NEPS
National Educational Panel Study

NEPS Working Papers

Götz Rohwer

Competencies as Dependent Variables in Regression Models

NEPS Working Paper No. 33

Bamberg, January 2014

A STUDY BY
LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

Working Papers of the German National Educational Panel Study (NEPS)
at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS Working Papers publish articles, expertises, and findings related to the German National Educational Panel Study (NEPS).
The NEPS Working Papers are edited by a board of researchers representing the wide range of disciplines covered by NEPS. The series started in 2011.

Papers appear in this series as work in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character.

Any opinions expressed in this series are those of the author(s) and not those of the NEPS Consortium.

The NEPS Working Papers are available at
https://www.neps-data.de/projektübersicht/publikationen/nepsworkingpapers

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg − Germany − contact.neps@lifbi.de

Competencies as Dependent Variables in Regression Models

Götz Rohwer, Ruhr-Universität Bochum

January 2014

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

Bibliographic data:

Rohwer, G. (2014). *Competencies as dependent variables in regression models* (NEPS Working Paper No. 33). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Competencies as Dependent Variables in Regression Models

Abstract

This paper considers methods to define dependent variables representing results of competence tests; as an example I refer to NEPS data on math competencies of 5th grade pupils. The simplest and easily comprehensible method is to use the number of correct responses in a competence test as values of a quantitative dependent variable in a regression model. Instead of simply using the number of correct responses one can define weighted versions which could take into account that items might have different importance for the competence that the test is intended to measure. However, I show that it is easily misleading to think of such weights as 'item difficulties' which can be derived from proportions of wrong responses.

Instead of using these simple approaches to the construction of a dependent variable, one can start from a probabilistic framework. As an example, I consider a Rasch model that allows one to construct a variable representing latent competencies which can subsequently be used as a dependent variable in regression models. I argue that this approach has two disadvantages, compared with using a simple summary index. The Rasch model introduces a nonlinear metric which is difficult to understand and therefore makes it difficult to interpret effects of explanatory variables. Moreover, the Rasch model employs a notion of 'item difficulties' which are derived from the distribution of competencies of the persons participating in the test.

I then discuss the proposal to use so-called plausible values for the construction of dependent variables. I distinguish between versions with and without conditioning variables. I show that using plausible values, when derived from models including conditioning variables, entail striking forms of statistical discrimination, and propose that this approach should not be used for sociological analyses.

Finally, I briefly consider models which avoid a reference to latent competencies and instead directly relate the observable response patterns to values of explanatory variables. While attractive at first sight, this approach has the drawback that such models must be supplemented by a procedure for aggregating item-specific probabilities.

Keywords

Competence data, scaling methods, overt and latent competencies, Rasch models, unobserved heterogeneity

Contents

1. Introduction

The National Educational Panel Study (NEPS) collects a large amount of statistical data on the development of competencies across the life course. These data can be used for many different research questions. This paper takes a sociological view which, for the present work, I demarcate by two ideas.

First, the main interest concerns people's living conditions (understood in a broad sense). With respect to competencies, this leads to two complementary questions: (a) How do people's living conditions contribute to the development of their competencies? (b) How do people's competencies contribute to the development of their living conditions? In this paper, I discuss some possibilities for approaching the first of these questions with statistical data and methods. The focus is on how to construct variables representing people's competencies that can be used as dependent variables in regression analyses.

In discussing this question, I presuppose a second connotation of the sociological view: One is not interested in the characterization of individual persons, but refers to groups of people defined by social categories. Of course, at the beginning the data relate to individual persons. This is true, in particular, for competence data which result from testing individual persons with competence tests. Since such tests consist of several items (questions or tasks), there is an aggregation problem already on the individual level. Explicitly, I refer to a test consisting of m items and assume that n persons participate in the test; x_{ij} denotes the response of person i to item j. To simplify the notations, I assume that there are only two possible responses: $x_{ij} = 1$ if the response is correct and otherwise $x_{ij} = 0$. So there are m responses, $x_{i1}, \ldots, x_{im}$, for each person i. These item-specific responses must be aggregated in some way in order to get an overall test result that can be interpreted as indicating a person's competence (in the domain to which the test relates).

However, the goal is not to predict the individual competencies of persons who participated in the NEPS competence tests. Instead, in statistical parlance, the interest concerns conditional distributions of competencies. The task is to compare such distributions and to find variables which can contribute to an explanation of differences.

It follows that the aggregation of item-specific responses should be done in such a way that the derived overall test results can be used as values of a dependent variable in regression models. There are basically two possibilities: (1) A two-step procedure. In a first step, based on the item-specific responses, one constructs a single variable representing the overall test result. Then, in a second step, this variable is used to compare groups of persons or, more general, as a dependent variable in a regression model. (2) One uses regression models which directly relate to the item-specific responses, and possibly adds an aggregation procedure afterwards.

I discuss both possibilities. In Section 2, I consider simple index constructions which do not presuppose a probabilistic model, and I show how such indices can immediately be used as dependent variables in regression models.

In Section 3, I begin with discussing the idea to use a probabilistic framework for the representation of competencies. I distinguish two versions. In one version, it is assumed that probabilities of correct responses directly reflect competencies. In another version, one assumes that competencies should be conceptualized in terms of latent variables which in some sense 'explain' the probabilities of correct responses. This is further discussed in Section 4 where a Rasch model is used to construct a dependent variable for further regression analyses.

In Section 5, I discuss the proposal to use so-called plausible values for the construction of dependent variables of regression analyses. I begin with discussing the interpretation of plausible values. I then show that using plausible values, when derived from models including conditioning

Table 1.1 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

variables, entails striking forms of statistical discrimination, and propose that this approach should not be used for sociological analyses.

Finally, in Section 6, I briefly consider models that avoid a reference to latent competencies and instead directly relate the observable response patterns to values of explanatory variables.

The paper ends with a short summary of the conclusions.

Illustrations with NEPS data

To illustrate the discussion, I use NEPS data on math competencies of 5th grade pupils.[1] I use the data file `SC3_xTargetCompetencies_D_1-0-0.sav` that is part of the SPSS version of the SC3 SUF.[2] 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 1.1 shows the distribution of their values.

There are three types of missing values: -97 (refused), -95 (implausible value), and -94 (not reached). I treat all three types equally as missing values. The following table shows the

[1]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.

[2]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).

distribution of missing values (M is the number of missing values, N the number of persons).

M	N	M	N	M	N	M	N	M	N	M	N	M	N
0	2076	3	405	6	142	9	35	12	14	15	7	19	1
1	1131	4	298	7	104	10	32	13	10	16	1	21	1
2	644	5	223	8	50	11	14	14	4	17	2	23	14

In this paper, I treat all missing answers as wrong answers, that is, negative values of response variables are substituted by zeros.[3]

In addition to the test results, I use a binary variable Z (0 for girls, 1 for boys) to illustrate the comparison of competencies between groups of persons. For 5144 of the 5208 pupils who participated in the math test one can find a valid value of Z in the data file `SC3_xTarget_D_1-0-0.sav`. 5130 of these pupils (2649 boys, 2481 girls) have a valid response to at least one test item. This is the number of cases that is used in all illustrations in the present paper: $n = 5130$, $m = 23$.

2. Nonprobabilistic index constructions

In this section, I consider index constructions which do not presuppose a probabilistic model.

