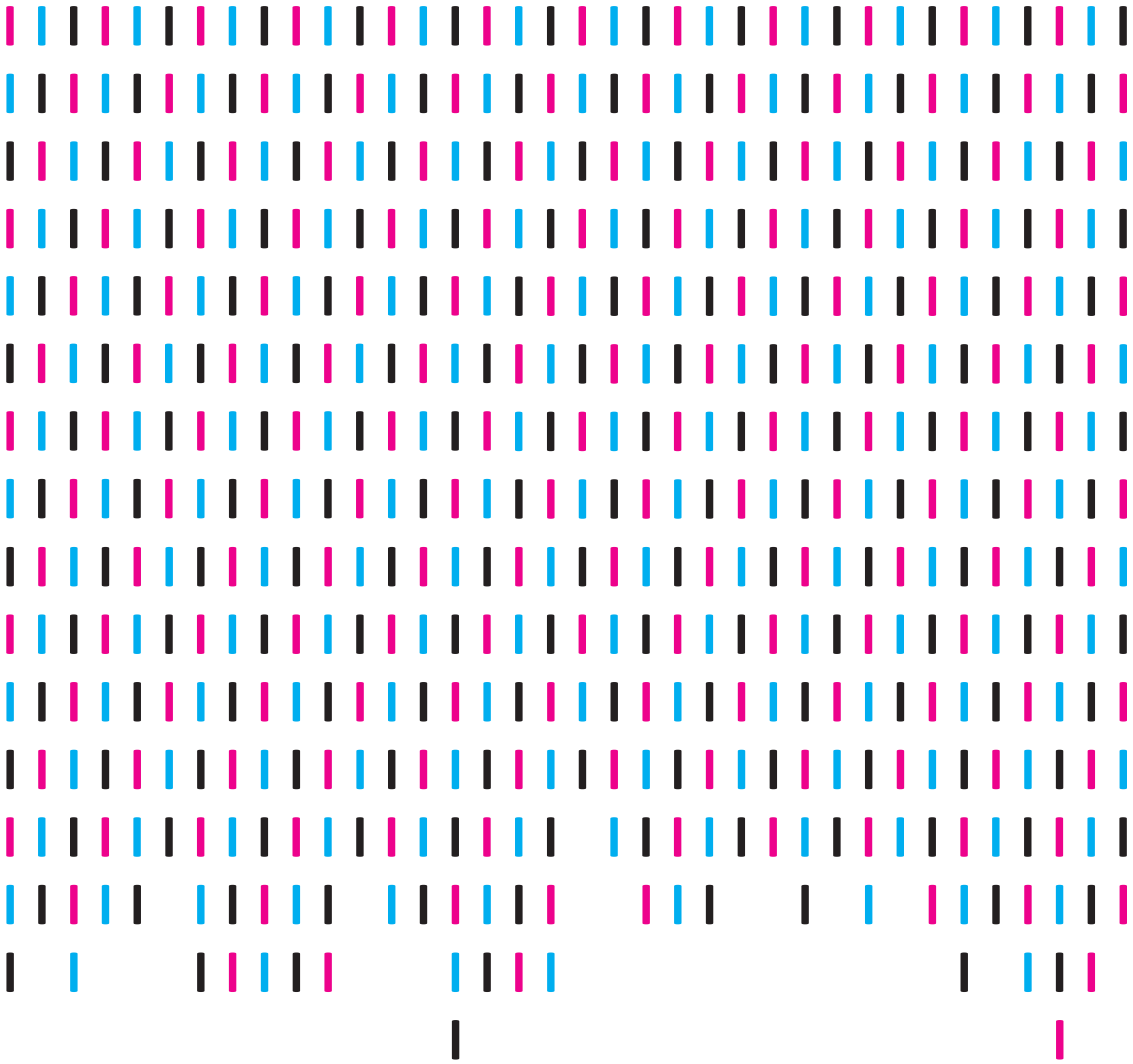


NEPS WORKING PAPERS

Götz Rohwer

USING NEPS DATA FOR COMPARING MATH COMPETENCIES AT GRADES 5 AND 7

NEPS Working Paper No. 60
Bamberg, October 2015

LIfBi
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

Using NEPS Data for Comparing Math Competencies at Grades 5 and 7

Götz Rohwer, Ruhr-Universität Bochum

October 2015

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

Bibliographic data:

Rohwer, G. (2015). *Using NEPS Data for Comparing Math Competencies at Grades 5 and 7.* (NEPS Working Paper No. 60). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study.

