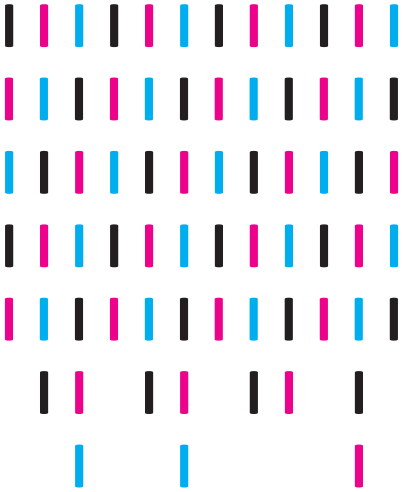

NEPS SURVEY PAPERS

Luise Fischer, Theresa Rohm, Timo Gnambs, &

Claus H. Carstensen

LINKING THE DATA OF THE COMPETENCE TESTS

NEPS Survey Paper No. 1
Bamberg, April 2016

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

NEPS
National Educational Panel Study

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

The NEPS Survey Paper Series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

The NEPS Survey Papers are edited by a review board consisting of the scientific management of LIfBi and NEPS.

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS Survey Papers are available at https://www.neps-data.de (see section “Publications“).

Editor-in-Chief: Corinna Kleinert, LIfBi/University of Bamberg/IAB Nuremberg

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

Linking the Data of the Competence Tests

Luise Fischer, Theresa Rohm, Timo Gnambs, & Claus H. Carstensen

Leibniz Institute for Educational Trajectories, Bamberg, Germany

E-mail address of lead author:

luise.fischer@lifbi.de

Bibliographic data:

Fischer, L., Rohm, T., Gnambs, T., & Carstensen, C. H. (2016). *Linking the data of the competence tests* (NEPS Survey Paper No. 1). Bamberg: Leibniz Institute for Educational Trajectories, National Educational Panel Study. doi:10.5157/NEPS:SP01:1.0

Linking the Data of the Competence Tests

Abstract

The National Educational Panel Study (NEPS) aims at investigating the development of competencies across the whole life span and developing tests for assessing different competence domains. In order to compare competencies across different measurement occasions and examine competence development over time the different measurements must be placed onto a common scale. This goal is achieved by linking two measurements of the same construct. The present document describes the linking procedure adopted for the NEPS. Moreover, this approach is demonstrated using data from Starting Cohort 3 that links mathematical and reading competences across Grades 5 and 7. First, the procedure on how to derive linked item parameters in an anchor-items design is described. After showing that the mathematical tests administered in Grades 5 and 7 are unidimensional, it is demonstrated that all common items administered in both grades showed negligible differential item functioning. Therefore, the two mathematical tests were linked using six measurement invariant items. Subsequently, the procedure on how to derive linked item parameters in an anchor-groups design is described. It is demonstrated that the reading competence tests administered in Grades 5 and 7 were unidimensional and showed no differential item functioning. Therefore, the two reading tests were linked using responses from an independent link sample. Finally, the naming conventions in the scientific use file for repeatedly administered items and different competence scores are presented.

Keywords

link design, link method, item response theory, longitudinal

1. Rationale for Linking

Within the National Educational Panel Study (NEPS; Blossfeld et al., 2011) different competences (e.g., reading, mathematics) are measured across the life span. The respective competence tests are constructed in such a way as to allow for an accurate assessment of these competences within each age group. Therefore, respondents from different age groups typically do not receive identical tests. For example, a test that is optimally challenging for students is likely to be too difficult for adolescents; similarly, a test targeted at adolescents is probably too easy for students. Rather, the NEPS administers different tests to participants that are specifically targeted at their age and competence level. As a consequence, the competence scores from different measurement occasions cannot be directly compared; differences in observed scores would reflect differences in competences as well as differences in item difficulties. In order to examine developmental trajectories and compare competences across different measurement occasions, the different measurements must be placed onto a common scale. This goal is achieved by linking two measurements of the same construct. Thus, linking is a necessary prerequisite to compare the competence data in the NEPS across the life course and investigate educational trajectories (Pohl, Haberkorn, & Carstensen, 2015).

The present study describes the linking procedure adopted for the NEPS. Moreover, this approach is demonstrated using data from Starting Cohort 3 that links mathematical and reading competences across Grades 5 and 7.

2. NEPS Linking Procedure

Specific test designs and statistical procedures are needed to place different measurements on a common scale. In the NEPS two different test designs, the *anchor-items design* and the *anchor-group design*, with their corresponding link methods are used.

