. If one is interested in person’s abilities to produce correct answers in the given time limit, missing answers of type 2 should be treated in the same way as missing answers of type 1. As was discussed in Section 5, this further depends on whether the items are of a multiple-choice type or not.

In the literature, several writers prefer an understanding of ‘ability’ in the second sense where time restrictions are hypothetically ignored (e.g., Lord 1974, Mislevy & Wu 1996: 12). In the following, I presuppose this understanding.

I also assume that missing answers of type 1 have been substituted in some way, so that all remaining missing answers are of type 2.

Estimation of item parameters

As was mentioned in Section 3, if items are presented in a random order, one can reasonably assume that missing answers of type 2 do not depend on item parameters. Such missing answers could then be ignored when estimating item parameters with a conditional likelihood.

But note that the argument does not apply when item parameters and person’s ability values are estimated simultaneously (e.g., marginal likelihood estimation of a Rasch model with a random term). I therefore consider, in Section 8, an alternative approach that is not restricted to conditional likelihood estimation.

Conditional likelihood estimation of item parameters

In the present subsection, to ease notation, I assume that $x_{ij} = 2$ if $m_{ij} = 2$. The conditioning uses variables S^0, S^1, and S^2 having values defined, respectively, by

$$w_i := \sum\nolimits_j I[x_{ij}=0],\ s_i := \sum\nolimits_j I[x_{ij}=1],\ u_i := \sum\nolimits_j I[x_{ij}=2]$$

Since all missing answers are of type 2, $w_i+s_i+u_i = m$. Person i’s contribution to the conditional likelihood can then be written as

$$\begin{aligned}
\mathcal{L}_i^{\text{con}} &= \Pr(X_1=x_{i1},\dots,X_m=x_{im} \mid c_i,\delta,S^0=w_i,S^1=s_i,S^2=u_i) \\
&= \frac{\Pr(X_1=x_{i1},\dots,X_m=x_{im},S^0=w_i,S^1=s_i,S^2=u_i \mid c_i,\delta)}{\Pr(S^0=w_i,S^1=s_i,S^2=u_i \mid c_i,\delta)} \\
&= \frac{\Pr(X_1=x_{i1},\dots,X_m=x_{im} \mid c_i,\delta)}{\sum_{x\in\mathcal{D}(w_i,s_i)} \Pr(X_1=x_1,\dots,X_m=x_m \mid c_i,\delta)}
\end{aligned}$$

where $\mathcal{D}(w_i, s_i)$ denotes the set

$$\{x = (x_1,\dots,x_m) \mid x_j \in \{0,1,2\}, \Sigma_j I[x_j=0] = w_i, \Sigma_j I[x_j=1] = s_i\}$$

When ignoring missing answers, this can be written as

$$\begin{aligned}\mathcal{L}_i^{\text{con}} &= \frac{\prod_{j\in V_i} \frac{\exp(c_i-\delta_j)^{x_{ij}}}{1+\exp(c_i-\delta_j)}}{\sum_{x\in\mathcal{D}(w_i,s_i)} \prod_{j\in V_i} \frac{\exp(c_i-\delta_j)^{x_j}}{1+\exp(c_i-\delta_j)}} \\ &= \frac{\prod_{j\in V_i} \exp(c_i-\delta_j)^{x_{ij}}}{\sum_{x\in\mathcal{D}(w_i,s_i)} \prod_{j\in V_i} \exp(c_i-\delta_j)^{x_j}} \\ &= \frac{\exp(s_i\, c_i) \prod_{j\in V_i} \exp(-\delta_j\, x_{ij})}{\sum_{x\in\mathcal{D}(w_i,s_i)} \exp(s_i\, c_i) \prod_{j\in V_i} \exp(-\delta_j\, x_j)} \\ &= \frac{\prod_{j\in V_i} \exp(-\delta_j\, x_{ij})}{\sum_{x\in\mathcal{D}(w_i,s_i)} \prod_{j\in V_i} \exp(-\delta_j\, x_j)}\end{aligned}$$

If missing answers of type 2 occur as contiguous blocks at the end of the response patterns, this simplifies to

$$\mathcal{L}_i^{\text{con}} = \frac{\prod_{j=1}^{v_i} \exp(-\delta_j\, x_{ij})}{\sum_{x\in\mathcal{D}_{s_i}^{v_i}} \prod_{j=1}^{v_i} \exp(-\delta_j\, x_j)} \tag{10}$$

where $v_i := w_i + s_i$ and $\mathcal{D}_{s_i}^{v_i}$ is the set of patterns of length v_i containing s_i ones.

