Longitudinal Measurement of Childrens' Receptive Vocabulary with NEPS Data

G. Rohwer, March 2016

Abstract. I consider two tests of receptive vocabulary administered, respectively, in the first and third wave of starting cohort 3 of the National Educational Panel Study (NEPS). The first test was administered in 2011 in a sample of children visiting a Kindergarten in that year; a second test took place two years later. I compare test results of 518 children who have participated in both tests. I first use 38 items which are identical in both tests. Based on sum scores, one sees a remarkable increase in the receptive vocabulary. I then show that the data are not compatible with a joint Rasch model which assumes time-invariant item parameters. As an alternative basis for longitudinal comparisons I suggest that a sufficient condition for two tests measuring the same kind of competence is that they consist of identical items.

1 Introduction

I consider two tests of receptive vocabulary administered, respectively, in the first and third wave of starting cohort 3 of the National Educational Panel Study (NEPS).[1] The first test was administered in 2011 in a sample of 2948 children who visited a Kindergarten in that year; the mean age of the children was five years (SD = 0.32). Two years later, 551 of these children could be followed to become part of the third wave sample where a second test of receptive vocabulary was administered. Subsequently, I use data on 518 children who have participated in both tests and have given a valid answer to at least one item.

The two tests are versions of a Peabody Picture Vocabulary Test, based on a previous research version developed by Roßbach, Tietze and Weinert (2005).[2] The tests consist of 77 and 66 binary items, respectively; 38 items are used in both tests. Corresponding to each item there is a spoken word, and the task is to select, out of four pictures, that picture which fits to the given word. Consequently, all items have a multiple-choice format with four alternatives.

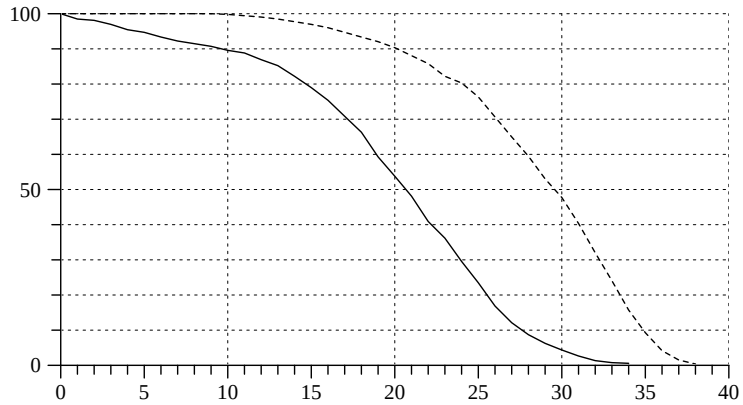

Figure 1 Proportion of children (in %) with a sum score greater than, or equal to, the number of items indicated on the abscissa. First test solid line, second test dashed line.

2 Repeating the same test

In order to assess changes of competencies one has to repeat 'the same test' at both points in time. An obvious possibility is to use identical items. This can be achieved with the 38 linking items which are identical in the two tests.

As a simple measure one can use the sum score, that is, the number of correctly solved items. (Here and subsequently, all missing responses are evaluated as wrong responses.) Figure 1 shows that a remarkable increase in the receptive vocabulary has taken place. For example, the proportion of children who were able to correctly answer at least 20 items has increased from 53 to 90 percent. The mean value changed from 19.2 (SD 7.1) to 27.9 (SD 5.8), indicating also a remarkable decrease in inequality. Figure 2 shows that almost all children increased there vocabulary.

[1] I acknowledge that this paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Kindergarten, 10.5157/NEPS:SC2:3.0.0. From 2008 to 2013, NEPS data was collected as part of the Framework Program for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF). As of 2014, NEPS is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. For a general introduction see Blossfeld, H.-P., Roßbach, H.-G, & von Maurice, J. (Eds.) (2011). Education as a Lifelong Process – The German National Educational Panel Study (NEPS). Zeitschrift f. Erziehungswissenschaft, 14.

[2] For background information see Berendes et al. (2013).