2.1 Link designs

In order to link competence scores from two tests common information on the two tests is needed; that is, the same respondents must provide answers to at least a subset of items from both tests. This common information can subsequently be used to place the two measurements on a common scale. In the NEPS competence tests for, among others, the domains reading, mathematics, scientific literacy, information and communication technologies (ICT), and English are administered. Because memory effects are to be expected in some domains if the same item is administered repeatedly to the same respondents (see Pohl et al., 2015), two different link designs are adopted to build an overlap of information between different measurement points.

2.1.1 Anchor-items design

For items measuring mathematical competence in secondary education in NEPS no memory effects are to be expected, since the items are similar to the tasks typically used in schools and the time interval between assessments is rather long with two years. Therefore, it is feasible to include a subset of items from a previously administered test in the test administered at a subsequent measurement occasion. If various assumptions are met (see section 2.3), these “common items” that are included in both tests can be used to link the two tests and create a common scale.

2.1.2 Anchor-group design

For the competence domains of reading, science literacy, and ICT an anchor-group design is needed because memory effects might distort responses if the same items are repeatedly administered to the same respondents. Therefore, common information on two tests is created using an independent link sample that is not part of the original sample whose responses are to be linked. This link sample is drawn from the same population as the respective starting cohort. In the NEPS Linking studies, the age of the link sample either matches the participants' age taking the latter test or falls somewhere between the age groups of the two measurement occasions. Both tests are administered the link sample within a single measurement occasion. Therefore, no developmental changes can occur that might influence the relationship between the two tests.

2.2 Link method

Because Scientific Use Files (SUFs) are published sequentially, the data of a former measurement point is already scaled and released, when data of a latter measurement point is available for scaling. In order to leave the earlier released reference scale unchanged, the data of subsequent measurement points are linked to that initial scale. Leaving the reference scale unchanged imposes some restrictions on potential link methods: If the data from both measurement points were scaled within a single analysis (i.e., "concurrent calibration"; Kolen & Brennan, 2014), the reference scale would change. Moreover, in the NEPS competence tests are scaled using models of the Rasch family (e.g., the Partial Credit model; PCM; Masters, 1982). Therefore, link methods that change the variance of item difficulties and person abilities (e.g., "mean/sigma linking"; Marco, 1977) would risk biasing the measurement of competence development. Therefore, these approaches were not taken into account. In a series of preliminary studies different procedures were evaluated in terms of their suitability for linking different competence domains, such as reading, science, or mathematics, across multiple measurement occasions (cf. Fischer, Rohm, & Carstensen, 2015a, 2015b). Based on these analyses the method of "mean/mean linking" (Loyd & Hoover, 1980) was adopted for the NEPS.

2.2.1 Linking in the anchor-items design

For competence domains using an anchor-items design two independently scaled tests (i.e., the mathematics tests at the first and the second measurement occasion) are linked using a linear transformation of the item parameters of the second test. To link and, therefore, map the latter scale to its reference scale it is assumed that the items' difficulties are the same in both assessments and that all changes in response frequencies can be attributed to change of the person competences over time. To set the item difficulties equal between assessments, a correction term c is derived using the common items that were included in both tests. After independent calibrations of both assessments, the correction term c is derived based on those respondents that participated at both measurement occasions (also referred to as the longitudinal subsample). The correction term c is computed as the difference of the mean of the item difficulty parameters at the first measurement occasion, $M(\sigma_{1j})$, and the mean of the item difficulty parameters at the second measurement occasion, $M(\sigma_{2j})$:

$$c = M(\sigma_{1j}) - M(\sigma_{2j}) \quad (1)$$

The correction term c is then added to each item difficulty parameter of the test to be linked. This approach links two Rasch model calibrated assessments in an anchor-items design using the "mean/mean" method (Loyd & Hoover, 1980). Because "mean/mean" linking depends upon the differences in difficulties between the common items, the choice of the common items will influence the link result to some degree; that is, another set of common items could result in a slightly different correction term c. This uncertainty in the link due to the sampling of common items is reflected by the link error. The link error is based on the differences between the linked item parameters for the k common items at the first and the second measurement occasion as $\Delta\sigma_j = \sigma_{1j} - (\sigma_{2j} + c)$. Following PISA 2006 (OECD, 2014) the link error is then given as the standard deviation of these differences, $SD(\Delta\sigma_j)$, standardized at the number of common items:

$$link\ error\ = \frac{SD(\Delta\sigma_j)}{\sqrt{k}} \tag{2}$$

2.2.2 Linking in the anchor-group design

For domains using an anchor-group design two independently scaled tests are also linked using a linear transformation of the item parameters in the second test. However, because no common items are included in the two tests, the correction term c is derived using the responses from an independent link sample. Let A and B represent the items from the two tests that were administered at the first respectively the second measurement occasion. In the main sample (i.e., the respective starting cohort) A and B are scaled independently at the two measurement occasions. Again, only the longitudinal subsample that participated at both measurement occasions is included in these analyses. In contrast, in the link sample that took both tests all items, A and B, are scaled concurrently; thus, all items are included in a single scaling model. Subsequently, means of the item difficulty parameters for A and B are computed in the main sample and the link sample, that is, (a) the mean difficulty parameters for A in the main sample, $M(\sigma_{MS,A})$, (b) the mean item difficulty parameters for B in the main sample, $M(\sigma_{MS,B})$, (c) the mean item difficulty parameters for A in the link sample, $M(\sigma_{LS,A})$, and (d) the mean item difficulty parameters for B in the link sample, $M(\sigma_{LS,B})$. Then, the correction term c is computed as

$$c = M(\sigma_{MS,A}) - M(\sigma_{MS,B}) + M(\sigma_{LS,B}) - M(\sigma_{LS,A}). \tag{3}$$

The first two terms in (3) insert the reference scale to the test that is to be linked, whereas the last two terms add the information from the link study. The latter reflects the relation between both tests that is not affected by time. The correction term c is then added to each item difficulty parameter of the test intended to be linked. Again, the uncertainty in the link due to the sampling of items is quantified by the link error. In an anchor-group design, the link error is calculated as the pooled link error for the k_A items in A and the k_B items in B. Thus, two sets of differences are derived: (a) the differences between the item parameters of A in the main sample and the link sample, $\Delta\sigma_j = \sigma_{MS,j} - \sigma_{LS,j}$ and (b) the differences between the linked item parameters of B in the main sample and the link sample, $\Delta\sigma_i = (\sigma_{MS,i} + c) - \sigma_{LS,i}$. The link error is then computed as:

$$link\ error\ = \sqrt{\left(\frac{SD(\Delta\sigma_j)}{\sqrt{k_A}}\right)^2 + \left(\frac{SD(\Delta\sigma_i)}{\sqrt{k_B}}\right)^2}. \quad (4)$$

2.3 Assumptions for linking two scales

2.3.1 Unidimensionality

To measure competence development within a domain the meaning of the underlying construct must not change over time. As a consequence, two tests that are to be linked need to be unidimensional. Unidimensionality is examined in two different ways:

In case of an anchor-items design the common items are used as an anchor test. Therefore, at each measurement occasion a one- and a two-dimensional model may be compared. For the two-dimensional model, the common items load on the first dimension and the unique items (i.e., the items included in only one test) load on the second dimension. If a model fit evaluation (i.e. Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC)) supports the one-dimensional model, unidimensionality can be assumed. Moreover, the residuals of the one-dimensional model should exhibit approximately zero-order correlations as indicated by Yen's (1984) Q_3. In case of an anchor-group design unidimensionality is evaluated with regard to the two tests administered to the link sample. Again, information criteria as well as residuals can be inspected.

2.3.2 Measurement invariance

Common items that are supposed to link two tests must exhibit measurement invariance. If a common item is not measurement invariant, it cannot be used to link two tests. An item is considered measurement invariant, when its relative position among the other items on the logit scale does not change between tests.

Using an anchor-items design, measurement invariance for a common item *j* is examined by comparing the item difficulty parameter resulting from a separate scaling of the first measurement occasion, σ_{1j}, and the parameter from the second measurement occasion, σ_{2j}. Because traditional significance tests yield an excessive power with large samples even for negligible effects, we adopted the DIF classification system of the Educational Testing Service (ETS; Holland & Wainer, 1993) that relies on an effect size in the delta metric and a significance test. The delta metric has a mean of 13 and a standard deviation of 4. Thus, one point on the delta scale corresponds to a quarter of a standard deviation. This is equivalent to Cohen's $d = 0.25$ and to $\eta^2 = 0.0154$. Following the ETS classification system items that showed a significant DIF effect greater than 1 point on the delta scale were classified as having moderate DIF. Instead of a classical null hypothesis test of no DIF, we adopted Murphy and Myors' (1999) minimum effect null hypothesis and tested for the presence of negligible DIF. Thus, a value of 1.54 percent of variance explained was used as criterion for negligible DIF. Following Lord (1980) we computed a Wald statistic for each item as

$$t_j = \frac{\sigma_{1j} - \sigma_{2j}}{\sqrt{SE(\sigma_{1j})^2 + SE(\sigma_{2j})^2}}. \quad (5)$$

The resulting t value was squared and, thus, transformed into an F distribution (with df_1 = number of measurement points and df_2 = number of participants). Adopting a value of 1.54 percent of variance explained as a minimum effect criterion, a non-central F test was used to test the assumption of non-negligible DIF (see Fischer, Gnambs, Rohm, & Carstensen, 2016).