Estimation of ability values

Having estimated item parameters, one could estimate ability values by solving the equation

$$\sum_{j=1}^{m} \frac{\exp(c_i - \hat{\delta}_j)}{1 + \exp(c_i - \hat{\delta}_j)} = s_i^* \tag{11}$$

which is implied by the first-order conditions for maximizing the likelihood of the Rasch model (1). s_i^* denotes person i's number of correct answers. If there were no missing answers of type 2 (and all missing answers of type 1 have been substituted in some way), one could use the observed value s_i. On the other hand, if there are missing answers of type 2, one cannot use the observed s_i, but would need an estimate of the number of correct answers the person might have produced if she had enough time to deal with all items.

Alternatively, one could again ignore missing answers of type 2 and use

$$\sum_{j\in V_i} \frac{\exp(c_i^* - \hat{\delta}_j)}{1 + \exp(c_i^* - \hat{\delta}_j)} = s_i^* \tag{12}$$

where V_i is the index set of person i's valid answers. Of course, implicitly also this approach entails an estimate of a person's number of correct answers for a hypothetical situation without time restrictions.

8. Using estimated score distributions

I assume that one intends to explicitly construct ability values which do not entail time restrictions. One then has to consider a hypothetical question: What would be the number of correct answers if a person had enough time to deal with all items? In other words, one has to estimate the distribution of a variable, say S^*, that represents the scores (= number of correct answers) in a counterfactual situation without time restrictions. In this section, I first describe a possible estimation procedure for S^*. I then discuss how the distribution of S^* can be used for the estimation of item parameters and ability values.

Box 8.1 Algorithm for the estimation of the score distribution.

```
$f(s,w) :=$ proportion of persons with $s_i = s$ and $w_i = w$

$b(s,w) := \{(k,l) \mid s \le k \le m{-}w, w \le l \le m{-}k, (k \ne s \text{ or } l \ne w)\}$

for $s = 0,\dots,m-1$
  for $w = 0,\dots,m-s-1$
    $F(s,w) \leftarrow \sum_{(k,l)\in b(s,w)} f(k,l)$
    for all $(k,l) \in b(s,w) : f(k,l) \leftarrow f(k,l) + f(s,w)\,\frac{f(k,l)}{F(s,w)}$

for $s = 0,\dots,m :\ \Pr(S^*{=}s) \leftarrow f(s,m{-}s)$
```

Estimation of the score distribution

Estimation of the distribution of S^* should be based on the observed values s_i and w_i. I assume that missing answers of type 1 have been substituted in some way. The number of missing answers of type 2 is then given by $m - v_i$ where $v_i = s_i + w_i$ is the number of valid answers.

To explain the approach, I refer to the variable (S^*, W^*) where $W^* := m - S^*$. The basic idea is to consider (s_i, w_i) as an exact observation of (S^*, W^*) if $s_i + w_i = m$, and as a censored observation if $s_i + w_i < m$. In order to take into account censored observations for the estimation of the distribution of (S^*, W^*), one can use a two-dimensional Kaplan-Meier procedure. Box 8.1 shows the algorithm for our application.[7]

Since $W^* = m - S^*$, it finally suffices to consider the distribution of S^*. This distribution can be used for two purposes. For each person i, one can calculate a conditional expectation

$$\begin{aligned} \mathrm{E}(S^* \mid S^* \ge s_i, W^* \ge w_i) &= \mathrm{E}(S^* \mid s_i \le S^* \le m{-}w_i) \\ &= \frac{\sum_{s_i \le s \le m-w_i} s \Pr(S^* = s)}{\sum_{s_i \le s \le m-w_i} \Pr(S^* = s)} \end{aligned}$$

These expectations equal the observed scores, s_i, if there are no missing answers of type 2. As a further application, one can use the conditional distributions $\Pr(S^* \mid s_i \le S^* \le m{-}w_i)$ for multiple imputations. This will be discussed next.

Estimation of item parameters

The procedure is shown in Box 8.2. Imputations are based on the estimated score distribution, conditional on known numbers of correct and wrong answers. The number of imputation steps is made dependent on the convergence of the sequence $\bar{\delta}_j^{(k)}$ for $k = 1, 2, \dots$ For checking convergence one can use, for example,

$$\frac{1}{m} \max_j \left\{ \left| \bar{\delta}_j^{(k)} - \bar{\delta}_j^{(k-1)} \right| \right\} \le \epsilon$$

where ϵ is a small number.

[7] For a discussion of this, and similar, procedures see Pötter 2008, esp. chap. 8.

