NEPS WORKING PAPERS

Götz Rohwer

COMPETENCE DISTRIBUTIONS, LATENT REGRESSION MODELS AND PLAUSIBLE VALUES

NEPS Working Paper No. 55
Bamberg, April 2015

LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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@lifbi.de

Competence Distributions, Latent Regression Models and Plausible Values

Götz Rohwer, Ruhr-Universität Bochum

April 2015

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

Bibliographic data:

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

Competence Distributions, Latent Regression Models and Plausible Values

Abstract

The paper discusses constructions of competence distributions presupposing a Rasch model. Constructions, as well as estimation procedures, differ depending on whether competence distributions refer to either observed sum scores or some latent competencies. It is shown that using posterior distributions for estimating distributions of latent competencies often leads to misleading results. The same problems occur when 'plausible values' are used to approximate posterior distributions. Some of these problems can be avoided with latent regression models which focus on mean values of latent competencies. These models, however, entail that the notion of competence becomes dependent on conditioning variables and therefore are in conflict with the idea of measuring competencies.

Keywords
Competence distributions, latent regression, plausible values, conditioning variables

Contents

The primary goal of large-scale assessments of competencies is to describe and compare distributions of competencies in and between groups of persons. This paper discusses some methods which have been proposed for the construction of competence distributions. I refer to a competence test, T_m, consisting of m binary items. The items are represented by variables, $X_1, \ldots, X_m$, having values 1 (if there is a correct answer) or 0 (otherwise). Values of these variables for the members of a population (or sample) $\mathcal{G}$ are given by vectors $x_i := (x_{i1}, \ldots, x_{im})$; i identifies members of $\mathcal{G}$. I also assume a variable G whose values can be used to demarcate subsets of $\mathcal{G}$. As a theoretical framework, I use a Rasch model

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

where $\delta = (\delta_1, \ldots, \delta_m)$ is a vector of item parameters, and u denotes values of a latent variable, U, often called 'latent competencies'.

In Section 1, I discuss how this model can be used to think of measures of competencies, as empirically assessed by the test T_m. In Section 2, I consider the estimation of distributions of sum scores (number of correctly answered items); and in Section 3, I consider two approaches to the estimation of distributions of latent competencies which are defined by a reference to the variable U in model (1). In Section 4, I discuss regression models which consider mean values of latent competencies as being dependent on covariates ('latent regression'). In Section 5, I discuss the proposal to use so-called 'plausible values' as a technical means for the estimation of latent competence distributions and regression models. I show that this method leads into several difficulties, both practical and theoretical. The paper ends with a brief conclusion.

1. Conceptualizations of Competencies

A straightforward approach to the conceptualization of competencies relates to their actual realization in the situation in which the test is performed (e.g. Holland, 1990). For each person i, they show up in the response vector x_i which can be summarized by the score sum

$$s_i^* := \sum_{j=1}^{m} x_{ij} \tag{2}$$

These are observed values resulting from a single application of the test T_m. They can be considered as values of a variable, S^*, having a frequency distribution: $\mathrm{P}(S^* = s)$, that is, the observed proportion of members of $\mathcal{G}$ having the realized sum score $s_i^* = s$.

A different approach is based on interpreting model (1) as postulating a probabilistic relationship between a latent variable, U, and observable test results, $X_1, \ldots, X_m$. The model then entails, for each value u of U, a distribution of *possible* test results (resulting from hypothetical repetitions of the test 'under the same conditions'). In this view, the variables $X_1, \ldots, X_m$ are random variables; and consequently, one can consider also the observable sum scores as values of a random variable

$$S := \sum_{j=1}^{m} X_j \tag{3}$$

Following this probabilistic interpretation, model (1) entails that, conditional on values of U, S has a generalized binomial distribution (I use Pr() for probabilities and P() for frequencies):

$$\Pr(S = s \,|\, U = u) = \sum_{x \in D_s} \prod_{j=1}^{m} \Pr(X_j = x_j \,|\, U = u) \tag{4}$$

for $s = 0, \ldots, m$; D_s is the set of all response patterns $x = (x_1, \ldots, x_m)$ where $\Sigma_j x_j = s$. The conditional mean value of S is simply

$$\mathrm{E}(S \,|\, U = u) = \sum_{j=1}^{m} \Pr(X_j = 1 \,|\, U = u) = \sum_{j=1}^{m} \frac{\exp(u - \delta_j)}{1 + \exp(u - \delta_j)} \tag{5}$$

This equation motivates the following definition:

> The latent competence u_i of a person i is *defined* by $\mathrm{E}(S \,|\, U = u_i)$, that is, the person's mean sum score in a sequence of hypothetical repetitions of the test T_m.

The definition entails that only observed test results are to be used for an assessment of competencies. Conditional on values of U, no further variables should play a role in the measurement of competencies; in particular,

$$\Pr(S = s \,|\, U = u, G = g) = \Pr(S = s \,|\, U = u) \tag{6}$$

However, even when accepting this definition of latent competencies, there is no immediate answer to the question of how to think of a distribution of competencies in the population $\mathcal{G}$. I consider two possibilities.

(A) One can be interested in the distribution of U in the population $\mathcal{G}$, that is, $\mathrm{P}(U = u)$. (Equivalently, one can refer to the conditional expectations, $\mathrm{E}(S \,|\, U = u)$, which are related to values of U by a nonlinear scale transformation.) Since U is a latent variable, there is no immediate way to estimate its distribution. In Section 3, I consider two approaches.