Using NEPS Data for Comparing Math Competencies at Grades 5 and 7

Abstract

Using data from two tests of math competencies administered at grades 5 and 7 in starting cohort 3 of the National Educational Panel Study, the paper discusses how to quantify changes of the competencies. It is shown that equating the two tests with a joint Rasch model cannot be justified statistically. Moreover, the conception of a longitudinally valid test has to take into account that there is a cumulative development of new math competencies across educational stages. It is proposed that this can be achieved by a reference to the same set of items at both grades which allows using simple sum scores for a quantification of changes of competencies.

Keywords

Math competencies, vertical equating, longitudinally valid test

Contents

1. Introduction

This paper uses data from two tests of math competencies administered at grade 5 and grade 7 in starting cohort 3 of the National Educational Panel Study (NEPS).[1] I discuss how these data can be used to investigate changes of math competencies between the two grades.

Based on information in the currently available SUF (version 3.0.0), the following table shows the items:

	Grade 5		Grade 7	
Item	Variable	Format	Variable	Format
1	MAG5D041	MC4	MAG9Q071	MC4
2	MAG5Q291	SCR	MAG7V071	MC4
3	MAG5Q292	SCR	MAG7R081	MC5
4	MAG5V271	MC4	MAG7Q051	MC4
5	MAG5R171	MC4	MAG5Q301	SCR
6	MAG5Q231	SCR	MAG9D151	MC4
7	MAG5Q301	SCR	MAG5D051	MC4
8	MAG5Q221	SCR	MAG5D052	MC4
9	MAG5D051	MC4	MAG9V011	MC5
10	MAG5D052	MC4	MAG9V012	MC5
11	MAG5Q14S	SCR	MAG7Q041	MC4
12	MAG5Q121	MC4	MAG7D042	MC4
13	MAG5R101	MC4	MAG7R091	MC5
14	MAG5R201	MC4	MAG9Q181	MC4
15	MAG5Q131	SCR	MAG7D011	MC4
16	MAG5D02S	SCR	MAG7V012	MC4
17	MAG5D023	SCR	MAG7V031	MC4
18	MAG5V024	SCR	MAG5R251	MC4
19	MAG5R251	MC4	MAG7D061	MC4
20	MAG5V01S	CMC	MAG5V321	SCR
21	MAG5V321	SCR	MAG9V091	MC4
22	MAG5V071	MC4	MAG5R191	MC4
23	MAG5R191	MC4	MAG7R02S	CMC
24	MAG5V091	MC5		

The test in grade 5 consists of 24 items; 11 have a short construction format (SCR), 12 have a multiple choice format (the number following MC provides the number of alternatives), and one item has a partial credit format (CMC). In grade 7 there are 23 items, almost all have a MC format. Except of the CMC items, all other items are binary. Six items (underlined in the table) are identical in the two tests and can be used for linking.

In the first wave (grade 5) there are 5194 pupils who have given a valid (not necessarily correct) answer to at least one item. Of these, 3833 pupils also have participated and given at least one valid answer in the grade 7 test. This is the number of cases used in this paper.

One has to decide how to evaluate missing answers.[2] While missing answers to SCR items can

[1] This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort Grade 5, 10.5157/NEPS:SC3:3.0.0. 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). [Special Issue] Zeitschrift für Erziehungswissenschaft: 14.