Box 8.2 Algorithm for the estimation of item parameters.

```
(1) $k \leftarrow 0$

(2) $k \leftarrow k + 1$

(3) for all $i$ with $s_i + w_i < m$:
    randomly draw a value $s_i^{(k)}$ from the conditional distribution
    $\Pr(S^* \mid s_i \leq S^* \leq m - w_i)$;
    randomly select $s_i^{(k)} - s_i$ of the $m - s_i - w_i$ missing $x_{ij}$-values
    and set these values to 1; set the remaining missing values to 0.

(4) estimate $\delta_1^{(k)}, \ldots, \delta_m^{(k)}$ with current $(x_{ij})$

(5) for $j = 1, \ldots, m$: $\bar{\delta}_j^{(k)} \leftarrow \sum_{l=1}^{k} \delta_j^{(l)} / k$

(6) check whether to stop the iterations, otherwise continue with (2).
```

Estimation of ability values

Having estimated item parameters, one can calculate, for each score group $s = 1, \ldots, m-1$, an ability value $c(s)$, by solving the equation

$$\sum_{j=1}^{m} \frac{\exp(c(s) - \hat{\delta}_j)}{1 + \exp(c(s) - \hat{\delta}_j)} = s \tag{13}$$

Persons without missing answers can then immediately be given the ability value $c(s_i)$. If there are missing answers, one can use a mean value

$$c^{+}(s_i, w_i) := \frac{\sum_{s_i \leq s \leq m - w_i} c(s) \Pr(S^* = s)}{\sum_{s_i \leq s \leq m - w_i} \Pr(S^* = s)} \tag{14}$$

based on the estimated score distribution, conditional on known numbers of correct and wrong answers.

9. Illustration with NEPS data

To illustrate the discussion, I use NEPS data on math competencies of 5th grade pupils.[8] I use the data file `SC3_xTargetCompetencies_D_1-0-0.sav` that is part of the SPSS version of the SC3 SUF.[9] The file contains information about 5208 pupils who participated in the competence tests. There are 24 items for math competencies. I use 23 of these items which are binary. Table 9.1 shows the distribution of their values.

[8] Acknowledgement: This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3 – 5th grade (From Lower to Secondary School), doi:10.5157/NEPS:SC3:1.0.0. The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States. For a general introduction to the NEPS, see Blossfeld, Roßbach and von Maurice (eds.) 2012.

[9] For a description of the SC3 SUF, see Skopek, Pink and Bela (2012). Additional information about the mathematics test is given by Duchhardt and Gerdes (2012).

Table 9.1 Valid answers and missing values in 23 items for math competencies.

Item	Variable	-97	-95	-94	0	1
X1	MAG5D041	66	17	14	2166	2945
X2	MAG5Q291	232	43	14	1442	3477
X3	MAG5Q292	256	37	14	1645	3256
X4	MAG5V271	436	4	14	3264	1490
X5	MAG5R171	179	11	16	2411	2591
X6	MAG5Q231	609	291	16	2515	1777
X7	MAG5Q301	116	93	16	3065	1918
X8	MAG5Q221	154	30	17	864	4143
X9	MAG5D051	84	4	17	562	4541
X10	MAG5D052	81	127	17	1986	2997
X11	MAG5Q14S	584	127	18	1553	2926
X12	MAG5Q121	431	10	21	3695	1051
X13	MAG5R101	130	69	23	2366	2620
X14	MAG5R201	115	6	28	1352	3707
X15	MAG5Q131	237	67	38	1131	3735
X16	MAG5D02S	318	154	46	656	4034
X17	MAG5D023	374	45	53	1923	2813
X18	MAG5V024	727	228	67	1920	2266
X19	MAG5R251	360	13	98	2571	2166
X20	MAG5V321	548	58	205	2997	1400
X21	MAG5V071	70	29	223	501	4385
X22	MAG5R191	47	217	284	2057	2603
X23	MAG5V091	0	23	465	2669	2051

Table 9.2 Crosstabulation of of number of missing values of type 1 (rows) and missing values of type 2 (columns).

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	16	19	23	
0	2076	24	14	5	27	13	5	4	2	3	1	1	2		1	1	14	2193
1	1107	34	11	7	19	4	2	2	1	1	1	1		1				1191
2	596	20	7	2	23	5	4		4		2					1		664
3	369	25	10	4	12	3	2	1		2	1		1					430
4	232	16	6		6	2	1			2								265
5	163	20	6		9	3			1									202
6	80	10	2		7	1				2								102
7	65	9	2		2													78
8	22	5	1															28
9	11	6																17
10	7	4	1		1													13
11	5	3			1													9
12	3	1																4
13	5	1	1															7
14	1	2																3
16		1																1
17	1																	1
	4743	181	61	18	107	31	14	7	8	10	5	2	3	1	1	2	14	5208

There are three types of missing values: -97 (refused), -95 (implausible value), and -94 (not reached). Based on the discussion in Section 1, I treat -95 and -97 as missing answers of type 1,

Table 9.3 Comparison of item parameters, estimated with a conditional likelihood (values are standardized such that $\Sigma_j \delta_j = 0$).