(B) One starts from the idea that a person's competence is given by a distribution of possible sum scores, conditional on the person's latent competence u_i. The distribution of competencies in a population $\mathcal{G}$ is then to be described as an unconditional distribution of S which is derived from postulating a fixed distribution of latent competencies in $\mathcal{G}$, say $\mathrm{P}(U = u)$. Note that this is a frequency distribution (with a finite support): $\mathrm{P}(U = u)$ is the proportion of members of $\mathcal{G}$ whose latent competence equals u. A reference to this distribution allows one to define the unconditional distribution

$$\Pr(S = s) = \sum\nolimits_u \Pr(S = s \,|\, U = u)\, \mathrm{P}(U = u) \tag{7}$$

which describes the distribution of competencies of the members of $\mathcal{G}$.

2. Distributions of Possible Sum Scores

I begin with the second approach (B). In this case, one intends to estimate the distribution of the random variable S, as defined by (3), in the population $\mathcal{G}$. The definition of this distribution presupposes a fixed distribution of U and relates to the test results (sum scores) of randomly drawn members of $\mathcal{G}$. However, since the distribution of U is fixed, and repetitions are postulated to be independent (not influenced by previous outcomes), persons with the same latent competence are exchangeable. Therefore, given a sufficiently large sample, it suffices to know the result of a single test for each person; in other words, one can use $\mathrm{P}(S^* = s)$ to estimate $\Pr(S = s)$.

For the illustration of this and subsequent arguments, I use model (1) to create artificial data. I assume $m = 15$ items having parameters $\delta_j := 0.5\, j - 4$. The population, $\mathcal{G}$, consists of two parts: $\mathcal{G}_0$ has $n_0 = 7000$ and $\mathcal{G}_1$ has $n_1 = 3000$ members. In order to distinguish the two

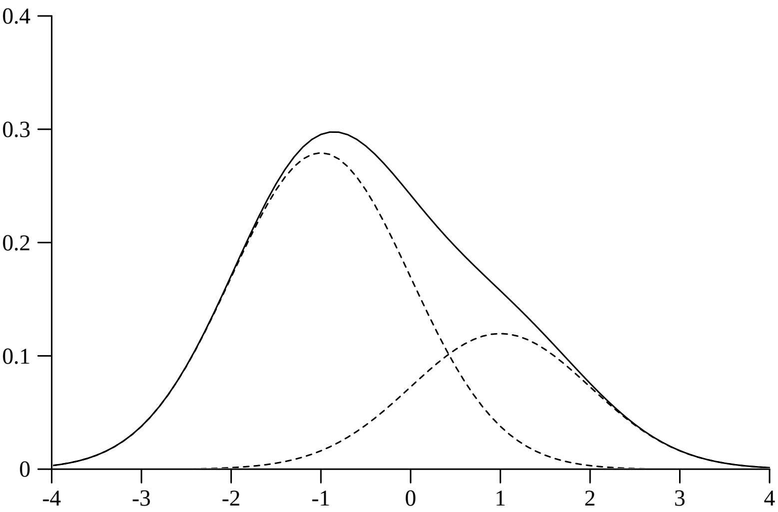

Fig. 2.1 Solid: the distribution $f(u)$ as defined in (8); dashed: distributions $f_0(u)$ and $f_1(u)$.

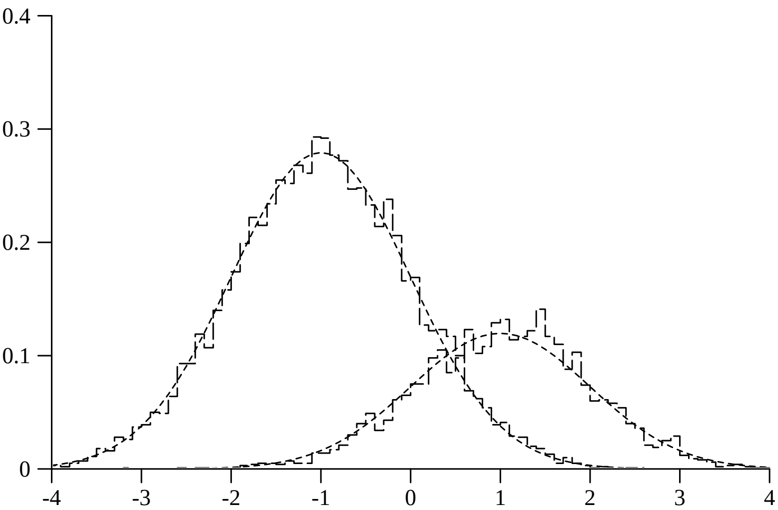

Fig. 2.2 Dashed: distributions $f_0(u)$ and $f_1(u)$; solid: histograms of the artificially generated values; mean values are -0.98 and 1.03, respectively.

subpopulations I use a variable, G, having values 0 and 1, respectively. Distributions of U are given by $f_0(u) = \phi(u+1)$ for $\mathcal{G}_0$ and by $f_1(u) = \phi(u-1)$ for $\mathcal{G}_1$ where ϕ denotes a standard normal density function. The distribution of U in $\mathcal{G}$ is

$$f(u) = 0.7\, f_0(u) + 0.3\, f_1(u) = 0.7\, \phi(u+1) + 0.3\, \phi(u-1) \tag{8}$$

For each person $i \in \mathcal{G}_g$ $(g = 0, 1)$, the distribution $f_g(u)$ is used to randomly generate a value u_i (see Figures 2.1 and 2.2). In $\mathcal{G}$, the mean value of the u_i values is -0.378, the standard deviation is 1.354. Finally, values of the variables X_j are created as follows. For each person i and each item j, one first draws a random number, r_{ij}, uniformly distributed in the interval $[0, 1]$; then

$$x_{ij} := \begin{cases} 1 & \text{if } r_{ij} \leq \frac{\exp(u_i - \delta_j)}{1+\exp(u_i - \delta_j)} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Using these data, Figure 2.3 illustrates that a single test result for each member of $\mathcal{G}$ suffices to estimate the distribution of S.