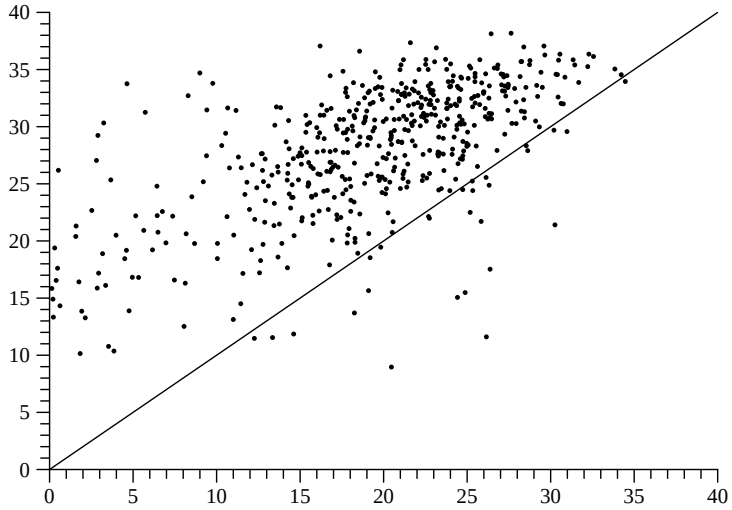

Figure 2 Scattergram of the sum scores in the first test (abscissa) and the second test (ordinate).

3 Using a joint Rasch model

I now consider a joint Rasch model for the two tests. Let J^c denote the set of common items, and J^a and J^b the item sets which are unique in the first and the second test, respectively. The variables (vectors) representing responses will be denoted, respectively, by X_t^c, X_t^a and X_t^b for the test T_t $(t = 1, 2)$; and corresponding vectors of individual values will be denoted by $x_{i,t}^c$, $x_{i,t}^a$ and $x_{i,t}^b$ $(i = 1, \ldots, 518)$. Of course, values of X_2^a and X_1^b are not available.

I begin with the common items. A cross-sectional Rasch model for person i at time t can be written

$$\Pr(X_t^c = x_{i,t}^c \,|\, \theta_{i,t}^c, \delta_t^c) = \prod_{j \in J^c} \frac{\exp(\theta_{i,t}^c - \delta_{j,t}^c)^{x_{ij,t}^c}}{1 + \exp(\theta_{i,t}^c - \delta_{j,t}^c)} \tag{1}$$

where δ_t^c is a vector of item parameters whose components are denoted by $\delta_{j,t}^c$; $x_{ij,t}^c$ denotes components of $x_{i,t}^c$, and $\theta_{i,t}^c$ is intended to represent person i's receptive vocabulary at the time of the test T_t.

The model entails an important claim: Given item parameters δ_t^c, a person's responses only depend on $\theta_{i,t}^c$. For a joint model of the two tests

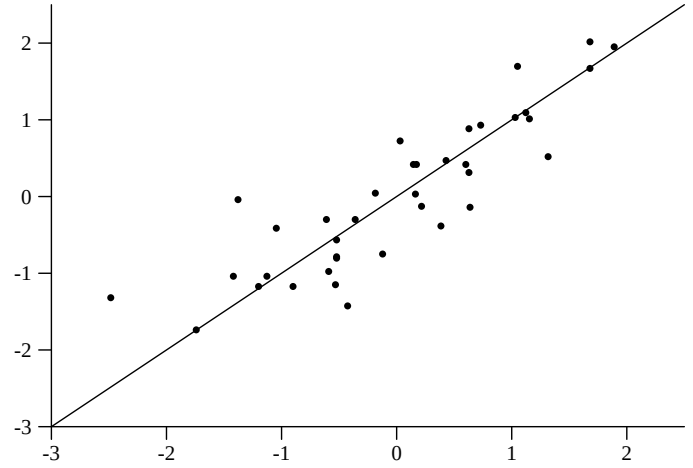

Figure 3 For each item, CML estimates of item parameters in the first and the second test are indicated on the abscissa and the ordinate, respectively.