Item	all missing answers treated as wrong	missing answers of type 2 ignored	missing answers of type 2 imputed
1	-0.0776	-0.0402	-0.0375
2	-0.6103	-0.5749	-0.5705
3	-0.3823	-0.3469	-0.3440
4	1.3648	1.3978	1.3908
5	0.2612	0.2957	0.2952
6	1.0555	1.0889	1.0826
7	0.9116	0.9453	0.9408
8	-1.4083	-1.3758	-1.3638
9	-2.0634	-2.0328	-2.0192
10	-0.1266	-0.0927	-0.0919
11	-0.0580	-0.0241	-0.0247
12	1.9104	1.9403	1.9272
13	0.2338	0.2653	0.2622
14	-0.8624	-0.8369	-0.8289
15	-0.8944	-0.8749	-0.8664
16	-1.2603	-1.2543	-1.2350
17	0.0504	0.0699	0.0667
18	0.5706	0.5885	0.5766
19	0.6673	0.6746	0.6587
20	1.4681	1.4502	1.3730
21	-1.7791	-2.0777	-1.9251
22	0.2500	0.1625	0.1501
23	0.7793	0.6518	0.5831

and -94 as missing answers of type 2.[10]

Table 9.2 shows a crosstabulation of the numbers of the two types of missing answers. There are 14 persons having no valid response at all. These persons will be omitted in subsequent calculations. 451 of the remaining 5194 persons have at least one missing answer of type 2.

9.1 Estimation of item parameters

If one is interested in the ability to produce correct answers in the given time limit, all missing answers should be evaluated as wrong answers. Alternatively, if time restrictions should be ignored, one could ignore missing answers of type 2 as was discussed in Section 7. Corresponding item parameters, estimated with a conditional likelihood, are shown in Table 9.3. As also seen from Figure 9.1, the estimates are highly correlated.

This is due to the fact that the number of missing answers of type 2 is small (less than 10 %). Note that the item parameters of a Rasch model are basically a nonlinear transformation of the proportions of wrong answers (often interpreted as 'item difficulties'). This is shown in Figure 9.2. Consequently, ignoring missing answers makes items 'less difficult'. However, due to the standardization $\Sigma_j \delta_j = 0$, this cannot immediately be seen from the values in Table 9.3.

[10] This is done in order to illustrate the discussed methods. Actually, it is not clear whether the NEPS data allow identifying missing answers of type 2 in a strict sense.

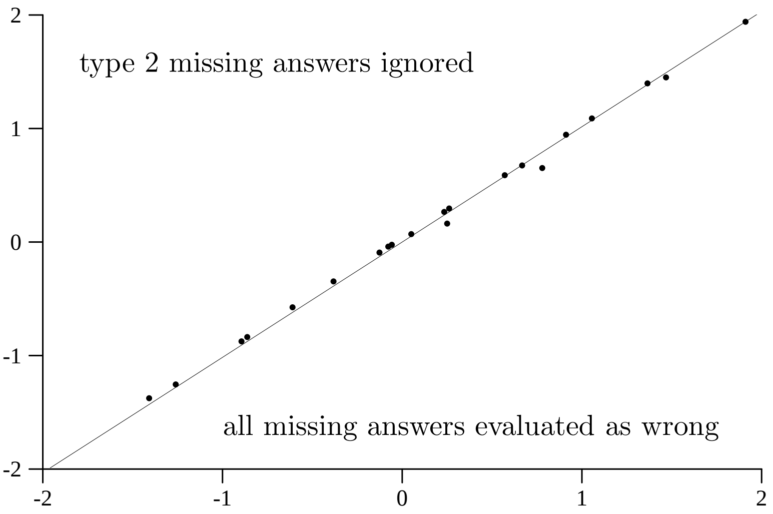

Fig. 9.1 Scatterplot of item parameters, also shown a linear regression line.