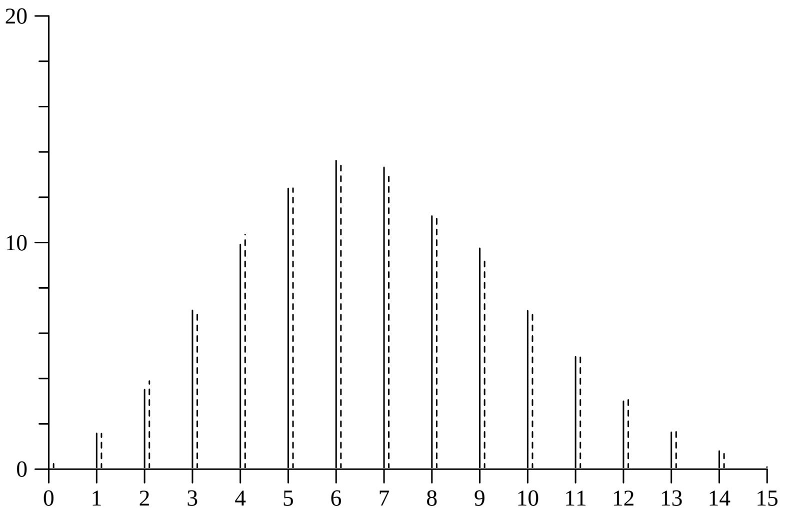

Fig. 2.3 Distributions of S, based on a single outcome (solid) or on the outcomes of 10 repetitions (dashed) for the persons in $\mathcal{G}$. Ordinate: percent.

The same argument applies when estimating distributions of S in subpopulations. The distribution of S in the subpopulation $\mathcal{G}_g$ is defined by

$$\Pr(S\!=\!s\,|\,G\!=\!g) = \sum\nolimits_u \Pr(S\!=\!s\,|\,G\!=\!g, U\!=u)\,\mathrm{P}(U\!=u\,|\,G\!=\!g) \tag{10}$$

Because of (6),

$$\Pr(S\!=\!s\,|\,G\!=\!g) = \sum\nolimits_u \Pr(S\!=\!s\,|\,U\!=u)\,\mathrm{P}(U\!=u\,|\,G\!=\!g) \tag{11}$$

and these conditional probabilities can be estimated by the frequencies $\mathrm{P}(S^*\!=\!s\,|\,G\!=\!g)$.

3. Distributions of Latent Competencies

In this section, I consider two approaches to the estimation of distributions of U in the population $\mathcal{G}$ and in subpopulations $\mathcal{G}_g$.

3.1 A Simple Estimation Approach

I begin with a simple method which is based on equation (5). For each person i in the population, a corresponding equation is

$$\mathrm{E}(S\,|\,U\!=u_i) = \sum_{j=1}^{m} \frac{\exp(u_i - \delta_j)}{1+\exp(u_i-\delta_j)} \tag{12}$$

Using s_i^* as an estimate of person i's expectation of S, suggests to consider $\hat{u}_i$, as defined by

$$\sum_{j=1}^{m} \frac{\exp(\hat{u}_i - \hat{\delta}_j)}{1+\exp(\hat{u}_i-\hat{\delta}_j)} = s_i^* \tag{13}$$

as an estimate of u_i.[1] The values $\hat{\delta}_j$ are estimates of the item parameters of model (1). Given such values, one can calculate a latent competence estimate corresponding to sum scores $s = 1, \ldots, m-1$.

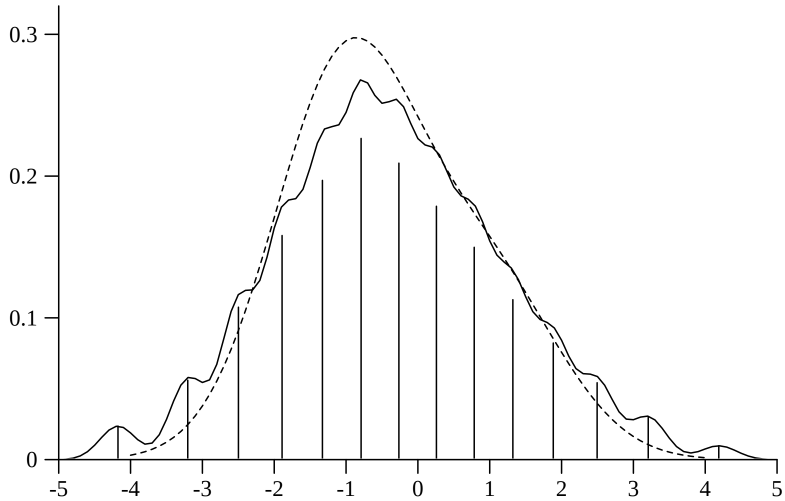

Fig. 3.1 Solid: observed and smoothed frequency distribution of the $\hat{u}_i$ values (derived from observed score sums $s = 1, \ldots, m-1$). Dashed: the presupposed distribution $f(u)$ as defined in (8). The ordinate relates to $f(u)$.

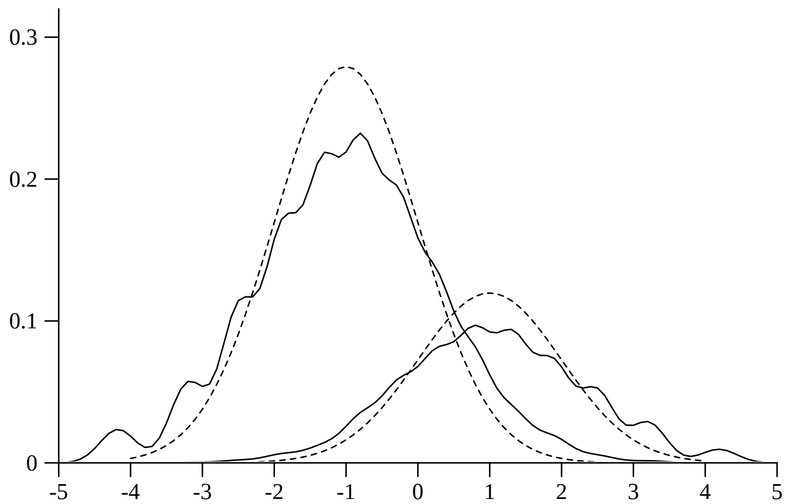

Fig. 3.2 Smoothed frequency distributions of the $\hat{u}_i$ values (solid) and the presupposed distributions $f_0(u)$ and $f_1(u)$ (dashed) in subpopulations $\mathcal{G}_0$ and $\mathcal{G}_1$. The ordinate relates to $f(u)$.