[2] For a description of missing answers in the first wave see Duchhardt and Gerdes (2012).

sensibly be evaluated as wrong answers, this seems not appropriate for MC items which in any case could have been answered simply by guessing. In subsequent calculations I therefore use two different methods (see Rohwer, 2013):

Method I All missing answers are evaluated as wrong answers.

Method II Missing answers to SCR items are evaluated as wrong answers; missing answers to an MC item with a alternatives are substituted by correct answers with probability $1/a$ and wrong answers with probability $1 - 1/a$.[3]

To ease the discussion and the calculations, I do not use the two CMC items.

2. Repeating the same test

In order to assess changes of competencies one has to use 'the same test' at both points in time. An obvious possibility is to use identical items. This can be achieved with the linking items. To ease notations, I represent these items by variables $X_{j,t}$ as follows:

Item	Grade 5	Grade 7
MAG5Q301	$X_{1,1}$	$X_{1,2}$
MAG5D051	$X_{2,1}$	$X_{2,2}$
MAG5D052	$X_{3,1}$	$X_{3,2}$
MAG5R251	$X_{4,1}$	$X_{4,2}$
MAG5V321	$X_{5,1}$	$X_{5,2}$
MAG5R191	$X_{6,1}$	$X_{6,2}$

Values of $X_{j,t}$ will be denoted by $x_{ij,t}$; $i = 1, \ldots, 3833$, $j = 1, \ldots, 6$ and $t = 1, 2$.

2.1 Comparing sum scores

A simple quantification of competencies uses sum scores. For person i in grade t, the sum score is defined as

$$s_{i,t} := \sum_{j=1}^{6} x_{ij,t} \tag{1}$$

Figure 1 shows frequency distributions of these sum scores in the two grades. Obviously, there is a shift to higher competencies. The mean sum score changes from 3.16 to 3.96. This is accompanied by a decreasing inequality: the coefficient of variation changes from 0.50 to 0.39. These figures are based on using *Method I* for the evaluation of missing answers.

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

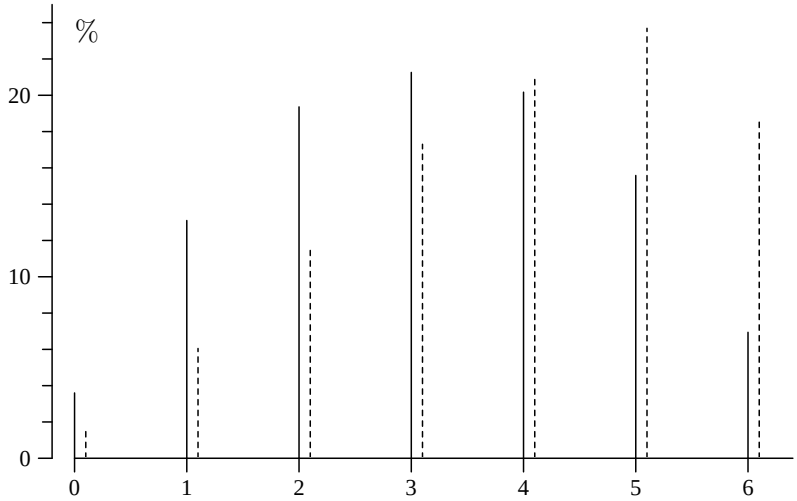


Fig. 1 Frequency distributions (in percent) of the sum scores in grade 5 (solid) and grade 7 (dashed).