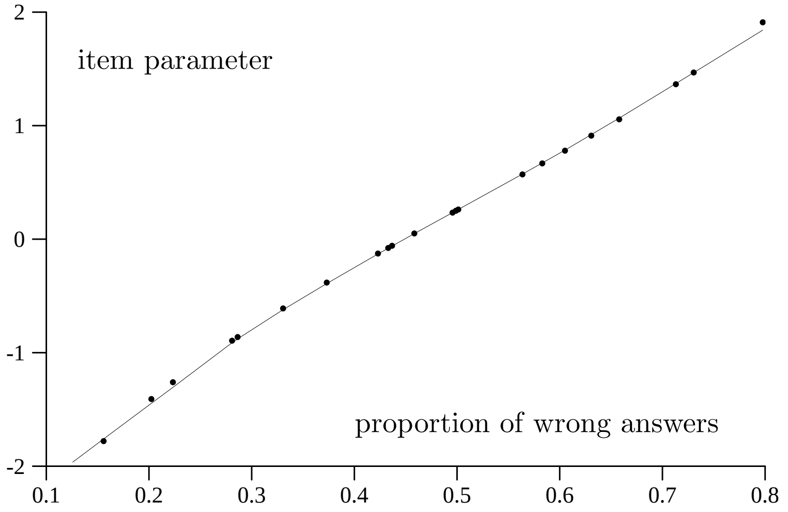

Fig. 9.2 Scatterplot of proportion of wrong answers and item parameters, based on a Rasch model that treats all missing answers as wrong.

9.2 Estimation of ability values

Having available estimates of item parameters, ability values can be calculated by solving equation (13) where s is a person's total number of correct answers For each sum score s, one gets an ability score $c(s)$.[11] Note that the calculation requires $0 < s < m$.[12]

Table 9.4 compares the ability scores for the two ways of treating missing answers. As also seen from Figure 9.3, the estimates are almost identical.

If all missing answers are evaluated as wrong answers, these ability values can immediately be used for individual assignment. If missing answers of type 2 are ignored in the calculation of

[11] If some, or all, items have a multiple-choice form, one could also use the formula (8) which was proposed by Lord. However, this cannot be illustrated here because the presently available SUF does not allow to identify multiple-choice items.

[12] In order to calculate ability scores also for persons having all items right, or wrong, one could use weighted likelihood estimates as proposed by Warm (1989).

Table 9.4 Ability scores calculated with equation (13) using the item parameters in Table 9.3.

S	all missing answers treated as wrong	missing answers of type 2 ignored	missing answers of type 2 imputed
1	-3.4870	-3.5517	-3.5265
2	-2.6937	-2.7518	-2.7314
3	-2.1874	-2.2397	-2.2237
4	-1.7979	-1.8453	-1.8332
5	-1.4718	-1.5153	-1.5066
6	-1.1851	-1.2254	-1.2196
7	-0.9245	-0.9623	-0.9592
8	-0.6820	-0.7179	-0.7171
9	-0.4521	-0.4866	-0.4879
10	-0.2306	-0.2641	-0.2672
11	-0.0141	-0.0469	-0.0517
12	0.2002	0.1680	0.1615
13	0.4152	0.3833	0.3754
14	0.6338	0.6022	0.5929
15	0.8594	0.8280	0.8174
16	1.0962	1.0649	1.0530
17	1.3495	1.3184	1.3053
18	1.6278	1.5969	1.5826
19	1.9440	1.9134	1.8980
20	2.3224	2.2922	2.2758
21	2.8167	2.7870	2.7696
22	3.5971	3.5681	3.5497

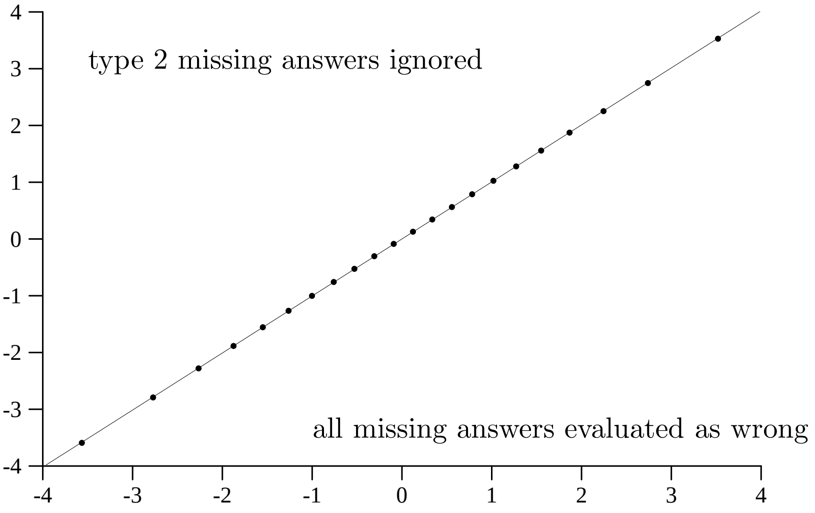

Fig. 9.3 Scatterplot of ability scores calculated with equation (13) using the item parameters in Table 9.3. Also shown is a linear regression line.

item parameters, one needs a further decision for the calculation of individual ability values. There are two possibilities: (a) One treats missing answers of type 2 as wrong answers and then can use formula (11), or (b) one ignores missing answers of type 2 and uses formula (12). The second approach entails that one intends to ignore time restrictions. Unfortunately, this remains implicit and without justification. As an explicit alternative, I therefore consider the approach discussed in Section 8..