this claim entails

$$\begin{aligned} &\Pr(X_1^c = x_{i,1}^c, X_2^c = x_{i,2}^c \,|\, \theta_{i,1}^c, \delta_1^c, \theta_{i,2}^c, \delta_2^c) = \\ &\quad \Pr(X_1^c = x_{i,1}^c \,|\, \theta_{i,1}^c, \delta_1^c) \Pr(X_2^c = x_{i,2}^c \,|\, \theta_{i,2}^c, \delta_2^c) = \\ &\quad \prod_{j \in J^c} \frac{\exp(\theta_{i,1}^c - \delta_{j,1}^c)^{x_{ij,1}^c}}{1 + \exp(\theta_{i,1}^c - \delta_{j,1}^c)} \prod_{j \in J^c} \frac{\exp(\theta_{i,2}^c - \delta_{j,2}^c)^{x_{ij,2}^c}}{1 + \exp(\theta_{i,2}^c - \delta_{j,2}^c)} \end{aligned} \tag{2}$$

So one can separately estimate the two cross-sectional models. I use conditional maximum likelihood (CML) estimation because this method does not require assumptions about the distribution of the childrens' competencies, and for identification of item parameters I use $\sum_{j \in J^c} \delta_{j,t}^c = 0$. Figure 3 shows estimates of the item parameters.

Linking the two tests with a Rasch model would require the assumption that the item parameters did not change:

$$\text{for } j \in J^c\text{: } \delta_{j,1} = \delta_{j,2} \tag{3}$$

Already Figure 3 makes this assumption questionable. In order to formally test whether the constraint (3) can be added to the model (2) one can use a likelihood ratio test, based on comparing the conditional log-likelihoods provided by CML estimation. Without the constraints, the conditional log-likelihood of model (2) has the value -17599.2; when adding the constraint

(3) the value is -17777.7, and the test statistic has the value 357. With 38 degrees of freedom there is strong evidence against assuming a common set of item parameters.

This result also illustrates that item parameters do not describe properties of items, but reflect the distribution of competencies in a population. In a cross-sectional view, presupposing a population with a fixed distribution of competencies, it might be possible to consider item parameters as fixed quantities (which, in a sense, define a 'measurement instrument'). In longitudinal applications, in contrast, one should expect that changes in the distribution of competencies will lead to changing item parameters. In fact, the assumption of time-invariant item parameters is then equivalent with an extremely restrictive assumption about the learning processes through which competencies could change. For example, think of a child whose probability of correctly understanding a word increased from 0.2 to 0.9. Now consider another word which the child could correctly understand with probability 0.6 at the time of the first test. The assumption of time-invariant item parameters would entail that the child must have learnt to correctly understand this word with probability 0.98.

4 Information from non-repeated items

So far I have only used the common items of the two tests. It is possible, of course, to estimate a joint Rasch model which includes all items. Without further constraints this will be equivalent with separate Rasch models for the two tests. Then one could add the constraint that the items in J^c have identical parameters in the two tests. However, since this constraint cannot be justified for model (2), it also cannot be justified for an enlarged version of that model.

A further point is noteworthy. Even if the assumption of equal parameters of the items in J^c could be justified, one could not draw any conclusions about the other item parameters. A joint model for the set of all observed items will always be compatible with arbitrary assumptions about parameters of items in J^a at the time of the second test, and with arbitrary assumptions about parameters of items in J^b at the time of the first test. For example, thinking of the items in J^a, the available data are compatible with assuming that several children have lost a knowledge of some of the corresponding words. And, what is presumably more important, thinking of the items in J^b, the available data are compatible with assuming that several, or perhaps all, children have not yet learned the meaning of the corresponding words at the time when the first test was administered.

5 Discussion

What follows from the fact that the assumption (3) cannot be justified in the present application? It has been argued that the condition of identical item parameters is a necessary part of the assumption that two tests measure 'the same construct' (e.g., Stocking and Lord, 1983; Rupp and Zumbo, 2006; Millsap, 2010). Accepting this view would entail that the identical 38 items assess different kinds of receptive vocabulary. I suggest that a reasonable alternative employs the following principle:

> A sufficient condition for two tests measuring the same kind of competence is that they consist of identical items. (4)

It remains then the question of how to quantify this competence. In Section 2 I have used sum scores. One might argue that a Rasch model is required for justifying sum scores because this model allows one to predict responses conditional on sum scores. However, this is a cross-sectional prediction and only requires cross-sectional models.