Figure 3.1 illustrates this method with the data introduced in Section 2.[2] The vertical lines show the frequencies of the $\hat{u}_i$ values (for score sums $s = 1, \ldots, m-1$). The solid line shows a smoothed version of this distribution, and the dashed line shows the distribution $f(u)$ as defined in (8). As illustrated in Figure 3.2, the same approach can be used to estimate distributions of U in subpopulations. This also allows one to estimate mean values of latent competencies in subpopulations: $\sum_{i \in \mathcal{G}_g} \hat{u}_i / n_g$. In the example, one finds the value -1.02 for the subpopulation $\mathcal{G}_0$ and the value 1.05 for the subpopulation $\mathcal{G}_1$.

[1] The same equation results from maximizing the likelihood of a version of the Rasch model (1) that considers the postulated latent competencies of individual persons, u_i, as model parameters.

[2] I use the following estimates of item parameters which result from a conditional maximum likelihood estimation of model (1): $\hat{\delta}_j$ $(j = 1, \ldots, 15)$ = -3.4993, -2.9872, -2.5064, -2.0194, -1.5140, -0.9557, -0.4925, -0.0152, 0.4861, 0.9804, 1.5171, 1.9640, 2.4773, 2.9629, 3.6020.

Although acceptable at first sight, estimation results are biased. As it is seen, in particular in Figure 3.1, the method overestimates the density in the tails of the distribution of U. This is a direct consequence of using the observed sum scores, s_i^*, as estimates of the conditional expectations $\mathrm{E}(S \mid U = u_i)$. The problem is obvious: with a single realization of a random variable one cannot reliably estimate its expectation. More observations are not available, however, because one cannot identify subsets of persons having the same latent competence (which, by assumption, are considered as exchangeable).

3.2 Using Posterior Distributions

I now consider the idea to estimate a parametric version of the distribution of U in the population $\mathcal{G}$ by a posterior distribution. This approach starts from a presupposed prior distribution, given by a density function $f^*(u)$, which is then used to define a joint distribution of U and $X_1, \ldots, X_m$:

$$g(x_1, \ldots, x_m, u) := \Pr(X_1 = x_1, \ldots, X_m = x_m \mid U = u)\, f^*(u) \tag{14}$$

This allows one to derive a distribution of U which is conditional on the observed test results:

$$g(u \mid x_1, \ldots, x_m) = \frac{g(x_1, \ldots, x_m, u)}{\int_u g(x_1, \ldots, x_m, u)\,\mathrm{d}u} \tag{15}$$

Using model (1), one gets

$$g(u \mid x_1, \ldots, x_m) = \frac{\prod_j \frac{\exp(x_j\,(u-\delta_j))}{1+\exp(u-\delta_j)}\, f^*(u)}{\int_u \prod_j \frac{\exp(x_j\,(u-\delta_j))}{1+\exp(u-\delta_j)}\, f^*(u)\,\mathrm{d}u} \tag{16}$$

Factoring out $\prod_j \exp(-x_j\,\delta_j)$ in both the numerator and the denominator, this simplifies into

$$g(u \mid x_1, \ldots, x_m) = \tag{17}$$

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

where $s = \Sigma_j x_j$. This shows that the posterior distribution of latent competencies only depends on the sum score, and it suffices to use the notation

$$g(u \mid S = s) \tag{18}$$

where $s = 0, \ldots, m$ denotes a sum score. Finally one can derive an estimate of the distribution of U in $\mathcal{G}$ by

$$g(u) = \sum_{s=0}^{m} g(u \mid S = s)\,\mathrm{P}(S^* = s) \tag{19}$$

where $\mathrm{P}(S^* = s)$ is the proportion of persons in $\mathcal{G}$ having the sum score $s_i^* = s$.

Prior distributions can be choosen arbitrarily. Most often a normal distribution is used. Since its mean value can be arbitrarily fixed (or is implied by a constraint on the item parameters), only its variance could be specified as an estimable parameter. To illustrate with the data introduced in Section 2, I use the marginal likelihood function

$$\mathcal{L}^m(\delta) = \prod_{i=1}^{n} \int_u \prod_{j=1}^{m} \frac{\exp(x_{ij}\,(u-\delta_j))}{1+\exp(u-\delta_j)}\, f^*(u)\,\mathrm{d}u \tag{20}$$

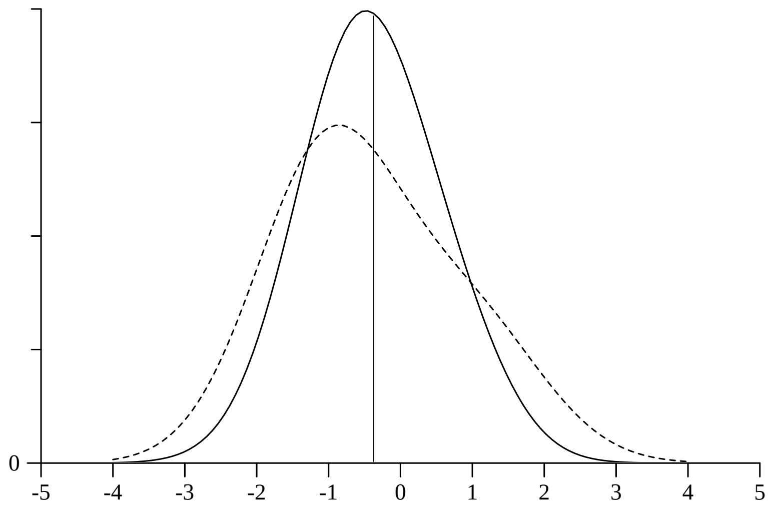

Fig. 3.3 Solid: the distribution $g(u)$ as defined in (19), based on the prior distribution $\phi(u; \mu = -0.37, \sigma = 1.36)$; dashed: the distribution $f(u)$ as defined in (8).