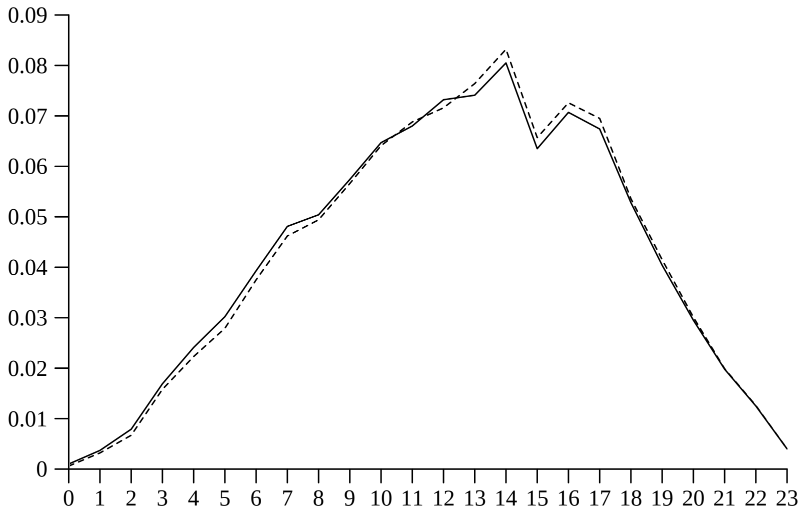

Fig. 9.4 Distributions of sum scores (number of correct answers). Solid: observed, dashed: estimated with the algorithm in Box 8.1.

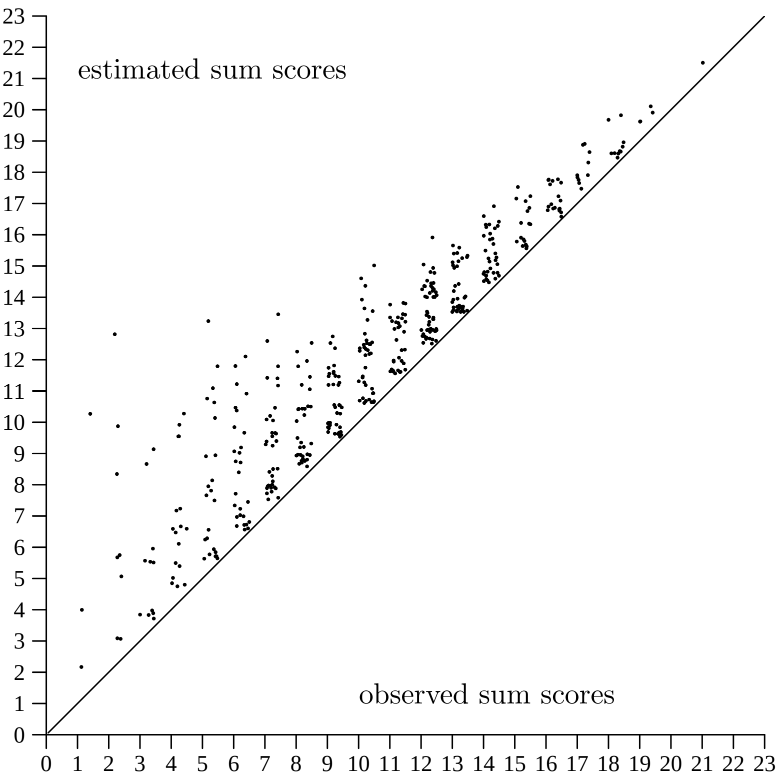

Fig. 9.5 Observed and estimated sum scores (number of correct answers) of 449 persons having at least one missing answer of type 2 and at least one correct answer.

9.3 Using estimated score distributions

I begin with the estimation of a score distribution with the algorithm depicted in Box 8.1. As shown in Figure 9.4, there is a slight shift to higher sum scores. However, for the 449 persons with at least one missing answer of type 2 (and at least one correct answer), the change is quite remarkable, see Figure 9.5.