One gets essentially the same results when using *Method II*. The following table compares the distributions of the sum scores (in percent):

	Method I		Method II	
s	Grade 5	Grade 7	Grade 5	Grade 7
0	3.6	1.7	3.1	1.6
1	13.1	6.1	12.2	5.8
2	19.4	11.5	19.1	10.9
3	21.3	17.3	21.3	17.4
4	20.2	21.1	21.3	21.3
5	15.6	23.7	15.9	24.0
6	6.9	18.7	7.2	19.0

(2)

It is remarkable, however, that the correlation between the two sum scores is relatively low: 0.6 (method I) or 0.59 (method II). The joint distribution (based on method I) is as follows.

		grade 7							
		0	1	2	3	4	5	6	total
grade 5	0	17	43	33	24	14	3	4	138
	1	28	97	150	124	63	30	10	502
	2	15	58	139	191	176	115	48	742
	3	3	22	79	171	224	217	99	815
	4	2	9	29	108	199	237	189	773
	5	1	3	8	40	103	229	213	597
	6	0	0	1	5	28	77	155	266
	total	66	232	439	663	807	908	718	3833

(3)

2.2 Using a Rasch model

Instead of sum scores, researchers often use a Rasch model for the quantification of competencies. For grade t, the model can be written

$$\Pr(X_{1,t} = x_{i1,t}, \ldots, X_{m,t} = x_{im,t} \,|\, \theta_{i,t}, \delta_t) = \prod_{j=1}^{m} \frac{\exp(\theta_{i,t} - \delta_{j,t})^{x_{ij,t}}}{1 + \exp(\theta_{i,t} - \delta_{j,t})} \tag{4}$$

where $m = 6$, $\delta_t := (\delta_{1,t}, \ldots, \delta_{m,t})$ is a vector of item parameters, and $\theta_{i,t}$ is intended to represent person i's competence. This is a cross-sectional model (it can be used to discriminate competencies in a single set of test-takers). Item parameters are identified only up to an additive constant, so that one has to add constraints

$$\sum_{j=1}^{m} \delta_{j,t} = \kappa_t \tag{5}$$

with arbitrary constants κ_t. Using the constraints with $\kappa_t = 0$, the left part of the following table shows estimates of the item parameters $\delta_{j,t}$.[4]

	Method I		Method II		Method I	Method II
j	$\hat{\delta}_{j,1}$	$\hat{\delta}_{j,2}$	$\hat{\delta}_{j,1}$	$\hat{\delta}_{j,2}$	$\hat{\delta}_j$	$\hat{\delta}_j$
1	0.79	0.80	0.84	0.83	0.79	0.83
2	−2.45	−2.19	−2.42	−2.17	−2.35	−2.32
3	−0.36	−0.85	−0.35	−0.82	−0.56	−0.55
4	0.53	0.68	0.46	0.65	0.60	0.55
5	1.41	1.54	1.46	1.57	1.48	1.51
6	0.07	0.02	0.01	−0.05	0.04	−0.02

(6)

Obviously, the item parameters changed between the two grades. This entails that estimates of $\theta_{i,t}$ cannot immediately be compared.

A sensible comparison would require a model having identical item parameters for the two grades. Starting from (4), one can add the constraints

$$\text{for } j = 1, \ldots, m\text{: } \delta_{j,1} = \delta_{j,2} =: \delta_j \tag{7}$$

Estimating a joint model with these constraints,[5] estimates of the item parameters δ_j are shown in the right part of Table (6).

However, from a statistical point of view, adding the constraints (7) cannot be justified. Based on the log-likelihoods

Model	Method I	Method II
$t = 1$	−6204.3	−6241.5
$t = 2$	−5390.1	−5369.2
joint model with constraints	−11637.8	−11653.6

(8)

[4] Here and subsequently, I use conditional maximum likelihood (CML) estimation.

[5] That is, using the likelihood

$$\prod_{i=1}^{n} \prod_{t=1}^{2} \prod_{j=1}^{m} \frac{\exp(\theta_{i,t} - \delta_j)^{x_{ij,t}}}{1 + \exp(\theta_{i,t} - \delta_j)}$$

with the constraint $\sum_j \delta_j = 0$.