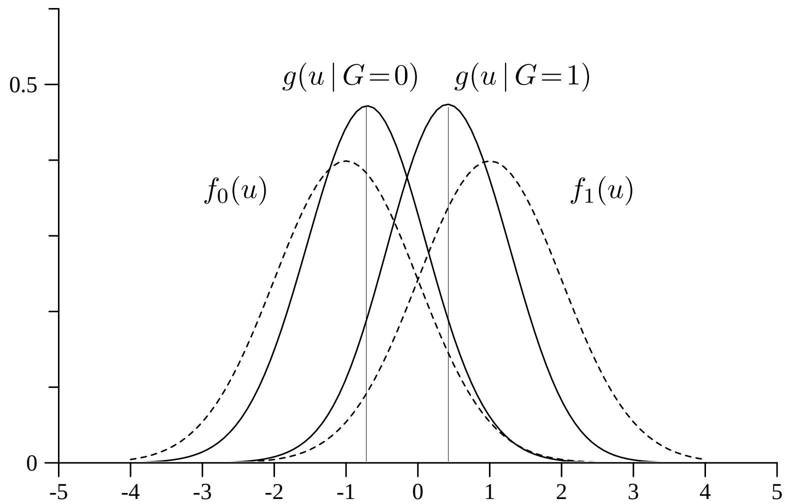

Fig. 3.4 Solid: the distributions $g(u|G = g)$ as defined in (21), based on the prior distribution $\phi(u; \mu = -0.37, \sigma = 1.36)$; dashed: the distributions $f_g(u)$.

where $f^*(u) = \phi(u; \mu, \sigma)$, and $\mu = -0.37$ and $\sigma = 1.36$ are fixed values. The resulting posterior distribution is shown in Figure 3.3. Obviously, it is not a good estimate of the distribution $f(u)$ as defined in (8).

Results are even more misleading when one uses this method for comparing distributions in subpopulations. Since the prior distribution, $f^*(u)$, does not depend on the variable representing the subpopulations, the posterior distribution in a subpopulation g is given by

$$g(u \,|\, G = g) = \sum_{s=0}^{m} g(u \,|\, S = s)\, \mathrm{P}(S^* = s \,|\, G = g) \tag{21}$$

where now $\mathrm{P}(S^* = s \,|\, G = g)$ is the proportion of persons in subpopulation $\mathcal{G}_g$ having the sum score $s_i^* = s$.

Figure 3.4 compares these distributions with the distributions $f_g(u)$ which are used for generating the data. In particular, one gets misleading estimates of mean values: -0.72 and 0.42, respectively. The difference is 1.14, but should approximately be 2.

4. Latent Regression Models

If one is interested only in the mean values of competence distributions in subpopulations, one can circumvent an explicit estimation of posterior distributions. Instead, one can start from a model which, in addition to item parameters, contains parameters representing mean competencies. This approach is called 'latent regression'.

To illustrate, I first continue with the example introduced in Section 2. Since there are only two subpopulations, $\mathcal{G}_0$ and $\mathcal{G}_1$, one can use the regression equation

$$u = g\,\beta + \epsilon \tag{22}$$

where $g \in \{0,1\}$ specifies the subpopulation and ϵ denotes values of a random variable with a standard normal distribution. Inserting this equation into the Rasch model (1), one gets the marginal likelihood function

$$\mathcal{L}^m(\delta,\beta) = \prod_{i=1}^{n} \int_{\epsilon} \prod_{j=1}^{m} \frac{\exp(x_{ij}\,(\epsilon + g_i\,\beta - \delta_j))}{1+\exp(\epsilon + g_i\,\beta - \delta_j)}\,\phi(\epsilon)\,\mathrm{d}\epsilon \tag{23}$$

where g_i denotes the subpopulation to which person i belongs. This likelihood exactly corresponds with the model that was used for the generation of data in Section 2. So it is not surprising that, in this example, one gets a good estimate of the difference of mean values: $\hat{\beta} = 2.00$. (The mean of the estimated item parameters is 0.978, so that the estimated mean values in the subpopulations are -0.98 and 1.02, respectively.)

A latent regression model can provide useful estimates of mean values even if the distributional assumptions are to some extent wrong. To illustrate, I use a modification of the example which assumes for the subpopulation $\mathcal{G}_0$ a log-normal distribution of latent competencies:

$$f_0'(u) := \frac{1}{u+3}\,\phi(\log(u+3) - 0.193) \tag{24}$$

beginning at $u = -3$; the mean value is again -1. For $\mathcal{G}_1$ I use again the distribution $f_1(u)$, so that the distribution for the whole population becomes

$$f'(u) = 0.7\,f_0'(u) + 0.3\,f_1(u) \tag{25}$$

Figure 4.1 shows these distributions. Although the likelihood (23) no longer conforms to the data generating process, one still finds an acceptable estimate $\hat{\beta} = 1.956$ (the mean value of the estimated item parameters is 1.00). In contrast, as illustrated in Figure 4.2, the method discussed in Section 3.2 would suggest a very misleading picture of the distribution of latent competencies in one of the subpopulations.