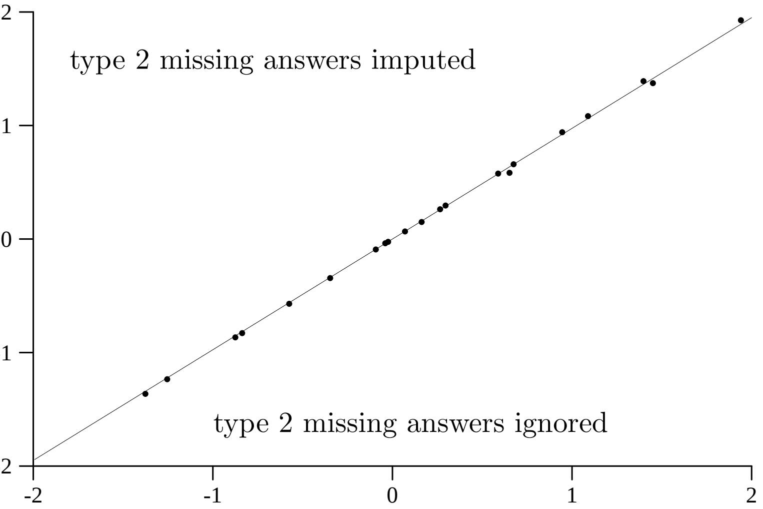

Fig. 9.6 Scatterplot of item parameters, also shown a linear regression line.

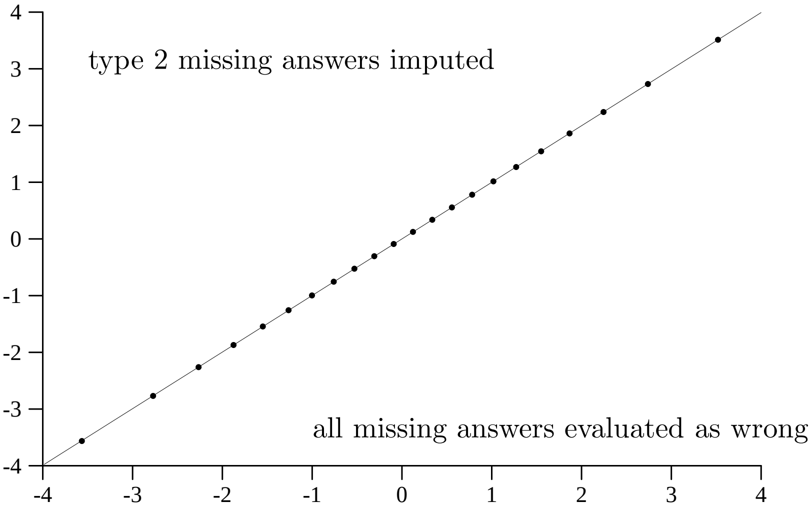

Fig. 9.7 Scatterplot of ability scores calculated with equation (13) using the item parameters in Table 9.3. Also shown is a linear regression line.

Having available an estimate of the distribution of S^*, one can use the algorithm depicted in Box 8.2 for the estimation of item parameters. The resulting estimates (after performing 10 iterations) are shown in Table 9.3. Obviously, one gets again quite similar values (see also Figure 9.6).

Finally, one can calculate ability scores. Using equation (13), for each score group s one gets a corresponding ability value $c(s)$. As can be seen from Table 9.4 and Figure 9.7 they are very similar to previously calculated values. However, when using formula (14), persons with missing answers of type 2 get substantial higher ability values. This is illustrated in Figure 9.8.

Results are similar when missing values of type 2 are simply ignored. As seen by Figure 9.9, ability values resulting, respectively, from (12) and (14) are highly correlated. Of course, both kinds of values only relate to hypothetical test situations without any time restrictions.