In anchor-group designs all items can be considered common items. Therefore, measurement invariance is tested the same way as described above by comparing the item difficulty parameters from the main sample (i.e., the respective starting cohort) and the link sample. Thus, examining measurement invariance in an anchor-group design is identical to examining differential item functioning (DIF) between the two groups.

Because the link information is derived on a subsample of respondents that participated at both measurement occasions (i.e., the longitudinal subsample), it is necessary that all items exhibit measurement invariance across the whole sample, that is, between respondents that participated at only one measurement occasion and the longitudinal subsample that provided data at both measurement occasions. This is tested by estimating the item parameters separately in each subsample and examining measurement invariance as described above.

2.3.2.1 Naming Conventions

In the SUFs repeatedly administered items retain their initial variable names. To identify common items across measurement occasions, a suffix is appended to the variable name that indicates the current sample. For example: Item “mag5q301_c” was first administered in Grade 5. Because the same item was also administered in Grade 7, the variable name for the second administration is extended by the suffix ”sc3g7” (= Starting Cohort 3, Grade 7) to “mag5q301_sc3g7_c”.

3. Linking of Starting Cohort 3 – Grade 5 and Grade 7

In the following section we demonstrate in two examples how to apply the linking procedures described above to the two link designs. First, we apply the method of “mean/mean” linking to two tests on mathematical competence measured in an anchor-items design. Then, the linking procedure in an anchor-items design is illustrated using two tests on reading competence. In Starting Cohort 3, mathematical and reading competences were measured in Grade 5 and two years later in Grade 7. First, the data were scaled independently for each grade following the NEPS scaling procedure described in Pohl and Carstensen (2012). Subsequently, the competence scores were linked across the two grades whereby the scale of Grade 5 served as the underlying reference scale for Grade 7. Mathematical competences were linked using an anchor-items design, whereas reading competences were linked using an anchor-groups design.

3.1 Linking the test of mathematical competence

The scaling results of the mathematics test in Grade 5 are described in Duchhardt and Gerdes (2012) and the respective results for Grade 7 are outlined in Schnittjer, Gerken, and Fischer (2015).

3.1.1 Sample

A sample of 5,193 participants received the mathematics test in Grade 5 and 6,191 participants finished the test in Grade 7. Overall, 3,833 respondents participated at both measurement occasions. Participants with less than three valid responses were excluded from the linking procedure.

3.1.2 Test instruments

The mathematics test in Grades 5 and 7 included 24 and 23 items, respectively. Six items were included in both tests (see Table 1).

3.1.3 Results

3.1.3.1 Unidimensionality

For each grade, we estimated a one-dimensional model that specified a single latent factor for all items and also a two-dimensional model that specified separate latent factors for the common items and the unique items (i.e., the items that were included at only one measurement occasion). In both grades the information criteria favored the two-dimensional model, AIC = 156983.15 and BIC = 157164.88 for Grade 5, and AIC = 132409.71 and BIC = 132599.81 for Grade 7, over the one-dimensional model, AIC = 157104.52 and BIC = 157272.79 for Grade 5, and AIC = 132466.18 and BIC = 132643.17 for Grade 7. Therefore, we also examined the residual correlations for the one-dimensional models. The corrected Q_3 statistics indicated largely unidimensional scales in Grade 5, $M(Q_3) = 0$, $SD(Q_3) = 0.03$, and Grade 7, $M(Q_3) = 0$, $SD(Q_3) = 0.02$. This indicates that unidimensional scales can be assumed for the mathematics tests in Grades 5 and 7.

3.1.3.2 Measurement invariance

First, the mathematics tests in Grades 5 and 7 were scaled separately in the longitudinal subsample. Subsequently, the difficulty parameters of the common items were centered in each grade (i.e., their means were set to zero). The differences in item difficulties between Grades 5 and 7, and the tests for measurement invariance based on the Wald statistic (see Equation 3) are summarized in Table 1. For the six common items measurement invariance was supported (i.e., the minimum effects hypothesis test was not significant).