The preliminary conclusion is that latent regression models can provide acceptable estimates of mean values of latent competencies in subpopulations. Several questions remain, however. Some of these which concern the theoretical status of latent regression models will be discussed in the next section. In the remainder of the present section I consider an ambiguity in the understanding of 'mean competencies' of the persons in a subpopulation demarcated by values of the regressor variables, say $G = g$. One has to distinguish between two versions:

a) One can start from the mean latent competence in the subpopulation g defined by

$$\bar{u}_g := \sum\nolimits_u u\,\mathrm{P}(U\!=u\,|\,G\!=g) \tag{26}$$

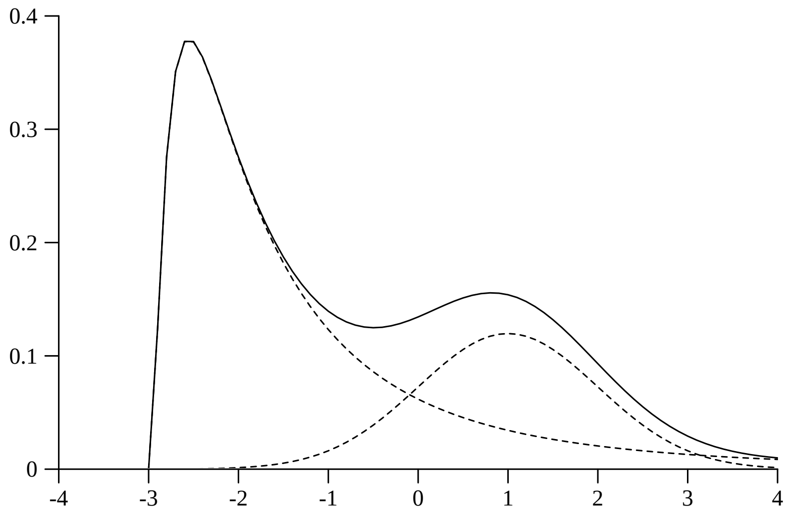

Fig. 4.1 Solid: the distribution $f'(u)$ as defined in (25); dashed: the components $f'_0(u)$ and $f_1(u)$.

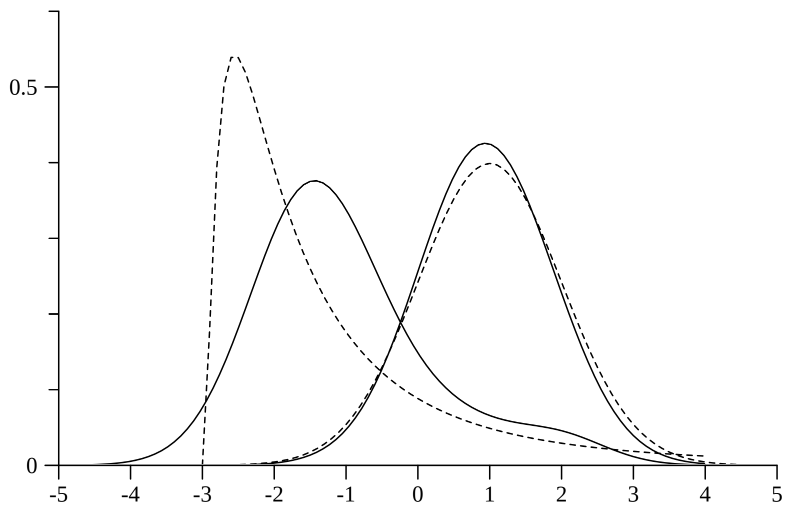

Fig. 4.2 Dashed: the distributions $f'_0(u)$ and $f_1(u)$ which are used for generating the data; solid: estimated posterior distributions.

and then think of a corresponding expectation of the number of correctly answered items:

$$\sum_{s=0}^{m} s \Pr(S = s \,|\, U = \bar{u}_g) = \sum_{s=0}^{m} s \sum_{x \in D_s} \prod_{j=1}^{m} \Pr(X_j = x_j \,|\, U = \bar{u}_g) \tag{27}$$

b) One can refer to the mean of the number of correctly answered items in the subpopulation g, that is,

$$\sum_{s=0}^{m} s \Pr(S = s \,|\, G = g) = \mathrm{E}(S \,|\, G = g) \tag{28}$$

which can immediately be estimated by the mean of the observed sum scores in $\mathcal{G}_g$.

In general, both versions of a 'mean competence' differ. For example, using the data introduced in Section 2, the mean sum score in subpopulation $G = 0$ is 5.44. On the other hand, the mean latent competence in this subpopulation is $\bar{u}_0 = -0.965$, so that the mean sum score defined in (27) takes the value 5.67. So one needs a decision about the notion of 'mean competence' to be used. Obviously, the mean value defined in (28) not only has a simpler interpretation but also can immediately be estimated without a latent regression model.

Probability metric vs. logit scale

Definition (28) corresponds to thinking of competencies in terms of probabilities (of correct answers). Instead of U, one refers to a variable $\bar{S} := h(U)$, where $h(u) := \mathrm{E}(S \mid U = u)$. Let $f(u)$ denote the distribution of U. Then

$$\mathrm{E}(S \mid G = g) = \sum_s s \int_u \Pr(S = s \mid G = g, U = u)\, f(u \mid G = g)\, du \tag{29}$$

Since $\mathrm{E}(S \mid G = g, U = u) = \mathrm{E}(S \mid U = u)$, as implied by (6), it follows:

$$\mathrm{E}(S \mid G = g) = \int_u \mathrm{E}(S \mid U = u)\, f(u \mid G = g)\, du = \tag{30}$$
$$\int_u h(u)\, f(u \mid G = g)\, du = \mathrm{E}(h(U) \mid G = g) = \mathrm{E}(\bar{S} \mid G = g)$$

This shows that $\mathrm{E}(S \mid G = g)$ can be interpreted as the mean value of $\bar{S}$ in the subpopulation g. Like U, also $\bar{S}$ is a 'latent variable' (defined by a statistical model). However, while U is defined on a logit scale, values of $\bar{S}$ can directly be interpreted as probabilities: $\bar{s}_i = h(u_i)$ is person i's probability of correctly solving the items of T_m, as postulated by the presupposed Rasch model; and $\mathrm{E}(S \mid G = g)$ corresponds to the mean of these probabilities in the subpopulation g.