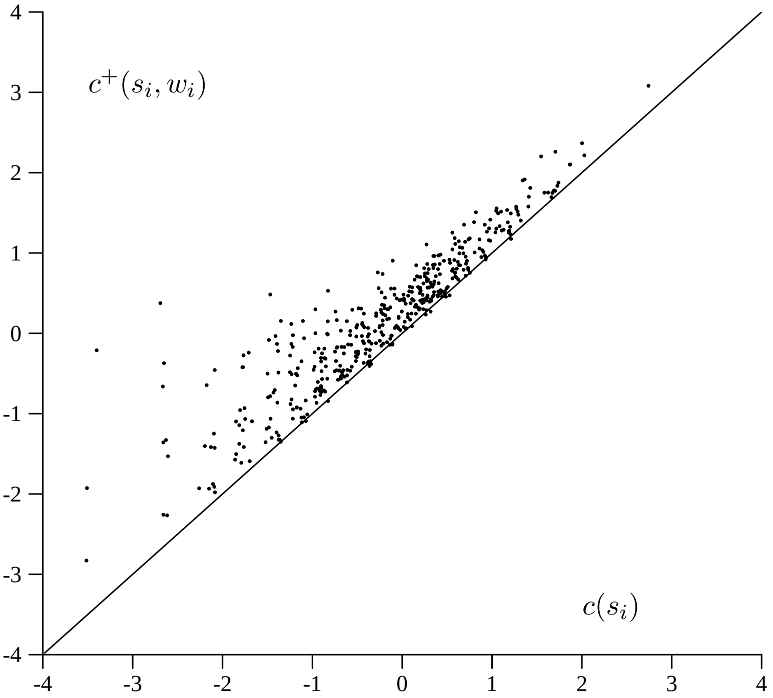

Fig. 9.8 Comparison of ability scores of 449 persons having at least one missing answer of type 2 and at least one correct answer: $c(s_i)$ calculated with (13), $c^+(s_i, w_i)$ calculated with (14).

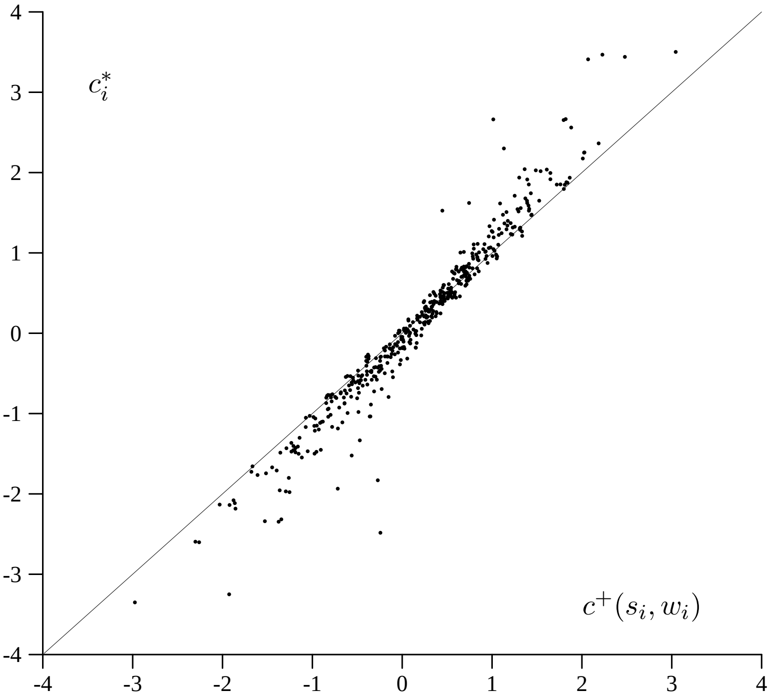

Fig. 9.9 Comparison of ability scores of 449 persons having at least one missing answer of type 2 and at least one correct answer: $c^+(s_i, w_i)$ calculated with (14); c_i^* calculated with (12).

References

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.

Culbertson, M. J. (2011). Is It Wrong? Handling Missing Responses in IRT. Paper presented at the Annual Meeting of the National Council of Measurement in Education. New Orleans, April 2011.

De Ayala, R. J., Plake, B. S., Impara, J. C. (2001). The Impact of Omitted Responses on the Accuracy of Ability Estimation in Item Response Theory. *Journal of Educational Measurement* 38, 213–234.

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.

Finch, H. (2008): Estimation of Item Response Theory Parameters in the Presence of Missing Data. *Journal of Educational Measurement* 45, 225–245.

Little, R. J. A., Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. 2nd Ed. New York: Wiley.

Lord, F. M. (1974). Estimation of Latent Ability and Item Parameters when there are Omitted Responses. *Psychometrika* 39, 247–264.

Lord, F. M. (1983). Maximum Likelihood Estimation of Item Response Parameters when Some Responses are Omitted. *Psychometrika* 48, 477–482.

Mislevy, R. J., Wu, P.-K. (1988). Inferring Examinee Ability when some Item Responses are Missing (RR 88-48-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Wu, P.-K. (1996). Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing (RR-96-30-ONR). Princeton, NJ: Educational Testing Service.

Pohl, S., Carstensen, C. H. (2012). NEPS Technical Report – Scaling the Data of the Competence Tests. *NEPS Working Paper* No. 14. Bamberg: Otto Friedrich Universität, Nationales Bildungspanel.

Pötter, U. (2008). *Statistical Models of Incomplete Data and their Use in the Social Sciences*. Bochum: Ruhr-Universität.

Rohwer, G., Pötter, U. (2002). *Methoden sozialwissenschaftlicher Datenkonstruktion.* Weinheim: Juventa.

Skopek, J., Pink, S., Bela, D. (2012). Data Manual. Starting Cohort 3 - From Lower to Upper Secondary School. NEPS SC3 1.0.0. *NEPS Research Data Paper*, University of Bamberg.

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