Table 1: DIF Analyses for the common items in the tests for mathematical competence in Grades 5 and 7

Grade 5	Grade 7	Δσ	$SE_{\Delta\sigma}$	*F*	*p (α = .05)*
mag5q301_c	mag5q301_sc3g7_c	0.03	0.05	0.24	*> .999*
mag5d051_c	mag5d051_sc3g7_c	-0.31	0.09	12.49	*> .999*
mag5d052_c	mag5d052_sc3g7_c	0.42	0.06	51.36	*0.72*
mag5r251_c	mag5r251_sc3g7_c	-0.12	0.05	5.30	*> .999*
mag5v321_c	mag5v321_sc3g7_c	-0.10	0.06	3.12	*> .999*
mag5r191_c	mag5r191_sc3g7_c	0.09	0.06	2.34	*> .999*

Note. Δσ = Difference in item difficulty parameters between Grades 5 and 7 (positive values indicate easier items in Grade 7); SEΔσ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test based on (5); Fcrit = Critical value for the minimum effects hypothesis test for an α of .05; the degrees of freedom (df1, df2) are based on the number

of measurement points (df1 = k-1) and the number of test takers taking both tests (df2 = n-1). The critical F0154(1, 3,832) = 88.3. A non-significant test indicates measurement invariance.

Furthermore, measurement invariance for Grade 7 was supported for the two groups of one-time participants (n = 2,358) and the longitudinal subsample (n = 3,833), with none of the F-statistics exceeding the critical value of F0154(1, 6,189) = 132.1.

3.1.3.3 Linking the two tests

The mathematics tests administered in the two grades were linked using the "mean/mean" method (see section 2.2.1). In the longitudinal subsample, the mean item difficulty parameters for the six common items were -0.384 in Grade 5 and -1.110 in Grade 7 (see Table 2).

Table 2: Original and linked item difficulty parameters for the mathematics test in Grade 7.

			Item difficulties σ_j	
Item	**Common item**	**Position**	**Original**	**Linked**
mag9q071_c	No	1	-0.364	0.362
mag7v071_c	No	2	0.499	1.225
mag7r081_c	No	3	0.207	0.932
mag7q051_c	No	4	0.287	1.013
mag5q301_sc3g7_c	Yes	5	-0.267	0.459
mag9d151_c	No	6	-1.356	-0.630
mag5d051_sc3g7_c	Yes	7	-3.130	-2.404
mag5d052_sc3g7_c	No	8	-1.828	-1.103
mag9v011_c	No	9	-0.523	0.203
mag9v012_c	No	10	0.243	0.969
mag7q041_c	No	11	-0.648	0.078
mag7d042_c	No	12	-1.828	-1.102
mag7r091_c	No	13	-0.120	0.606
mag9q181_c	No	14	-1.819	-1.093
mag7d011_c	No	15	-1.346	-0.621
mag7v012_c	No	16	-0.150	0.576
mag7v031_c	No	17	-0.395	0.331
mag5r251_sc3g7_c	Yes	18	-0.502	0.224
mag7d061_c	No	19	0.660	1.386
mag5v321_sc3g7_c	Yes	20	0.284	1.010
mag9v091_c	No	21	1.189	1.914

mag5r191_sc3g7_c	Yes	22	-1.217	-0.491
mag7r02s_c	No	23	-1.258	-0.532

Note. Original item difficulty parameters were derived by an independent scaling of the item responses (see Schnittjer et al., 2015). Linked item difficulty parameters were derived by adding c to the original item parameters.

Using Equation (1) this resulted in a correction term of c = -0.384 - (-1.110) = 0.726. The correction term c was added to each item difficulty parameter derived in Grade 7 and, thus, resulted in the linked item parameters (see Table 2). The corresponding link error according to Equation (2) was 0.09 (for more detailed information see Fischer et al., 2016.).

Person abilities were subsequently estimated using the linked item difficulty parameters in Grade 7. In the SUF, manifest scale scores are provided in the form of two different weighted maximum likelihood estimates (WLE; see Pohl & Carstensen, 2012), “mag7_sc1” and “mag7_sc1u”, including their respective standard error, “mag7_sc2” and “mag7_sc2u”. Both WLE scores are linked to the underlying reference scale of Grade 5. The uncorrected score “mag7_sc1u” (uncorrected for the position of the math test within the booklet) can be used, if the research focus lies on longitudinal issues, such as competence development, since the positon of the domains stays the same over subsequent assessment and therefore, resulting differences in WLE scores can be interpreted as development trajectories across measurement points. Conversely, the corrected score “mag7_sc1” is corrected for the position of the math test within the booklet and can thus not be used for longitudinal purposes but for cross-sectional research questions.

3.2 Linking the test of reading competence

The scaling results of the reading competence tests in Grade 5 are presented in Pohl, Haberkorn, Hardt, and Wiegand (2012), whereas the respective results for Grade 7 can be found in Krannich et al. (in prep.). Moreover, assumptions for the linking of reading competences are discussed in Pohl et al. (2015).