2.1 A simple additive index

A simple index is based on counting the number of correct responses. Denoting the variable by S, its value for person i is

$$s_i := \sum_{j=1}^{m} I[x_{ij} = 1] \tag{1}$$

m denotes the number of items, $I[\ldots]$ is the indicator function. An index that uses the proportion of correctly answered items can be defined by $S^* := S/m$. Mean values of S are 13.23 for boys and 11.77 for girls. The distributions are shown in Figure 2.1.

2.2 Regression models

The variable S (or S^*) can immediately be used as the dependent variable in a regression model. The basic idea is to think of the distribution of S as being dependent on values of explanatory variables. Assuming p explanatory variables, say $Z_1, \ldots, Z_p$, one considers the conditional probabilities $\Pr(S = s \mid Z_1 = z_1, \ldots, Z_p = z_p)$.[4] Such models correspond to the second connotation of a sociological view mentioned at the beginning: They are not concerned with the characterization of individuals but are tools for investigating how the distribution of a dependent variable depends on values of other (explanatory) variables.

Actually, most often one ignores the form of the distribution, and regression models only relate to conditional mean values (expectations). Parametric regression models then have the general form

$$\mathrm{E}(S \mid Z_1 = z_1, \ldots, Z_p = z_p) \approx h(z_1, \ldots, z_p; \beta)$$

[3] For further discussion of missing answers in competence tests see Rohwer (2013).

[4] These probabilities are posited by the regression model; the variable S was constructed without a probabilistic framework.

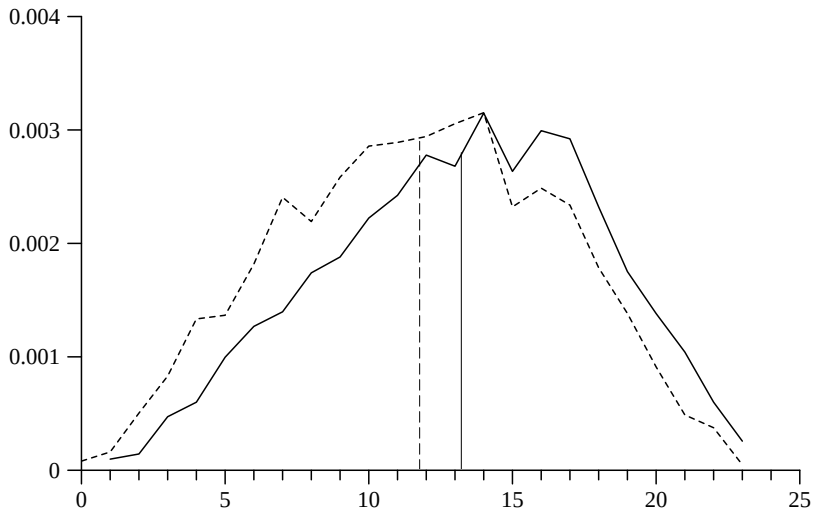

Fig. 2.1 Frequency distributions of S^* (proportion of correct answers) for boys (solid) and girls (dashed).

where h is a parametric function of the explanatory variables that contains a parameter vector β to be estimated from the data. In particular, a simple linear regression model without interaction terms could be written as

$$\mathrm{E}(S \mid Z_1 = z_1, \ldots, Z_p = z_p) \approx \alpha + z_1\beta_1 + \cdots + z_p\beta_p$$

To illustrate, I use the data on math competencies with a single explanatory variable, Z (= 0 for girls and = 1 for boys). For this application, a simple linear model

$$\mathrm{E}(S \mid Z = z) \approx \alpha + z\beta \tag{2}$$

suffices. Using OLS estimation, one gets the parameter values $\hat{\alpha} = 11.77$ and $\hat{\beta} = 1.46$, which correspond to the mean values mentioned in Section 2.1.

2.3 Weighting with item difficulties?

The index S (or S^*) can be criticized with the argument that it takes all items as equally difficult. The argument suggests to think about a more refined index that takes item difficulties into account. Assume that one has available values $d_1, \ldots, d_m$ indicating the difficulty of the test items. One can then define a weighted version of S^*, say S^w, having values

$$s_i^w := \frac{\sum_{j=1}^m I[x_{ij} = 1]\, d_j}{\sum_{j=1}^m d_j} \tag{3}$$

Of course, this index requires a foregoing definition of item difficulties. There are basically two possibilities.

One method is based on the expert knowledge of the persons who created the test items. The argument is simple: These persons are responsible for selecting more or less difficult test items, and therefore should provide indicators of the difficulty of the actually chosen items. In fact, since 'competence' and 'difficulty' are complementary notions, it would be better to speak of 'importance' instead of 'difficulty': Each item should be given a weight indicating the importance of the item as part of the overall competence that the test is intended to measure. Of course, formulations of the test items and proposed values of their importance should be publicly

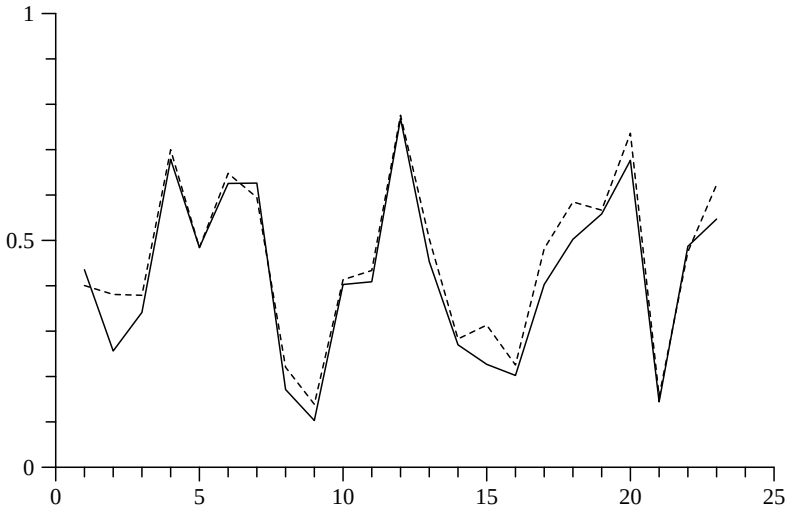

Fig. 2.2 Values of d_j (proportion of not correct answers) for items $j = 1, \ldots, 23$, for boys (solid) and girls (dashed).

accessible, including the participants of the tests. This entails that such values must be fixed in advance.

A quite different method uses observed responses to assess the difficulty of test items. In the present nonprobabilistic framework, one could use the proportion of wrong answers:

$$d_j := \frac{1}{n} \sum_{i=1}^{n} I[x_{ij} = 0] \tag{4}$$

n being the number of respondents. Values for boys and girls are shown in Figure 2.2.

While the definition of item difficulties according to (4) seems sensible, on first sight, it actually creates an essential problem: The approach makes item difficulties dependent on respondents' competencies, and vice versa. To illustrate the problem, consider the following example with four items and five persons:

$$\mathbf{X}' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \qquad \mathbf{X}'' = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

The same test is done at the beginning ($\mathbf{X}'$) and at the end ($\mathbf{X}''$) of a class. Persons 1, 3 and 4 performed equal at both times, persons 2 and 5 performed better at the end of the class. So one would conclude that the competencies of persons 1, 3 and 4 did not change, and the competencies of persons 2 and 5 increased. However, if one uses the definition (4), the increased competencies of persons 2 and 5 lead to a change in item difficulties:

Item difficulty	1	2	3	4
at the beginning	0.2	0.6	0.4	0.6
at the end	0.2	0.4	0.2	0.6

and the values of S^w change as follows:

Person	1	2	3	4	5
at the beginning	0.11	0.44	0.67	0.67	.55
at the end	0.14	0.57	0.57	0.71	.86

While there is no change in the response patterns of persons 1, 3 and 4, the index S^w suggests that the competence of person 3 has decreased and the competencies of persons 1 and 4 have increased.