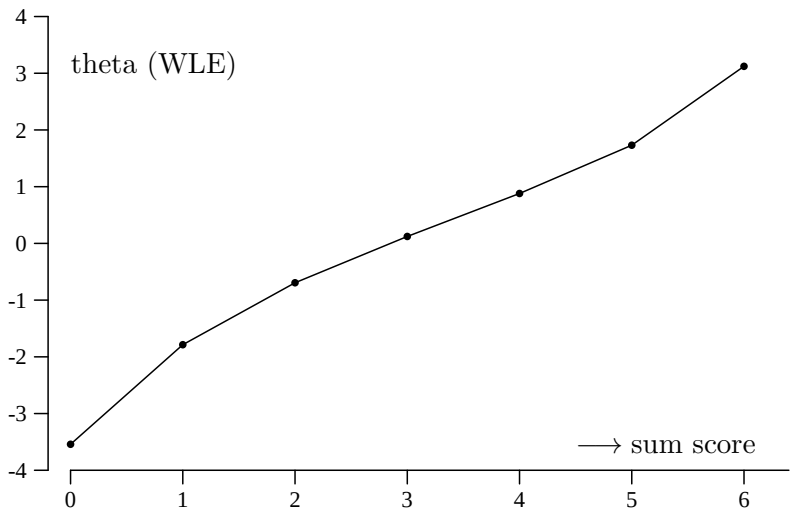


Fig. 2 Correspondence between sum scores and WLEs calculated by solving (10).

the test statistic for a likelihood ratio test is 86.8 (method I) or 85.8 (method II), and with 5 degrees of freedom is highly significant against the equating constraints. Therefore, from a statistical point of view, the assumption that a common Rasch model fits the test results in the two grades must be rejected.

What follows from this result? 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 six items assess different constructs in grade 5 and grade 7. 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. (9)

Based on this principle, one can first of all justify the comparisons discussed above using sum scores. Moreover, regardless of the fact that item parameters of separate Rasch models changed between the two grades, one can make sense of item parameters δ_j resulting from a joint model with equating constraints. They can be used for a scale transformation of the sum scores, for example in the form of weighted ML estimates. As proposed by Warm (1989), these estimates result from solving equations

$$s_{i,t} = h(\theta_{i,t}) := \sum_{j=1}^{m} \pi_{ij} - \frac{\sum_{j=1}^{m} \pi_{ij}\,(1-\pi_{ij})\,(1-2\,\pi_{ij})}{2\,\sum_{j=1}^{m} \pi_{ij}\,(1-\pi_{ij})}, \quad \pi_{ij} := \frac{\exp(\theta_{i,t} - \hat{\delta}_j)}{1+\exp(\theta_{i,t} - \hat{\delta}_j)} \tag{10}$$

For each possible sum score $s_{i,t}$ there is a corresponding value of $\theta_{i,t}$, regardless of i and t. Based on the estimates of δ_j in the right half of (6), the scale transformation is shown in Figure 2.[6] Taking care of the different metrics, both scales may be used to quantitatively represent the competencies.

[6] Methods I and II for the evaluation of missing answers lead to essentially the same scale transformation.

3. Tests with partially overlapping item sets

I now consider the complete tests (except for the two CMC items).

The test T1 in grade 5 consists of $m_1 = 23$ items represented by variables $X_{1,1}, \ldots, X_{m_1,1}$
The test T2 in grade 7 consists of $m_2 = 22$ items represented by variables $X_{1,2}, \ldots, X_{m_2,2}$

As before, the variables $X_{1,t}, \ldots, X_{6,t}$ relate to the linking items; the ordering of the remaining variables equals the ordering of the items shown in the Introduction. Values of the variables will be denoted by $x_{ij,1}$ and $x_{ij,2}$, respectively.

Obviously, since the two tests consist of mainly different items, sum scores (divided by the number of items) cannot be used to compare competencies. In the following, I first show that, in the present application, equating the tests via the linking items leads into severe difficulties. I then consider an alternative based on principle (9) suggested above.