5. Plausible Values

I now consider 'plausible values' which have been proposed to support the estimation of distributions of latent competencies (Mislevy, 1991; Mislevy et al., 1992; von Davier et al., 2009). The euphemistic term 'plausible value' denotes values randomly drawn from the posterior distributions which are intended to represent latent competencies. In order to discuss possible uses of such values I first refer to the construction of posterior distributions described in Section 3.2. For each person i, there is a posterior distribution $g(u \mid S = s_i^*)$, and one can think that 'plausible values' $p_i^{(k)}$ $(k = 1, \ldots, K)$ are random draws from these distributions.

Such 'plausible values' can be used to calculate K versions of a distribution of latent competencies. Their mean approximates the distribution $g(u)$ as defined in (19). This shows that 'plausible values' are simply a technical means that allows one to approximate posterior distributions (and quantities derived from such distributions) without the need of explicitly estimating the parametric densities. Of course, when posterior distributions are misleading (as was argued in Section 3.2), this remains true when they are approximated with 'plausible values'.[3]

'Plausible values' can also be used as a technical means for latent regressions. This modeling approach presupposes, for each subpopulation (as demarcated by the regressor variables), a particular distribution of latent competencies, say $f^*(u \mid G = g)$. In the examples discussed in Section 4, $f^*(u \mid G = g) = \phi(u - g\,\beta)$. Thinking of these as prior distributions, one can

[3] As an example, one can think of the proposal to use 'plausible values' for the estimation of 'proficiency levels' which are derived from quantiles of posterior distributions of latent competencies; see, e.g., OECD (2009: 100).

derive corresponding posterior distributions. In parallel to the derivation of (17), one can derive distributions

$$g(u \mid S = s, G = g) = \frac{\dfrac{\exp(s\,u)}{\prod_j 1 + \exp(u - \delta_j)}\, f^*(u \mid G = g)}{\displaystyle\int_u \dfrac{\exp(s\,u)}{\prod_j 1 + \exp(u - \delta_j)}\, f^*(u \mid G = g)\,\mathrm{d}u} \tag{31}$$

These distributions can be used to generate 'plausible values', and these values can then be used as values of the depending variable of a standard regression model. In this way, instead of estimating a latent regression model, one can estimate K versions of a standard regression model. If it is a linear model, the mean values of the estimated model parameters will be approximately identical to the corresponding parameters of the latent regression model (where the degree of approximation depends on K).[4] So it is seen that also in this context 'plausible values' are only a technical means which allows one to calculate a latent regression model without the need to maximize the corresponding likelihood function. There are, however, both practical and conceptual difficulties.

A difficulty of practical application

A first difficulty concerns the practical application. The model of the posterior distribution that is used for the generation of 'plausible values' must correspond to the latent regression model that one aims to estimate with these values. For example, it would obviously be wrong to use, instead of (23), a standard regression model with 'plausible values' drawn from a single posterior distribution for the whole population. In fact, one has to incorporate *the same* regression model that one intends to estimate into the model to be used for the generation of 'plausible values'. So one needs a new set of 'plausible values' for each latent regression model. In other words, 'plausible values' cannot be provided with any degree of generality.

A conceptual difficulty concerning the regression model

Another difficulty derives from a remarkable feature of latent regression models: The definition of the dependent variable, which is intended to represent measured competencies, depends on the specification of the model. So there is an important difference compared to ordinary regression models which presuppose that the dependent variable has a definition that is independent of the regressor variables and the model specification. This independence is essential for any explanatory claim of the model. In contrast, when using a latent regression model, already the notion of a person's competence (as defined by a measurement procedure) depends, not only on the outcome of a competence test, but also on the person's values of the regressor variables (so-called 'conditioning variables' in this context).[5]

A seemingly similar problem occurs when missing values of a dependent variables are substituted by imputed values generated by a model-dependent procedure. However, given that there is a model-independent *definition* of the dependent variable, the imputation procedure can be considered (and possibly justified) as a method for *estimating* the missing values. Mislevy and other authors have proposed that also 'plausible values' can be considered as 'imputed values'

[4]If the regression model is linear, one could also use EAP ('expected a posteriori') values

$$\mathrm{EAP}_i := \int_u u\, g(u \mid S = s_i^*, G = g_i)\,\mathrm{d}u \tag{32}$$

instead of 'plausible values', for the dependent variable.

[5]For an illustration, see Rohwer (2014).

substituting latent competencies which can never be observed (Mislevy, 1991; Mislevy et al., 1992). The analogy is misleading, however, because one no longer estimates values of a defined variable but, instead, defines the values of a variable by an imputation procedure.

In Section 1 I have proposed the following definition: the latent competence u_i of a person i is defined by $\mathrm{E}(S \,|\, U = u_i)$, that is, the expectation of the sum score in hypothetical repetitions of a test T_m. This definition entails that u_i is to be considered as a (hypothetically) fixed quantity. In contrast, a latent regression model relates to values of a random variable U_i' defined by

$$u_i' := \sum_{l=1}^{L} z_{il}\, \beta_l + \epsilon \tag{33}$$

where $z_i = (z_{i1}, \ldots, z_{iL})$ are the person's values of the regressor variables $Z = (Z_1, \ldots, Z_L)$, $\beta = (\beta_1, \ldots, \beta_L)$ is a corresponding parameter vector, and ϵ is the realization of a residual variable. This residual variable, and its distribution f_ϵ, is postulated by the model.