The example shows that using item difficulties derived from frequencies of wrong (or correct) answers makes the measure of competence in a problematic way dependent on the distribution of competencies in the group of persons that participated in the test (or belong to a reference sample).[5] If one intends an objective measure, it should be defined by a procedure that is independent of the actual competencies of the persons participating in a competence test. This would be the case if item difficulties are defined in advance as a fixed part of the test.

3. Using a probabilistic framework

In this section, I begin with discussing the idea to use a probabilistic framework for the representation of competencies. Corresponding models will be considered in subsequent sections.

Consider the response to a single test item, say X_j. Conceiving of X_j as a random variable, one can think of the probability that X_j takes the value 1, formally: $\Pr(X_j = 1)$. In order to understand this probability, one has to refer to a person whose probability for giving a correct answer to the jth item is $\Pr(X_j = 1)$. Of course, this probability can well be different for different persons, and one therefore should in some way characterize the person to which the probability relates. I use again Z to denote the variable whose value characterizes the referenced person. The basic expression then becomes

$$\Pr(X_j = 1 \mid Z = z) \tag{5}$$

to be understood as the probability that a person characterized by the value z of the variable Z will give a correct response to item j.

Note that in this set-up Z is a variable, and this entails that the probability (5) refers to a *generic* person, that is, a person who is characterized *only* by a value of Z. This corresponds to the second connotation of a sociological view mentioned in the introduction. Of course, Z can consist of several components. For example, $Z = z$ could mean a person being a female 5th grade pupil in a specified type of school.[6]

Starting from (5) not only provides an interpretation of the probabilities for giving correct answers but also gives these probabilities an operational meaning: $\Pr(X_j = 1 \mid Z = z)$ can be estimated by a proportion of persons, in a sample demarcated by $Z = z$, who can answer the jth item correctly.

Now, given this probabilistic set-up, there are basically two different possibilities for the definition of quantitative measures of competence.

a) One can directly use $\Pr(X_j = 1 \mid Z = z)$ as a quantitative measure of the competence (to give correct answers to item j) of a person specified by $Z = z$.

b) One can attempt to construct a latent variable behind the item-specific probabilities in such a way that its values can be interpreted as measures of competence.

In the remainder of this section I briefly consider the first approach which is concerned with

[5] For further discussion of distribution-dependent index constructions see Rohwer & Pötter 2002: 71f.

[6] Note also that in this set-up Z is not a random variable, but is only used to formulate a condition that is required for making the probability intelligible. To think of a distribution of Z would require the reference to a specified sample or population, entailing that it would be a frequency distribution, not a probability distribution; or alternatively, one would need to specify a random mechanism that generates values of Z.

'overt competencies'. Beginning with considering the items separately, one can think of

$$Y_j(z) := \Pr(X_j = 1 \mid Z = z) \tag{6}$$

as a measure of the competence for giving a correct response to item j. When referring to all m items, one can use the vector

$$Y(z) := (Y_1(z), \ldots, Y_m(z)) \tag{7}$$

Note that (6) and (7) are definitions. They do not presuppose, or entail, the independence assumption

$$\Pr(X_1 = x_1, \ldots, X_m = x_m \mid Z = z) \approx \prod_{j=1}^{m} \Pr(X_j = x_j \mid Z = z) \tag{8}$$

In general, this assumption will not be true. Nevertheless, it is well possible to aggregate the components of the vector $Y(z)$ into a single measure of competence. One can use, for example, an index

$$Y^*(z) := \frac{1}{m} \sum_{j=1}^{m} \Pr(X_j = 1 \mid Z = z) \tag{9}$$

which can be interpreted as a mean probability of correct responses. As defined, the index treats all items as equally important for an assessment of competencies. Instead, one could use varying weights representing the importance of the items.

4. Competencies derived from a Rasch model

A widespread probabilistic framework for competencies is the Rasch model. In this section I consider using this model for a two-step procedure: In a first step the model is used to construct a variable representing latent competencies which, in a second step, is used as the dependent variable of a regression model.

4.1 Description of the Rasch model

I interpret the Rasch model as a measurement model providing the second part of a measurement procedure. The first part consists in a method for generating the data that are relevant for the quantity to be measured; in the present context this is the test that generates a person's observed scores. The measurement model then specifies how to use these data for the calculation of the final measure. Given this understanding, the Rasch model aims to construct, for each person i, a value θ_i that can be interpreted as a quantitative measure of the competence that corresponds to her observed test result $(x_{i1}, \ldots, x_{im})$. In the standard Rasch model θ_i is a scalar quantity. The basic idea is to consider the test results as values of random variables, $X_1, \ldots, X_m$, whose joint distribution depends on model parameters, formally:

$$\Pr(X_1 = x_{i1}, \ldots, X_m = x_{im}; \theta_i, \delta) \tag{10}$$

In addition to θ_i, there is a vector of item parameters: $\delta := (\delta_1, \ldots, \delta_m)$.

The Rasch model is a specific version of (10) which results from assuming (a) conditional independence, and (b) a specific parametric form:

$$\Pr(X_1 = x_{i1}, \ldots, X_m = x_{im}; \theta_i, \delta) = \prod_{j=1}^{m} \frac{\exp(\theta_i - \delta_j)^{x_{ij}}}{1 + \exp(\theta_i - \delta_j)} \tag{11}$$

Given this model, the likelihood of data $(x_{i1}, \ldots, x_{im})$, for persons $i = 1, \ldots, n$, is

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

The first-order conditions for maximizing this likelihood entail

$$\sum_{j=1}^{m} \frac{\exp(\theta_i - \delta_j)}{1 + \exp(\theta_i - \delta_j)} = s_i \tag{12}$$

where $s_i = \sum_j x_{ij}$ is the sum score of person i. This equation has two important implications. First, all persons having the same number of correctly answered items (= belong to the same 'score group') get the same value of θ_i. So one can simply refer to a mapping $s \mapsto \theta(s)$ where $\theta(s) = \theta_i$ iff $s = s_i$. Second, one cannot calculate values of θ_i for persons having all answers wrong, or all answers right. This implication can be avoided by using weighted maximum likelihood estimates; this will be discussed in Section 4.2.

In order to find estimates of the model parameters one can proceed in two steps. In a first step, one estimates the item parameters, in a second step one calculates values of θ_i. In order to perform the first step without the need to use values of θ_i, one begins with a conditional likelihood which, for person i, is defined by

$$\mathcal{L}_i^{\text{con}}(\delta) := \Pr(X_1 = x_{i1}, \ldots, X_m = x_{im} \mid S = s_i; \theta_i, \delta)$$

The additional conditioning is done with values of the variable S, the total number of correctly answered items. Now define

$$\mathcal{D}_s := \{x = (x_1, \ldots, x_m) \mid x_j \in \{0, 1\}, \Sigma_j x_j = s\}$$

Each set $\mathcal{D}_s$ contains all response patterns where the total number of correctly answered items is s. This allows one to derive (see, e.g., Rohwer & Pötter 2002: 286)

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

It follows that one can find estimates of $\delta_1, \ldots, \delta_m$ by maximizing the conditional likelihood

$$\mathcal{L}^{\text{con}}(\delta) = \prod_{i=1}^{n} \mathcal{L}_i^{\text{con}}(\delta)$$

without the need to use values of θ_i. Then, having found estimates $\hat{\delta}_j$, one can calculate values of θ_i by solving equation (12).