3.1 Limitations to vertical equating

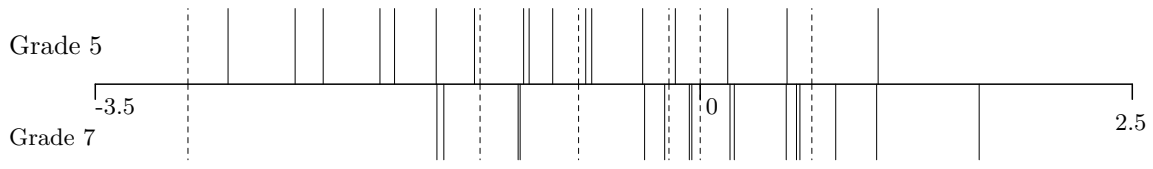
One can start from a joint Rasch model

$$\Pr(X_{1,1} = x_{i1,1}, \ldots, X_{m_1,1} = x_{im_1,1}, X_{1,2} = x_{i1,2}, \ldots, X_{m_2,2} = x_{im_2,2} \,|\, \theta_{i,1}, \theta_{i,2}, \delta_1, \delta_2) = \prod_{j=1}^{m_1} \frac{\exp(\theta_{i,1} - \delta_{j,1})^{x_{ij,1}}}{1 + \exp(\theta_{i,1} - \delta_{j,1})} \prod_{j=1}^{m_2} \frac{\exp(\theta_{i,2} - \delta_{j,2})^{x_{ij,2}}}{1 + \exp(\theta_{i,2} - \delta_{j,2})} \qquad (11)$$

where δ_1 and δ_2 are vectors of item parameters in T1 and T2, respectively. For parameter identification (as an alternative to (5)), I add the constraint $\delta_{1,1} = \delta_{1,2} = 0$. Without further constraints, this joint model is equivalent with two separate Rasch models. CML estimation provides the log likelihood -73892.3.

Equating can be achieved by adding the constraint that the linking items have identical item parameters (condition (7)). CML estimation of the joint model then results in the log likelihood -73931.9. The test statistic for a likelihood ratio test is 79.2 (method I) or 80.4 (method II), and with 5 degrees of freedom is again highly significant against the equating constraints. Therefore, again, from a statistical point of view, equating the two tests with a Rasch model cannot be justified.

However, even if there were no statistical objections against the assumption of identical parameters of the linking items, one could not conclude that both tests measure the same kind of math competence.[7] Consider the following graphic which shows the distribution of item parameters resulting from equating parameters of the linking items (indicated by dashed lines).



Assuming that both tests assess the same kind of math competence would entail that students in grade 5 are already able to solve the items in the second test, depending on their value of $\theta_{i,1}$.

[7] This is often considered as a precondition for a sensible equating of two tests, see e.g. von Davier and von Davier (2007).

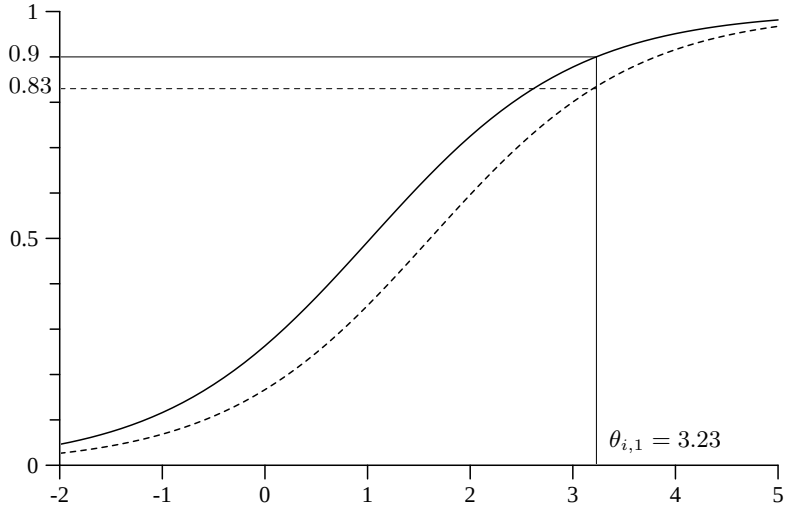


Fig. 3 Comparison of `MAG5Q121` in T1 ($\delta_{15,1} = 1.03$, solid) and `MAG9V091` in T2 ($\delta_{22,2} = 1.61$, dashed).