In order to understand the requirement of time-invariant item parameters, it is helpful to consider the relationship between sum scores and the person parameters $\theta^c_{i,t}$. When using weighted likelihood estimation, as proposed by Warm (1989), the relationship is

$$s^c_{i,t} = h_t(\theta^c_{i,t}) := \sum_{j\in J^c} \pi_{ij} - \frac{\sum_{j\in J^c} \pi_{ij}\,(1-\pi_{ij})\,(1-2\,\pi_{ij})}{2\,\sum_{j\in J^c} \pi_{ij}\,(1-\pi_{ij})} \tag{5}$$

where $\pi_{ij} := \exp(\theta^c_{i,t} - \delta^c_{j,t})/(1+\exp(\theta^c_{i,t} - \delta^c_{j,t}))$, and $s^c_{i,t}$ is person i's sum score w.r.t. J^c of the test T_t. The equation shows that estimates of the parameters $\theta^c_{i,t}$ result from the observed sum scores:

$$\theta^c_{i,t} = h_t^{-1}(s^c_{i,t}) \tag{6}$$

The function h_t^{-1} can be viewed as a scale transformation of the observed sum scores which depends on estimates of the item parameters δ^c_t. Therefore, if these parameters change between two tests, the same sum score will be transformed into two different theta values.

I conclude that, when item parameters change, theta values can be used for cross-sectional comparisons between persons, but it is questionable whether they can be used for representing changes of competencies. It is possible, however, that the bias due to changing item parameters is small. For the present application, Figure 4 shows the scale transformations $h_t^{-1}(s)$, for $t = 1, 2$, resulting from two separate Rasch models for the common items of the two tests. They are almost indistinguishable. Therefore, instead of using sum scores for investigating changes of

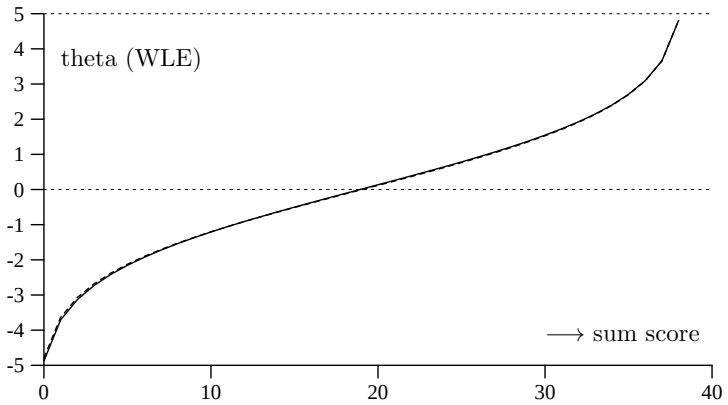

Figure 4 Scale transformations $h_t^{-1}(s)$, for $t = 1$ (solid) and $t = 2$ (dashed), resulting from two separate Rasch models for the common items of the two tests.

receptive vocabulary, one also could use theta values resulting from a joint model with the constraint (3).

This conclusion concerns comparisons based on the set of common items, J^c. Accepting the principle (4), in order to use the complete tests, one would need a reference to a comprehensive test, say T_t^*, which includes the complete set of items, $J := J^c \cup J^a \cup J^b$. This requires assumptions about how the children would have responded to items in J^a if presented at $t = 2$, and how they would have responded to items in J^b if presented at $t = 1$. As I have argued above, such assumptions cannot be justified with the available data. Responses to items in J^a and J^b can therefore only be used for cross-sectional comparisons of the receptive vocabulary.

References

Berendes, K., Weinert, S., Zimmermann, S., Artelt, C. (2013). Assessing Language Indicators across the Lifespan within the German National Educational Panel Study (NEPS). *Journal for Educational Research Online*, 5, 15–49.

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.

Millsap, R. E. (2010). Testing Measurement Invariance Using Item Response Theory in Longitudinal Data: An Introduction. *Child Development Perspectives* 4, 5–9.

Roßbach, H. G., Tietze, W., Weinert, S. (2005). Peabody Picture Vocabulary Test – Revised. Deutsche Forschungsversion des Tests von L. M. Dunn & L. M. Dunn von 1981. Universität Bamberg, FU Berlin.

Rupp, A. A., Zumbo, B. D. (2006). Understanding Parameter Invariance in Unidimensional IRT Models. *Educational and Psychological Measurement* 66, 63–84.

Stocking, M. L., Lord, F. M. (1983). Developing a Common Metric in Item Response Theory. *Applied Psychological Measurement* 7, 201–210.

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