To illustrate this procedure, I use the NEPS data on math competencies and treat all missing answers as wrong answers. In order to find estimates of the item parameters, I use the constraint $\sum_j \delta_j = 0$. The estimated values are shown in Table 4.1. These values mainly reflect the item difficulties d_j as defined in (4); this is illustrated in Figure 4.1 (the linear correlation is 0.9975).

Competence values are calculated by solving equation (12). Results are shown in Table 4.2. For each score group s, the table shows the corresponding value of $\theta(s)$. The table also shows the number of corresponding response patterns and the number of persons. The total number of persons in the table is 5130. Of these, 5 persons have all items wrong ($s = 0$), and 21 persons have all items right ($s = 23$).

Table 4.1 Estimated item parameters.

j	$\hat{\delta}_j$	j	$\hat{\delta}_j$	j	$\hat{\delta}_j$	j	$\hat{\delta}_j$
1	−0.0746	7	0.9171	13	0.2335	19	0.6635
2	−0.6122	8	−1.4015	14	−0.8605	20	1.4687
3	−0.3819	9	−2.0642	15	−0.8990	21	−1.7771
4	1.3648	10	−0.1276	16	−1.2706	22	0.2460
5	0.2624	11	−0.0591	17	0.0488	23	0.7810
6	1.0604	12	1.9161	18	0.5661		

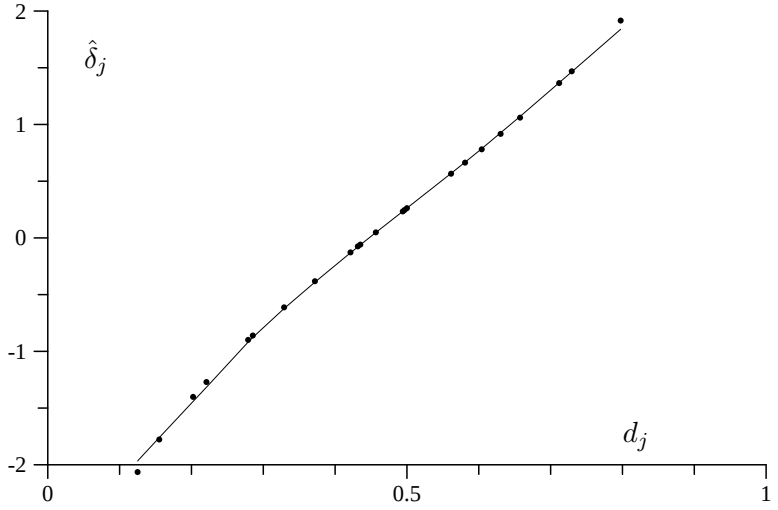

Fig. 4.1 Relationship between d_j as defined in (4) and the item parameters $\hat{\delta}_j$ in Table 4.1.

Table 4.2 Estimated values of $\theta(s)$ for score groups $s = 1, \ldots, 22$.

s	patterns	persons	$\theta(s)$	s	patterns	persons	$\theta(s)$
0	1	5		12	373	376	0.12
1	10	17	−3.57	13	371	376	0.34
2	34	41	−2.77	14	409	415	0.56
3	72	84	−2.27	15	324	328	0.78
4	117	124	−1.88	16	361	363	1.02
5	147	154	−1.55	17	340	349	1.27
6	198	201	−1.26	18	251	273	1.55
7	241	246	−1.00	19	182	208	1.87
8	253	257	−0.76	20	119	153	2.25
9	290	291	−0.53	21	68	103	2.74
10	329	332	−0.31	22	19	65	3.52
11	346	348	−0.09	23	1	21	

4.2 Weighted ML estimation

When using (12), one gets maximum likelihood estimates of values of θ_i. An alternative, so-called weighted maximum likelihood estimates (WLE), was proposed by Warm (1989). Since

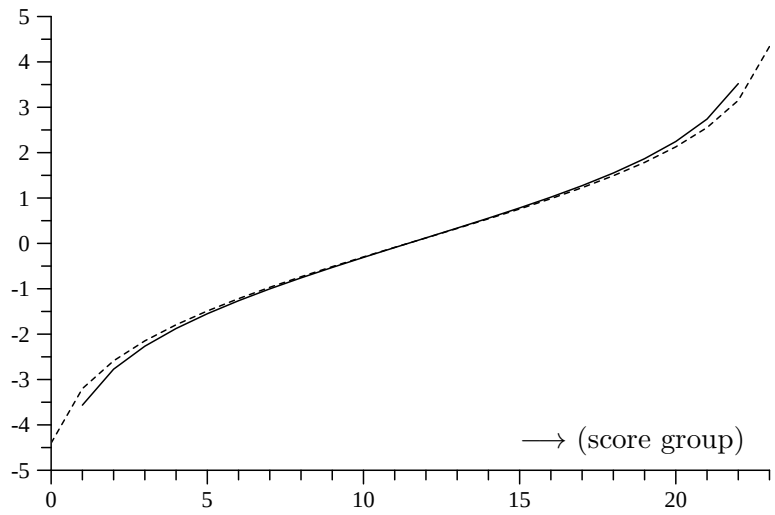

Fig. 4.2 MLE (solid) and WLE (dashed) estimates of $\theta(s)$ for score groups $s = 0, \ldots, 23$.

these are often used,[7] I briefly describe the calculation.

Warm's proposal concerns the second step, after the item parameters have been calculated. For the calculation of values of θ_i, he proposes to use the weighted likelihood function

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

where the weights are defined by

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

Maximizing this likelihood entails the equation

$$\sum_{j=1}^{m} e_{ij} - \frac{\sum_{j=1}^{m} e_{ij} \, (1 - e_{ij}) \, (1 - 2\, e_{ij})}{2 \, \sum_{j=1}^{m} e_{ij} \, (1 - e_{ij})} = \sum_{j=1}^{m} x_{ij} \tag{13}$$

where

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

WLEs are found by solving (13) instead of (12). For our application, Figure 4.2 compares the MLE and WLE estimates of values of $\theta(s)$. One obviously gets very similar values.[8]

4.3 How to construct a dependent variable?

Having estimated a Rasch model, one can define a variable, say C, having values $c_i := \hat{\theta}_i$ which represent the latent competencies. This variable can then be used as a dependent variable in further regression analyses. However, one could also use the variable S which simply records the number of correctly answered items (see Section 2.2). For each value of S (except 0 and m

[7] As explained by Pohl and Carstensen (2012), also the NEPS SUFs contain WLEs.

[8] For further discussion, see Linacre (2009).

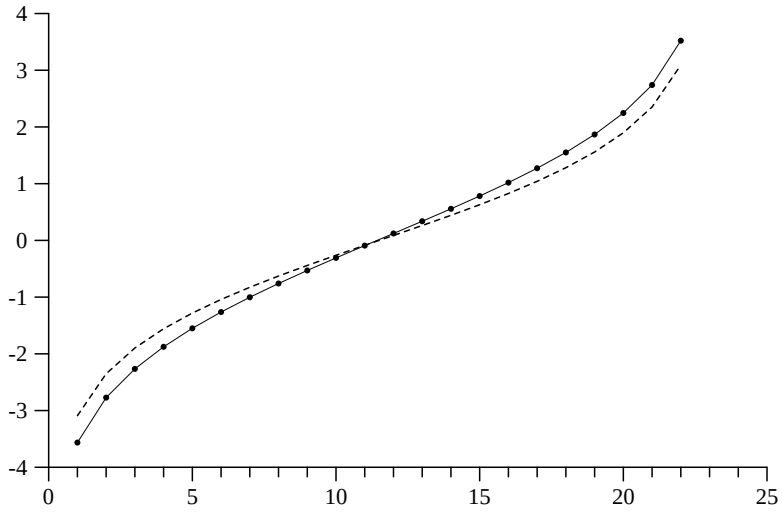

Fig. 4.3 Plot of the function $s \longrightarrow c_s$ with $d = 0$ (dashed line), and $s \longrightarrow C(s)$ (solid, already shown in Figure 4.2).

if not using WLEs) there is exactly one value of C, so one can consider C as a function of S. This is already shown in Figure 4.2.