To illustrate, I consider the most difficult items: `MAG5Q121` in T1 ($\delta_{15,1} = 1.03$) and `MAG9V091` in T2 ($\delta_{22,2} = 1.61$); Figure 3 provides the item characteristic curves. Now think of a student who, in grade 5, has learnt to solve the first of these items with probability 0.9. Given the model, this entails $\theta_{i,1} = 3.23$. However, if all items correspond to the same kind of competence, this student would already in grade 5 be able to solve the second item (the most difficult in T2) with probability 0.83.

This conclusion is highly implausible when thinking of the development of math competencies during educational processes. This development has two aspects: students become more proficient in solving already known tasks, and they learn to solve new tasks. Therefore, T2 probably contains some items which students have not already learned to solve in grade 5, and therefore are of a different kind.[8]

Unfortunately, the meaning of the test items has not been published yet. However, the discussion of the NEPS math tests by Neumann et al. (2013) suggests that the items correspond to what students *should know* in the respective grades. For example, with regard to the content area "Data and chance" the authors say:

> In Grade 5, children should be able to deal with data more systematically and purposefully than in kindergarten. Competence in this area is indicated by the extent to which children are able to collect data from simple experiments or observations and represent them in tables or figures such as bar charts or line charts. In the subarea "chance" it is required to compare the probabilities of different events in random experiments and to know the basic concepts of "certain", "impossible", or "likely". Children should also be able to assess winning chances in dice games. (p. 90)

The authors continue with remarks about grade 9:

> In Grade 9, students should be able to plan simple statistical studies, measure data systematically (e.g. distances covered by paper planes with different characteristics), organize data, and represent them graphically (e.g. by histograms or scatter plots).

[8] I use the following definition: Two items, I_1 and I_2, belong to different kinds of math competencies if learning to solve I_1 does not entail learning to solve I_2 (or vice versa). Of course, the definition presupposes an institutional setting in which the learning takes place.

> In order to analyze data, student of that age should be able to choose and apply suitable statistical methods (e.g. means or variance). This includes, for example, making conjectures on possible correlations between characteristics of a sample that are based on scatter plots. (p. 91)

Obviously, a test suitable for grade 9 should contain several items which students in grade 5 have not yet learned to solve.[9] One can assume that the same is true for grade 7 (which is not explicitly discussed by the authors).

3.2 Reference to a longitudinally valid test

In order to quantify the development of math competencies during educational processes one needs an approach which can satisfy two requirements:

a) The approach should be compatible with the fact that during educational processes students acquire new kinds of mathematical competence.

b) A comparison of competencies at two occasions should be based on the idea that, ideally, results of *the same test* should be compared.

Principle (9) provides a basis. For the present application, one has to think of a test which includes *all* items assumed to be relevant for an assessment of math competencies at grades 5 *and* 7. This test, subsequently called T*, consists of all items which are part of T1 or T2.

If this test would have been administered at both grades, one not only could convincingly argue that the same measurement instrument was used, but also could use simple sum scores for a quantification of competence changes. I therefore suggest that this test is considered as the relevant reference.

This leads to the question: How to use the results of T1 and T2 for estimating students' responses to the items not administered, as indicated by the question marks in the following graphic (the symbols $s^x_{i,t}$ denote the respective sum scores).

	Grade 5	Grade 7
Items not in T2	$s^a_{i,1}$	$s^a_{i,2}$?
Items in T1 and T2	$s^c_{i,1}$	$s^c_{i,2}$
Items not in T1	$s^b_{i,1}$?	$s^b_{i,2}$

$s^a_{i,2}$ denotes a student's sum score which he or she would have achieved at grade 7 from the items of T1 which actually were not presented. For a rough estimate one might assume that

[9] This agrees with thinking of a "cumulative development of competencies across educational stages" (Weinert et al. 2011: 68).