The interpretation proposed by Mislevy et al. can be understood in the following way: One starts from using (33) for a latent regression model that provides estimates of β and item parameters. The resulting distribution of U_i' (which depends on Z, β and the presupposed distribution f_ϵ) is then used as a prior distribution for deriving a posterior distribution $g_i'(u \,|\, Z = z_i; \beta)$; and this distribution is finally used to generate the 'plausible values' for person i. The mean of a sufficiently large number of such 'plausible values' corresponds to the EAP value

$$\int_u u\, g_i'(u \,|\, Z = z_i; \beta)\, \mathrm{d}u \tag{34}$$

There is no reason, however, why this value should be considered as an estimate of u_i; and consequently, there is no reason why random draws from $g_i'(u \,|\, Z = z_i; \beta)$ should be considered as sensible imputations. Quite the contrary, the procedure makes *the definition* of latent competencies to depend on conditioning variables and the specification of a regression model.

Measurement versus imputation

It has been suggested that 'plausible values' can also be understood as incorporating 'measurement errors' (e.g. Wu, 2005). However, 'plausible values' are random draws from posterior distributions of latent competencies which are defined by a statistical model. Using 'plausible values' instead of directly referring to the underlying distributions only creates an approximation error; there is no relationship with a measurement error.

In order to introduce a notion of measurement error, one first has to define the quantity that one intends to measure. I have proposed to refer to $\mathrm{E}(S \,|\, U = u_i)$, that is, a person's expectation of the number of correctly answered items (in a series of hypothetical repetitions of a test). The measurement error associated with an observed sum score s_i^* is then given by

$$s_i^* - \mathrm{E}(S \,|\, U = u_i) \tag{35}$$

This allows one to interpret $\Pr(S = s \,|\, U = u_i)$ as a distribution of measurement errors (around its mean). Entailed by the definition of competencies, this distribution only depends on outcomes of the competence test, not on any further covariates Z; consequently

$$\Pr(S = s \,|\, U = u_i, Z = z_i) = \Pr(S = s \,|\, U = u_i) \tag{36}$$

This approach to the definition of measurement errors can be carried over to latent competencies. As u_i corresponds to $\mathrm{E}(S \,|\, U = u_i)$, so does $\hat{u}_i$ correspond to $\mathrm{E}(S \,|\, U = \hat{u}_i)$. So one can think of

$\hat{u}_i - u_i$ as corresponding to the measurement error (35). Again, also $\hat{u}_i$ depends only on the test results, not on any further covariates.

This is different, however, when one refers to the distributions $g_i'(u \mid Z = z_i; \beta)$. These distributions depend not only on test results, but also on conditioning variables. For an ordinary imputation method it is, of course, essential that imputed values depend on all covariates to be used for the imputation. However, the idea is that the imputed value can be considered (and justified) as a (good) estimate of the missing value. And it is obvious, then, that the imputed value is to be interpreted, not as a measured value, but as an estimated value.

This understanding of an imputation procedure is not appropriate for the assessment of competencies. Since the distributions $g_i'(u \mid Z = z_i; \beta)$ depend on covariates, they cannot be interpreted as representing measurement errors. And this entails that 'plausible values' which are drawn from these distributions cannot be interpreted as more or less accurate measurements.

6. Conclusion

In this paper I have discussed approaches to the definition and estimation of distributions of competencies (based on outcomes of a single competence test). The focus is on approaches which presuppose a Rasch model. I have proposed to interpret its random variables as representing hypothetical repetitions of the actually used test. This allows one to think of a correspondence between values of a variable, U, representing latent competencies, and expectations of observable sum scores, formally: $\mathrm{E}(S \mid U = u)$. There are then two possibilities to refer to individual competencies and their distribution in a population $\mathcal{G}$: (A) Individual competencies are referred to by values of U, and the interest concerns the distribution of U in $\mathcal{G}$. (B) The competencies of a person is referred to by a distribution of possible sum scores, conditional on the person's value of U. The distribution of competencies in a population $\mathcal{G}$ is then to be described as an unconditional distribution of S which is derived from postulating a fixed distribution of latent competencies in $\mathcal{G}$.

Notwithstanding their intimate relationship, possibilities to estimate the distributions are remarkably different. While distributions of S can easily be estimated from observed sum scores, there is no straightforward approach to the estimation of distributions of latent competencies (or expectations of S). In particular, methods which require parametric assumptions will often suggest misleading results.

To some extent, these problems can be circumvented when the interest only concerns mean values of latent competence distributions. This is the motivating idea of latent regression models. There are, however, two drawbacks. First, this approach is based on a notion of 'mean latent competence' which is difficult to understand. Moreover, there is an alternative notion which simply refers to the mean of observable sum scores, and when using this notion one no longer needs a latent regression model.

Second, a latent regression model makes the definition of its dependent variable to depend on the actually used regressor variables, the model specification, and an arbitrarily presupposed distribution of residuals. This method therefore conflates the definition of latent competencies with their explanation (prediction). For example, the definition of math competencies then depends on a person's sex (Rohwer, 2014).

Finally, I have considered the concept of 'plausible values'. These are random draws from posterior distributions and can be used as a technical means that allows one to avoid an explicit reference to parametric representations of such distributions. However, when such distributions misrepresent latent competencies the same will be true when one uses 'plausible values' generated from such distributions. On the other hand, for the estimation of latent regression models, one would need ‘plausible values’ which are derived from models incorporating the actually intended model specification. Consequently, ‘plausible values’ which possibly could ease the estimation of a latent regression model cannot be provided with any degree of generality.

References

Holland, P. W. (1990). On the Sampling Theory Foundations of Item Response Theory Models. *Psychometrika* 55, 577–601.

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.

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.

von Davier, M., 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.

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