So the question remains, Why not use S instead of C? Both are quantitative variables, the difference concerns the metric. S simply counts the number of correct answers, C entails a nonlinear transformation. Is there an argument for using C instead of S?

One argument could be that C, in contrast to S, takes item difficulties into account. However, the Rasch model derives the item parameters from frequencies of correct and wrong answers, and as I have shown above, this is an inherently problematic approach. Also note that most of the difference between C and S does not result from differences among item difficulties but from a logit transformation. To illustrate, I use the Rasch model with the constraint that all item parameters are equal, say equal to d. Starting from (12), and using c_s to denote the person parameter in score group s, one finds

$$\frac{\exp(c_s - d)}{1 + \exp(c_s - d)} = \frac{s}{m}$$

and consequently

$$c_s = \log\left(\frac{s}{m - s}\right) + d$$

As shown by Figure 4.3, already without taking into account item difficulties one gets an essentially different metric.

In order to construct a variable that is better comparable with S, one can apply the inverse logit transformation to C, resulting in a variable C^* with values

$$c^* = \frac{\exp(c_s)}{1 + \exp(c_s)}$$

As shown by Figure 4.4, S and C^* are not much different, and both could be used as a dependent variable. Using instead the variable C would require to refer to a not easily understandable metric, entailing that also effects of explanatory variables are difficult to interpret. To illustrate, I use

$$\mathrm{E}(C \mid Z = z) \approx \alpha_c + z\beta_c \tag{14}$$

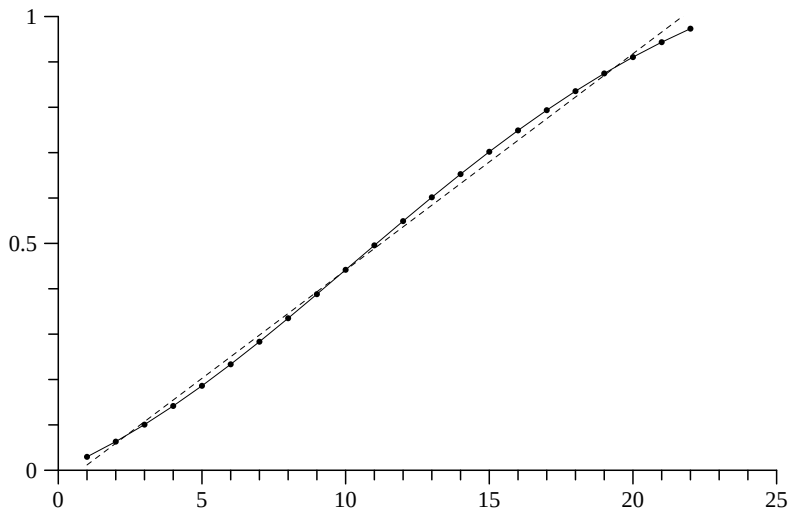

Fig. 4.4 Plot of the function $s \longrightarrow C^*(s)$. The dashed line shows a linear relationship.

which is analogous to (2). The OLS estimates are $\hat{\alpha}_c = 0.0784$ and $\hat{\beta}_c = 0.3495$. How do these figures relate to girls' and boys' competencies for solving the items? One could apply the inverse logit transformation:

$$\frac{\exp(0.0784)}{1+\exp(0.0784)} = 0.52 \quad \text{and} \quad \frac{\exp(0.4278)}{1+\exp(0.4279)} = 0.61$$

These figures are then similar to the proportions of items which girls ($11.77/23 = 0.51$) and boys ($13.23/23 = 0.57$) can solve correctly.

5. Plausible values

In this section I discuss the proposal to use so-called plausible values for the dependent variable of regression analyses. In order to understand this proposal one needs to consider the models to be used for the construction of plausible values. There are two kinds: models with and without additional covariates for conditioning. I begin with models which do not include conditioning variables.

5.1 Models without conditioning variables

As a starting point, I briefly describe the marginal maximum likelihood method. This method was introduced to separate the estimation of item parameters from the calculation of competence values (Bock & Aitkin 1981). Only for Rasch models this can already be achieved with the method of conditional likelihood estimation (see Section 4.1). Here I describe the marginal likelihood method, and subsequently the calculation of plausible values, by referring again to the Rasch model.

The basic idea is to substitute the model parameters which represent the latent competencies by a random variable, say U. Analogous to (11), the model can then be written as

$$\Pr(X_1{=}x_1, \ldots, X_m{=}x_m \mid u; \delta) = \prod_{j=1}^{m} \frac{\exp(x_j\,(u-\delta_j))}{1+\exp(u-\delta_j)} \tag{15}$$

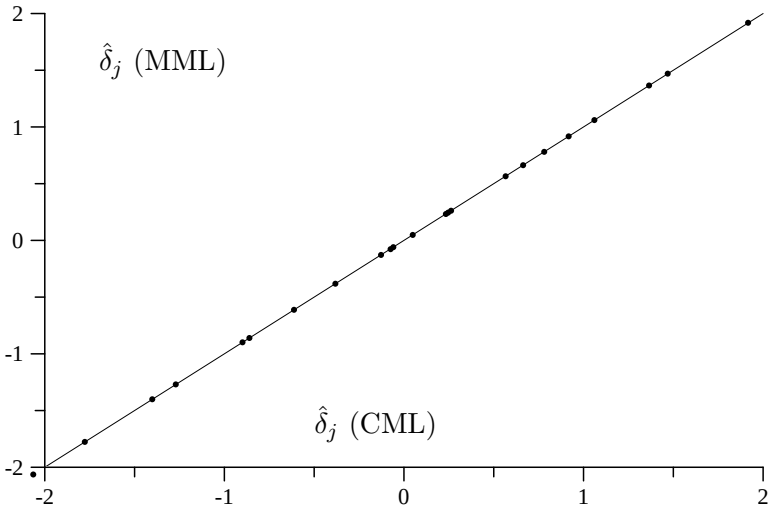

Fig. 5.1 Scatterplot of item parameters estimated with conditional (CML) and marginal (MML) maximum likelihood method, respectively.

where u is a value of the random variable U. If one now assumes a distribution of U, say $f(u)$, one can define a marginal likelihood function

$$\mathcal{L}^m(\delta) := \prod_{i=1}^{n} \int_u \prod_{j=1}^{m} \frac{\exp(x_{ij}\,(u-\delta_j))}{1+\exp(u-\delta_j)}\, f(u)\,\mathrm{d}u \tag{16}$$

By maximizing this function one gets estimates of the item parameters δ_j.

For the Rasch model, these estimates are similar to the estimates produced by conditional likelihood estimation. In fact, using a standard normal distribution for integration, one gets almost identical estimates for the NEPS data, see Figure 5.1.[9]

Having estimated item parameters, there are different possibilities for the definition of latent competencies. For example, one can use again equation (12) to calculate for each person a value of her latent competence. The proposal to represent latent competencies by plausible values follows a different approach that is based on the distribution of U conditional on observed test results.