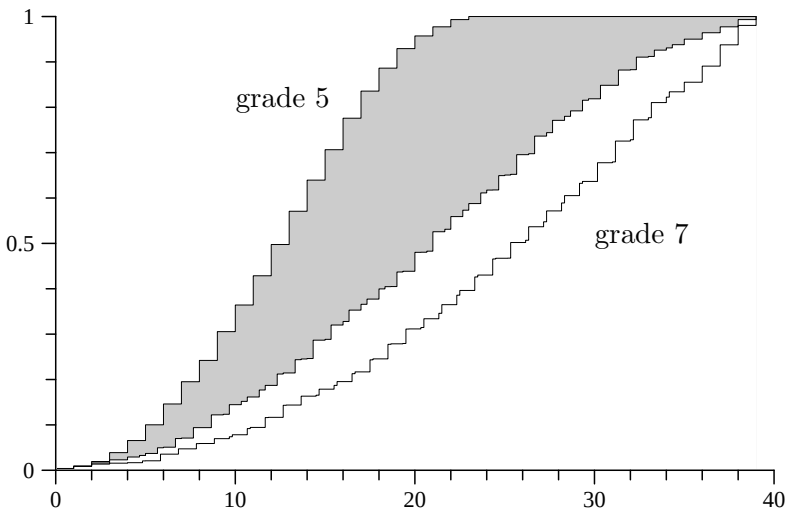


Fig. 4 Distribution functions of $s^*_{i,2}$ and the interval-valued sum score $s^*_{i,1}$ (grey-scaled).

the proportion of correctly answered items equals the corresponding proportion for the observed items: $s^a_{i,2} \approx s^c_{i,2}\, 17/6$.[10]

Without knowledge about the meaning of items not administered at grade 5 one cannot make any specific assumptions about $s^b_{i,1}$. However, one can consider the following interval: $0 \leq S^b_{i,1} \leq s^c_{i,1}\, 16/6$. The lower limit conforms to the assumption that students at grade 5 have not yet learned to solve items from T2 which were not presented to them. The upper limit conforms to the assumption that one can use the presented items from T2 as a basis for extrapolation.

Let $s^*_{i,1}$ and $s^*_{i,2}$ denote estimates of the sum scores in T* in grade 5 and grade 7, respectively. Using the rough estimation approach just depicted, one gets:

$$s^*_{i,1} \in [\, s^a_{i,1} + s^c_{i,1}, s^a_{i,1} + s^c_{i,1}\, 22/6\,]$$
$$s^*_{i,2} \approx s^b_{i,2} + s^c_{i,2}\, 23/6$$

Figure 4 shows distribution functions of these sum scores. For grade 5, it is the distribution function of an interval-valued variable.[11] Although there is a large range of uncertainty, one might conclude that math competencies, as defined by T*, have increased. Based on knowledge of the meaning of the items in T*, it might be possible to reduce the range of indeterminacy.

[10] This ignores the different item difficulties. It is an open question, however, whether a joint Rasch model would provide better estimates. Note that, in the present application, equating with a Rasch model that only uses the linking items and the remaining items from T1 will not fit.

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

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.

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.

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

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

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

Pohl, S., Carstensen, C. H. (2013). Scaling of Competence Tests in the National Educational Panel Study – Many Questions, Some Answers, and Further Challenges. *Journal for Educational Research Online* 5, 189–216.

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

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.

von Davier, M., von Davier, A. A. (2007). A Unified Approach to IRT Scale Linking and Scale Transformations. *European Journal of Research Methods for the Behavioral and Social Sciences* 3, 115–124.

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

Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of Competencies Across the Life Span. In: H.-P. Blossfeld, H.-G. Rossbach, J. von Maurice (Eds.), Education as a Lifelong Process. The German National Educational Panel Study (NEPS). *Zeitschrift für Erziehungswissenschaft*, Special Issue 14, 67–86.