3.2.1 Sample

Overall, 5,193 participants were administered the reading test in Grade 5, 6,186 participants took the test in Grade 7, and 3,829 respondents participated at both measurement occasions. Participants with less than three valid responses were excluded from the link procedure.

3.2.2 Test instruments

The reading test in Grade 5 included 32 items whereas the test in Grade 7 consisted of 40 items. Because retest effects are expected for the reading items, no common items could be administered in the two tests. Instead, an overlap of information was accomplished by using an independent link sample including 608 participants attending Grade 7. While participants in the main study took the two reading competence tests with a time-lag of two years between Grade 5 and Grade 7, participants of the link study took both tests at one measurement point in Grade 7. In the link sample the two tests were presented in random order: 309 participants received the test for Grade 5 first and subsequently the Grade 7 test; 299 participants took the Grade 7 test before working on the Grade 5 test. Moreover, because in Grade 7 two different test versions (i.e., an easy and a difficult test) were administered to participants (see Krannich et al., in prep.), the participants in the link sample

were randomly assigned either test version. In the link sample the test was scaled concurrently, whereas in the main sample the tests in Grades 5 and 7 were scaled independently.

3.2.3 Results

3.2.3.1 Unidimensionality

In the link sample we estimated a one-dimensional model that specified a single latent factor for all items and also a two-dimensional model that specified separate latent factors for the two tests. The information criteria slightly favored the two-dimensional model, AIC = 32125.42 and BIC = 32606.13, over the one-dimensional model, AIC = 32158.13 and BIC = 32630.02. Therefore, we also examined the residual correlations of the one-dimensional model. The corrected Q_3 statistics indicated largely unidimensional scales, $M(Q_3) = 0$, $SD(Q_3) = 0.06$. This indicates that unidimensional scales can be assumed for the reading tests in Grades 5 and 7.

3.2.3.2 Measurement invariance

Measurement invariance was examined using the same procedure as used for the mathematics test. We tested whether the item parameters derived in the link sample showed a non-negligible shift in item difficulties as compared to the longitudinal subsample from the main sample. The respective results are summarized in Tables 4 and 5 in the appendix. The analyses supported measurement invariance for all items.

Furthermore, measurement invariance for Grade 7 was supported for the two groups of one-time participants (n = 2,357) and the longitudinal subsample (n = 3829), with none of the *F*-statistics exceeding the critical value of $F_{0154}(1, 6{,}184) = 131.9$.

3.2.3.3 Linking the two tests

The reading competence tests administered in the two grades were linked using the "mean/mean" method for the anchor-group design (see section 2.2.2). The mean item difficulty parameters in the longitudinal subsample were -1.416 in Grade 5 and -1.293 in Grade 7; in the link sample, the respective mean parameters were -2.321 and -1.531 for Grades 5 and 7. Following equation (3) the correction term was calculated as c = -1.416 – (-1.293) + (-1.531) – (-2.321) = 0.667. The correction term c was added to each difficulty parameter derived in Grade 7 and, thus, resulted in the linked item parameters (see Table 3). The resulting link error according to Equation (4) was 0.07 (for more detailed information see Fischer et al., 2016.).

Table 3: Item difficulty parameters of the linked reading competence test in Grade 7

		Item difficulties σ_i	
Item	**Position**	**Original**	**Linked**
reg70110_c	1	-0,375	0,292
reg70120_c	2	-2,524	-1,856
reg7013s_c	3	-2,594	-1,927
reg70140_c	4	-3,456	-2,789
reg7015s_c	5	-2,940	-2,273
reg7016s_c	6	-1,099	-0,432
reg70210_c	7	-2,792	-2,125
reg70220_c	8	-1,941	-1,274
reg7023s_c	9	-1,932	-1,265
reg7024s_c	10	-0,754	-0,087
reg70250_c	11	-1,003	-0,336
reg7026s_c	12	-1,419	-0,752
reg70310_c	13	-2,629	-1,961
reg70320_c	14	-1,627	-0,960
reg7033s_c	15	-1,215	-0,548
reg70340_c	16	-1,533	-0,866
reg70350_c	17	-2,040	-1,373
reg70360_c	18	-1,252	-0,585
reg70410_c	19	-2,350	-1,683
reg70420_c	20	-1,892	-1,225
reg70430_c	21	-2,403	-1,736
reg70440_c	22	-1,917	-1,249
reg7045s_c	23	-0,469	0,198
reg70460_c	24	0,801	1,468
reg7051s_c	25	-1,963	-1,295
reg70520_c	26	-1,292	-0,625
reg7053s_c	27	-1,164	-0,496
reg7055s_c	28	0,124	0,791
reg70560_c	29	0,522	1,190
reg70610_c	30	-2,847	-2,180