It is assumed that one can refer to a joint distribution of U and the random variables $X_1, \ldots, X_m$ by a probability/density function

$$g(x_1, \ldots, x_m, u; \delta) := \Pr(X_1 = x_1, \ldots, X_m = x_m \mid u; \delta)\, f(u) \tag{17}$$

One can then consider the distribution of U conditional on observed test results:

$$g(u \mid x_1, \ldots, x_m; \delta) = \frac{g(x_1, \ldots, x_m, u; \delta)}{\int_u g(x_1, \ldots, x_m, u; \delta)\,\mathrm{d}u} \tag{18}$$

Using the parametric form assumed in (15), this becomes

$$g(u \mid x_1, \ldots, x_m; \delta) = \frac{\prod_j \frac{\exp(x_j\,(u-\delta_j))}{1+\exp(u-\delta_j)}\, f(u)}{\int_u \prod_j \frac{\exp(x_j\,(u-\delta_j))}{1+\exp(u-\delta_j)}\, f(u)\,\mathrm{d}u} \tag{19}$$

[9] For comparison, the item parameters estimated with (16) have been transformed to get the same mean and standard deviation as the parameters in Table 4.1.

In both the numerator and the denominator one can factor out the term $\prod_j \exp(-x_j\,\delta_j)$, and so one can write

$$g(u \mid x_1, \ldots, x_m; \delta) = \tag{20}$$

$$\frac{\prod_j \dfrac{\exp(x_j\,u)}{1+\exp(u-\delta_j)}\,f(u)}{\displaystyle\int_u \prod_j \frac{\exp(x_j\,u)}{1+\exp(u-\delta_j)}\,f(u)\,\mathrm{d}u} = \frac{\dfrac{\exp(s\,u)}{\prod_j 1+\exp(u-\delta_j)}\,f(u)}{\displaystyle\int_u \frac{\exp(s\,u)}{\prod_j 1+\exp(u-\delta_j)}\,f(u)\,\mathrm{d}u}$$

where $s = \sum_j x_j$ is the total number of correctly answered items. This shows that one gets the same conditional distribution for all response patterns that belong to the same score group. It therefore suffices to consider the conditional distributions $g(u|s;\delta)$ where s is a value of the variable S (the total number of correctly answered items).

For each person i who participated in the test there is a sum score s_i and one can therefore think of a conditional distribution of U that has the density $g(u|s_i;\delta)$. Plausible values are random draws from these conditional distributions (see, e.g., Mislevy et al. 1992: 138f). Of course, in order to actually calculate such values one needs an assumption about the parametric form of the unconditional distribution of U (denoted above by $f(u)$). Often used is a normal distribution. One can then calculate K sets of plausible values, say

$$\{p_i^{(k)} \mid i = 1, \ldots, n\}$$

for $k = 1, \ldots, K$. These plausible values can be used to estimate characteristics of the distribution of U conditional on the test results of all sampled persons.

Plausible values can also be used for further regression analyses. This can simply be done by defining dependent variables $C^{(k)}$ having values $p_i^{(k)}$. One then performs K regression analyses and finally uses mean values of the calculated parameters. As an alternative, one can use mean values of the distributions $g(u|s_i;\delta)$, often called 'expected a posteriori estimators (EAPs)'. One then defines just one dependent variable, say C^e, having values

$$c_i^e := \mathrm{E}(U \mid s_i) = \int_u u\,g(u \mid s_i;\delta)\,\mathrm{d}u \tag{21}$$

Which of the two methods should be preferred? An argument for preferring plausible values concerns the variance of the distribution of latent competencies. Using EAPs instead of plausible values tends to underestimate this variance.[10]

A more important question concerns whether one should use plausible values (or EAPs) instead of competence scores which can directly be derived from numbers of correct responses (as discussed in previous sections). The main argument seems to be that plausible values should be used when the tests for generating competence data do not allow 'sufficiently reliable' measuring of individual competencies (see, e.g., Mislevy, Johnson & Muraki 1992: 137f). This could be the case in large-scale assessments which employ the method of matrix-sampling of items, and each person is administered only a very small number of items. It is proposed, then, that one should no longer attempt to measure individual competencies but directly estimate a presupposed population distribution of latent competencies.

But how might it be possible to estimate the distribution of a quantity which is not derived from individual values of that quantity? As argued by Mislevy, Johnson and Muraki (1992: 138), this becomes possible by considering competencies as 'missing values', in fact, values which are

[10] This has also been shown in simulation studies, see, e.g., OECD (2009: 98), von Davier, Gonzalez & Mislevy (2009).

not 'observed', or 'measured', *for all* persons (see also Mislevy 1991). The argument is that, instead of intending to measure competencies, one can construct a model for the distribution of the completely missing values, and then estimate that model with whatever information is available.

Following this reasoning, it seems that the basic question concerns our understanding of the NEPS competence data: Should we assume that these data do not provide 'sufficiently reliable' information about individual competencies and therefore use the method of plausible values? The alternative is easily misleading, however. It is surely important to think about measurement errors and how they can be taken into account. But this is not done, at least not explicitly, by the method of plausible values. While it is sometimes suggested to think of the distributions $g(u|s;\delta)$ as representing measurement errors (e.g., Wu 2005, OECD 2009: 96), there is no conceptual foundation for this interpretation.[11] Whether plausible values are useful for coping with measurement errors should therefore be considered as an open question.

5.2 Conditioning variables

As described in the previous section, the leading idea of the method of plausible values is to consider competencies as completely missing values and to estimate distributions of these values based on whatever information is available. The approach suggests to use not only information from competence tests but also from any number of further variables which might be correlated with competencies.

Following this idea, one does not start from just one distribution of latent competencies, but from distributions conditional on these additional covariates (often then called 'conditioning variables'). Let Z denote a vector of such covariates. The latent competence of a person i is then assumed to be $h(z_i;\beta) + u_i$ where $h(z_i;\beta)$ is a parametric function of the person's value of Z.

As before, one assumes a joint distribution of $X_1, \ldots, X_m$ and U, now conditional on values of Z. It is assumed that the distribution of U is independent of Z (and δ), so that one can write

$$g(x_1, \ldots, x_m, u \,|\, z;\delta) = \Pr(X_1 = x_1, \ldots, X_m = x_m \mid u, z;\delta)\, f(u) \tag{22}$$

Instead of (15), the model for the response patterns becomes

$$\Pr(X_1 = x_1, \ldots, X_m = x_m \mid u, z;\delta) = \prod_{j=1}^{m} \frac{\exp(x_j\,(h(z;\beta) + u - \delta_j))}{1 + \exp(h(z;\beta) + u - \delta_j)} \tag{23}$$

and completely analogous to the derivation of (20) one gets

$$g(u \mid s, z;\delta) = \frac{\dfrac{\exp(s\,(h(z;\beta) + u))}{\prod_j 1 + \exp(h(z;\beta) + u - \delta_j)}\, f(u)}{\displaystyle\int_u \frac{\exp(s\,(h(z;\beta) + u))}{\prod_j 1 + \exp(h(z;\beta) + u - \delta_j)}\, f(u)\,\mathrm{d}u} \tag{24}$$

For each person i there is now a distribution of her latent competencies, described by the density $g(u - h(z_i;\beta) \,|\, s_i, z_i;\delta)$, which depends on both her test result and her values of the conditioning variables.

To illustrate this approach with the NEPS data, I use $h(z;\beta) = z\beta$ with $z = 0$ for girls and

[11] Consider, for example, the following statement: Plausible values "are Monte Carlo draws from posterior proficiency distributions for each individual, and hence incorporate all sources of uncertainty (including measurement error)." Schofield et al. 2013: 3.

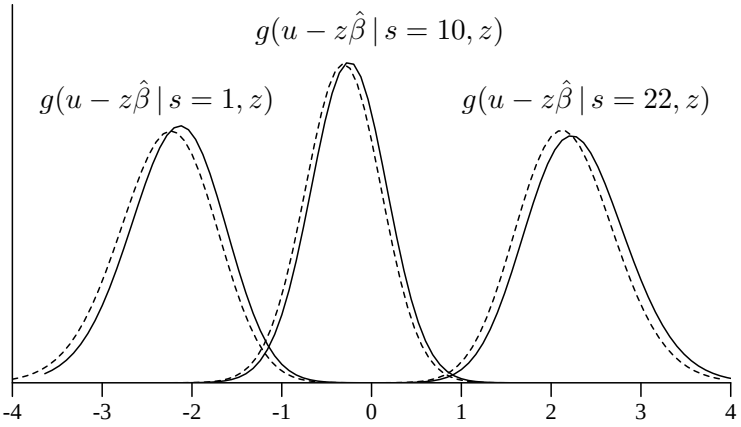

Fig. 5.1 Conditional distributions of latent competencies for boys (solid) and girls (dashed), calculated according to (24).

Table 5.1 Estimated latent competence scores for score groups $s = 0, \ldots, 23$.

	number of		latent scores			number of		latent scores	
s	boys	girls	boys	girls	s	boys	girls	boys	girls
0	0	5	−2.46	−2.59	12	195	181	0.11	0.04
1	7	10	−2.18	−2.29	13	188	188	0.29	0.22
2	10	31	−1.91	−2.01	14	221	194	0.47	0.40
3	33	51	−1.66	−1.75	15	185	143	0.66	0.59
4	42	82	−1.43	−1.51	16	210	153	0.85	0.78
5	70	84	−1.21	−1.29	17	205	144	1.05	0.98
6	89	112	−1.01	−1.08	18	163	110	1.26	1.18
7	98	148	−0.81	−0.88	19	123	85	1.49	1.40
8	122	135	−0.62	−0.69	20	97	56	1.73	1.64
9	132	159	−0.44	−0.50	21	73	30	2.00	1.90
10	156	176	−0.25	−0.32	22	42	23	2.29	2.18
11	170	178	−0.07	−0.14	23	18	3	2.62	2.49

$z = 1$ for boys. Figure 5.1 shows the conditional distributions, separately for girls and boys, for three score groups.[12]

Analogous to (21), one can define a mean value (EAP)

$$c_i := \mathrm{E}(U \mid s_i, z_i) = h(z_i; \beta) + \int_u u\, g(u \mid s_i, z_i; \delta)\, \mathrm{d}u \tag{25}$$

as an estimate of a person's latent competence. These values depend on score groups and covariates. Estimates for boys and girls are shown in Table 5.1.

The table reveals a striking form of statistical discrimination: In each score group girls get lower latent competence values than boys (already seen in Figure 5.1). In fact, even with exactly the same response pattern, girls get a lower latent competence than boys. So one should conclude that this approach does not lead to reasonable measures of individual competencies.

Possibly, this could be ignored if one does not intend to measure individual competencies. However, if derived from a model that contains conditioning variables, plausible values entail the

[12] Model parameters have been estimated by maximizing a marginal likelihood which is calculated with numerical integration.

same discrimination. Plausible values of a person i are then random draws from the distributions $g(u-h(z_i;\beta)\,|\,s_i,z_i;\delta)$ as illustrated in Figure 5.1. I therefore conclude that, for subsequent regression analyses, plausible values, or EAPs, should only be used if they are not derived from models containing conditioning variables.

The interest in regression analyses suggests a further argument for not using conditioning variables. In particular in sociological applications, such analyses are often intended to contribute to explanations. In the present context, one is interested in variables on which the development of competencies depends. When using regression models for studying such dependencies, a minimal condition for acceptable explanations is that the dependent variable can be defined independently of the explanatory variables. Quantitative measures of competence which are to be used as values of a dependent variable should therefore be derived only from observable test scores.

6. Models without latent competencies

So far I have considered two-step procedures: In a first step one constructs a variable representing competencies of individual persons, in a second step this is used as the dependent variable in regression models. Another approach is to use regression models which directly make probabilities of observed test results dependent on explanatory variables. In this section, I briefly consider this approach.

6.1 A model assuming conditional independence

A simple regression model for the response patterns assumes conditional independence and can be written as

$$\Pr(X_1=x_1,\ldots,X_m=x_m\mid Z=z;\beta,\delta)\approx\prod_{j=1}^{m}\frac{\exp(h(z;\beta)-\delta_j)^{x_j}}{1+\exp(h(z;\beta)-\delta_j)} \tag{26}$$

It is assumed that the probability on the left-hand side depends on item parameters $\delta:=(\delta_1,\ldots,\delta_m)$, and on explanatory variables. The dependence on explanatory variables is specified by a parametric function $h(z;\beta)$ where z is the value of a covariate Z, possibly consisting of several components, and β is a parameter vector.

An obvious advantage is that model (26) does not require to think in terms of latent competencies. On the other hand, there is no reference to any summary measure of competence at all, and so it is again difficult to interpret effects of explanatory variables.

To illustrate, I use model (26) for comparing the math competencies of boys and girls: $h(z;\beta)=z\,\beta$, where $z=0$ for girls and $z=1$ for boys. Treating missing values as wrong answers, and using maximum likelihood estimation, one gets $\hat\beta=0.297$ (std.err. 0.013). How is this value to be interpreted?

In contrast to regression models with a quantitative dependent variable representing competencies (e.g. S or C), model (26) requires to think in terms of probabilities of correct answers. The model allows one to compare girls' and boys' probabilities for all possible response patterns. These are, however, very small numbers. As an alternative, one can use mean values, for example the index $Y^*(z)$ defined in (9), based on item-specific probabilities

$$\Pr(X_j=1\mid Z=z)\approx\frac{\exp(z\hat\beta-\hat\delta_j)}{1+\exp(z\hat\beta-\hat\delta_j)}$$

(see Table 6.1). Based on formula (9), estimated mean probabilities of correct answers are 0.51 for girls and 0.58 for boys (actually almost equal to the observed proportions).