		Item difficulties σ_i	
Item	**Position**	**Original**	**Linked**
reg70620_c	31	-0,613	0,055
reg7063s_c	32	-2,706	-2,039
reg70640_c	33	0,464	1,131
reg70650_c	34	0,229	0,897
reg7066s_c	35	-1,208	-0,541
reg7071s_c	36	-1,482	-0,815
reg70720_c	37	0,918	1,585
reg70730_c	38	0,631	1,299
reg70740_c	39	-0,911	-0,244
reg7075s_c	40	0,318	0,985

Note. Original item difficulty parameters were derived by an independent scaling of the item responses (see Krannich et al., in prep.). Linked item difficulty parameters were derived by adding c to the original item parameters.

Person abilities were subsequently estimated using the linked item difficulty parameters in Grade 7. In the SUF, manifest scale scores are provided in the form of two different WLE estimates, ”reg7_sc1” and “reg7_sc1u”, including their respective standard errors “reg7_sc2” and “reg7_sc2u”. Both WLE scores are linked to the underlying reference scale of Grade 5. The uncorrected score “reg7_sc1u” (uncorrected for the position of the reading test within the booklet) can be used, if the focus of the research lies on longitudinal issues, such as competence development since differences in WLE scores can be interpreted as development trajectories across measurement points. Again, the corrected score “reg7_sc1” was corrected for the position of the reading test within the booklet and can be used, if the research interest lies on cross-sectional issues.

4. Summary

The NEPS repeatedly measures different competences (e.g., reading, mathematics) across the life span. To study competence development and compare competences scores from different measurement occasions, the different scores must be placed onto a common scale. Otherwise changes in competences would be confounded with differences in test difficulties. Therefore, in the NEPS repeatedly measured competences are linked and placed onto a common scale. Depending on the specific competence domain the NEPS uses two different link strategies. Competences such as mathematical competence that are unlikely to be prone to memory effects are linked using an anchor-items design. In contrast, for competences that might be susceptible to memory effects (e.g., reading competence) an anchor-group design is used. After outlining the basics of these link procedures, the present study demonstrated how to link mathematical and reading competences across Grades 5 and 7. It was shown that the two mathematical tests administered in the two grades were essentially unidimensional. Moreover, six items that were included in both tests were measurement invariant and, thus, could be used to link the two tests. Therefore, the item parameters of the mathematical test administered in Grade 7 were linked to the item parameters of the

respective test administered in Grade 5. As a consequence, person abilities estimated using these linked item difficulty parameters were on the same scale as the person abilities derived in Grade 5. These ability estimates can be used for longitudinal comparisons; ability differences using these scores can be interpreted as development trajectories across the two measurement occasions. Subsequently, we also demonstrated that the reading competence tests administered in Grades 5 and 7 could be linked using an anchor-groups design. It was shown that both tests were essentially unidimensional and showed no differential item functioning. Therefore, the item parameters of the reading test administered in Grade 7 were linked to the item parameters of the respective test administered in Grade 5 using the responses of an independent link sample. As a consequence, person abilities estimated using these linked item difficulty parameters were on the same scale as the person abilities derived in Grade 5. These scores can be used to compare ability estimates across the two grades. In conclusion, the paper summarized the two link strategies adopted in the NEPS and showed how to derive ability estimates that can be compared across different measurement occasions.

References

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS). Zeitschrift für Erziehungswissenschaft, 14, 1-4.

Duchhardt, C., & Gerdes, A. (2012). NEPS Technical Report for Mathematics - Scaling Results of Starting Cohort 3 in Fifth Grade (NEPS Working Paper No. 19). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Fischer, L., Rohm, T., & Carstensen, C. H. (2015a, July). Comparing link methods based on their link errors. Poster presented at the International Meeting of the Psychometric Society, Beijing, China.

Fischer, L., Rohm, T., & Carstensen, C. H. (2015b, September). Ein Vergleich verschiedener Linkmethoden und ihrer Linkfehler anhand Rasch skalierter Kompetenzdaten. Paper presented at the 12th meeting of the section Methoden & Evaluation of the DGPs, Jena.

Fischer, L., Gnambs, T., Rohm, T., & Carstensen, C. H. (2016) Evaluating link methods on Rasch scaled longitudinal data in large scale assessments. Unpublished manuscript, Bamberg: Leibniz Institute for Educational Trajectories.