Table 6.1 Observed frequencies and estimated probabilities of correct answers.

	observed		model 26		model 27	
Item	boys	girls	boys	girls	boys	girls
1	0.56	0.57	0.60	0.53	0.60	0.53
2	0.74	0.59	0.70	0.64	0.70	0.64
3	0.66	0.60	0.66	0.59	0.66	0.59
4	0.32	0.25	0.32	0.26	0.32	0.25
5	0.52	0.48	0.54	0.46	0.54	0.46
6	0.37	0.31	0.37	0.31	0.37	0.31
7	0.37	0.37	0.40	0.33	0.40	0.33
8	0.83	0.76	0.82	0.77	0.82	0.77
9	0.90	0.85	0.89	0.86	0.89	0.86
10	0.60	0.56	0.61	0.54	0.61	0.54
11	0.59	0.54	0.60	0.53	0.60	0.53
12	0.23	0.17	0.23	0.18	0.23	0.18
13	0.55	0.46	0.54	0.47	0.54	0.47
14	0.73	0.70	0.74	0.68	0.75	0.68
15	0.77	0.67	0.75	0.69	0.75	0.69
16	0.80	0.76	0.80	0.75	0.81	0.75
17	0.60	0.49	0.58	0.51	0.58	0.50
18	0.50	0.38	0.47	0.40	0.47	0.40
19	0.44	0.40	0.45	0.38	0.45	0.38
20	0.32	0.21	0.30	0.24	0.30	0.24
21	0.86	0.83	0.86	0.82	0.87	0.82
22	0.51	0.49	0.54	0.47	0.54	0.46
23	0.45	0.33	0.43	0.36	0.43	0.36

6.2 A version with unobserved heterogeneity

For comparison with the model discussed in Section 5.2, one can consider to specify the dependence on covariates by a function $h(z;\beta)+u$ where u is the value of a further random variable, U, representing unobserved heterogeneity. The model then becomes

$$\Pr(X_1=x_1,\ldots,X_m=x_m \mid Z=z, U=u;\beta,\delta) \approx \prod_{j=1}^{m} \frac{\exp(h(z;\beta)+u-\delta_j)^{x_j}}{1+\exp(h(z;\beta)+u-\delta_j)} \tag{27}$$

Assuming for the distribution of U a parametric density, say $f(u;\phi)$, independent of Z, one can use a marginal likelihood

$$\mathcal{L}^m(\delta,\beta,\phi) := \prod_{i=1}^{n} \int_u \prod_{j=1}^{m} \frac{\exp(h(z_i;\beta)+u-\delta_j)^{x_{ij}}}{1+\exp(h(z_i;\beta)+u-\delta_j)}\, f(u;\phi)\,\mathrm{d}u \tag{28}$$

for estimating the model parameters. The contribution of person i equals the expectation, taken w.r.t. the distribution $f(u;\phi)$, of the probability of the person's response pattern.

To illustrate, I use the same data as in Section 4 and assume a standard normal distribution for U. Treating missing values as wrong answers, the estimated item parameters are almost equal to those resulting from model (26), see Figure 6.1.

The estimated value of β is 0.36 (std.err. 0.029). For interpretation, one has to refer to the question how the probabilities of giving correct answers depend on values of Z. The quantities of interest are then $\Pr(X_j=1 \mid Z=z)$ to be aggregated in some way (see Section 6.1). However,

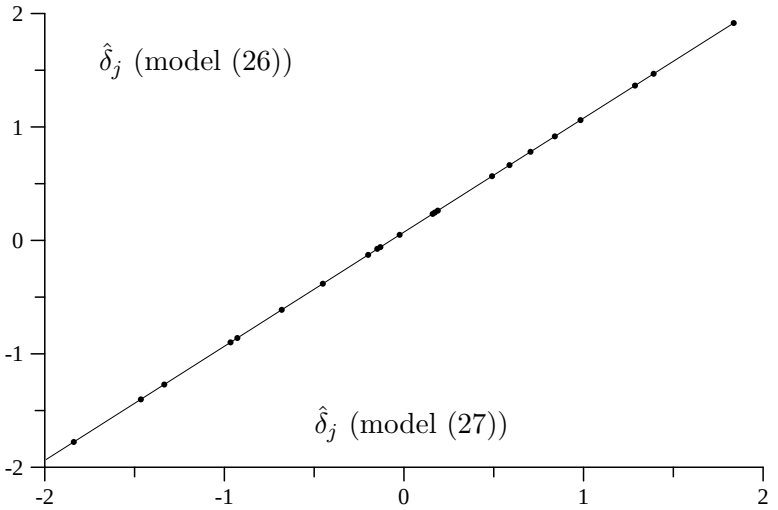

Fig. 6.1 Scatterplot of item parameters estimated with models (26) and (27), respectively.

initially the model only provides

$$\Pr(X_j=1 \mid Z=z, U=u) \approx \frac{\exp(z\hat{\beta}+u-\hat{\delta}_j)}{1+\exp(z\hat{\beta}+u-\hat{\delta}_j)}$$

So one needs a distribution of U for calculating a mean probability. Should one use the unconditional or a conditional distribution? In contrast to the model discussed in Section 5.2, in the present model U is not considered as providing information about latent competencies; and there is, therefore, no reason for using conditional distributions of U. Instead, U is interpreted as representing unobserved heterogeneity in the sense of omitted explanatory variables. One can therefore simply use the distribution $f(u;\phi)$ that served to set up the model, and since this distribution was assumed to be independent of Z, one can use

$$\Pr(X_j=1 \mid Z=z) \approx \int_u \frac{\exp(z\hat{\beta}+u-\hat{\delta}_j)}{1+\exp(z\hat{\beta}+u-\hat{\delta}_j)}\, f(u \mid \hat{\phi})\, \mathrm{d}u$$

to estimate the mean probability of a correct answer to item j. For the example with a standard normal density, such estimates are shown in Table 6.1 in the columns labelled 'model 27'. They are almost identical with those resulting from model (26) that was used in Section 6.1.

7. Conclusion

In this paper, I consider several methods to define dependent variables representing results of competence tests and illustrate these methods with NEPS data on math competencies of 5th grade pupils. The simplest and easily comprehensible method is to use the number of correct responses as values of a quantitative dependent variable in a regression model. Instead of simply using the number of correct responses one can define weighted versions which could take into account that items might have different importance for the competence that the test is intended to measure. However, it is easily misleading to think of such weights as 'item difficulties' which can be derived from proportions of wrong responses.

Instead of using these simple approaches to the construction of a dependent variable, one can start from a probabilistic framework. As an example, I consider the Rasch model that can be

used to construct a variable representing latent competencies which, in a second step, can be used as a dependent variable in regression models. I argue that this approach has two disadvantages, compared with using a simple summary index. The Rasch model introduces a nonlinear metric which is difficult to understand and therefore makes it difficult to interpret effects of explanatory variables. Moreover, the Rasch model employs a notion of 'item difficulties' which are derived from the distribution of competencies of the persons participating in the test.

I then discuss the proposal to use so-called plausible values for the construction of dependent variables of further regression analyses. I distinguish between versions with and without further conditioning variables. I show that using plausible values, when derived from models including conditioning variables, entails striking forms of statistical discrimination, and propose that this approach should not be used for sociological analyses.

Finally, I briefly consider models which avoid a reference to latent competencies and instead directly relate the observable response patterns to values of explanatory variables. While attractive at first sight, this approach has the drawback that such models must be supplemented by a procedure for aggregating item-specific probabilities.

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.

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

Davier, M. von, Gonzalez, E., Mislevy, R. J. (2009). What are Plausible Values and Why are they Useful? *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments*. Vol. 2, 9–36.

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.

Linacre, J. M. (2009). The Efficacy of Warm's MLE Bias Correction. *Rasch Measurement Transactions* 23, 1188-89.

Mislevy, R. J. (1991). Randomization-Based Inference About Latent Variables from Complex Surveys. *Psychometrika* 56, 177–196.

Mislevy, R. J., Johnson, E. G., Muraki, E. (1992). Scaling Procedures in NAEP. *Journal of Educational Statistics* 17, 131–154.

OECD (2009). *PISA Data Analysis Manual: SPSS*. 2nd ed.

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.

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

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

Schofield, L. S., Junker, B., Taylor, L. J., Black, D. A. (2013). Predictive Inference Using Latent Variables with Covariates.

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.

Wu, M. (2005). The Role of Plausible Values in Large-Scale Surveys. *Studies in Educational Evaluation* 31, 114–128.