Holland, P. W., & Wainer, H. E. (1993). Differential item functioning. Hillsdale, NJ: Erlbaum.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking. New York, NY: Springer.

Krannich, M., Jost, O., Rohm, T., Koller, I., Carstensen, C. H., & Fischer, L. (in prep.). NEPS Technical Report for Reading – Scaling results of Starting Cohort 3 in seventh grade (NEPS Working Paper). Bamberg: Leibniz Institute for Educational Trajectories.

Lord F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. Journal of Educational Measurement, 17, 179-193.

Marco, G. L. (1977). Item characteristic curve solutions to the three intractable testing problems. Journal of Educational Measurement, 16, 139-160.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Murphy, K. R., & Myors, B. (1999). Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. Journal of Applied Psychology, 84, 234-248.

Organisation for Economic Cooperation and Development (OECD). (2014). PISA 2012 Technical Report. Paris, France: OECD Publishing. Retrieved from https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

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.

Pohl, S., Haberkorn, C., & Carstensen, C. H. (2015). Measuring competencies across the lifespan - Challenges of linking test scores. In M. Stemmler, A. von Eye, & W. Wiedermann (Eds), Dependent data in social science research (pp. 281-308). Berlin, Germany: Springer.

Pohl, S., Haberkorn, K., Hardt, K., & Wiegand, E. (2012). NEPS Technical Report for Reading – Scaling results of Starting Cohort 3 in fifth grade (NEPS Working Paper No. 15). Bamberg: Otto-Friedrich-Universität, Nationales Bildungspanel.

Schnittjer, I., Gerken, A., & Fischer., L. (2015). NEPS Technical Report for Mathematics – Scaling results of Starting Cohort 3 in seventh grade (NEPS Working Paper No. XX).

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of thethree-parameter logistic model. Applied Psychological Measurement, 8, 125-145.

Appendix

Table 4: DIF Analyses for Reading Competence between the longitudinal subsample in Grade 5 and the Link Sample (LS).

Item	Δσ	$SE_{\Delta\sigma}$	*F*
reg50110_c	0.003	0.302	0.00
reg5012s_c	-0.372	0.274	1.84
reg50130_c	0.006	0.189	0.00
reg50140_c	-0.047	0.156	0.09
reg50150_c	-0.037	0.126	0.09
reg5016s_c	-0.112	0.143	0.61
reg50170_c	0.426	0.114	13.89
reg50210_c	-0.108	0.237	0.21
reg50220_c	-0.385	0.115	11.15
reg50230_c	0.219	0.239	0.84
reg50240_c	0.008	0.153	0.00
reg50250_c	-0.312	0.126	6.14
reg5026s_c	0.145	0.115	1.61
reg50310_c	0.166	0.211	0.62
reg50320_c	0.249	0.275	0.82
reg50330_c	-0.148	0.215	0.47
reg50340_c	0.315	0.175	3.25
reg50350_c	-0.251	0.124	4.08
reg50360_c	0.673	0.265	6.44
reg50370_c	-0.238	0.142	2.81
reg50410_c	0.267	0.135	3.90
reg5042s_c	-0.292	0.196	2.21
reg50430_c	0.218	0.119	3.39
reg50440_c	0.063	0.120	0.27
reg50460_c	0.102	0.130	0.61
reg50510_c	0.165	0.239	0.48
reg5052s_c	-0.106	0.184	0.33
reg50530_c	-0.461	0.132	12.27
reg50540_c	0.086	0.182	0.22
reg5055s_c	0.343	0.211	2.64

Item	Δσ	$SE_{\Delta\sigma}$	*F*
reg50560_c	-0.256	0.142	3.23
reg50570_c	-0.347	0.153	5.14

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in Grade 5 and the link sample (positive values indicate easier items in the link sample); $SE_{\Delta\sigma}$ = Pooled standard error; *F* = Test statistic for the minimum effects hypothesis test based on(5). the critical value for the minimum effects hypothesis test using an α of .05 is F_{0154} (1, 4,435) = 99.7. A non-significant test indicates measurement invariance.

Table 5: DIF Analyses for Reading Competence between the longitudinal subsample in Grade 7 and the Link Sample (LS).

Item	Δσ	$SE_{\Delta\sigma}$	*F*
reg70110_c	-0.742	0.152	23.87
reg70120_c	-0.690	0.199	12.02
reg7013s_c	0.228	0.292	0.61
reg70140_c	0.382	0.382	1.00
reg7015s_c	-0.331	0.246	1.81
reg7016s_c	0.120	0.188	0.41
reg70210_c	-0.123	0.168	0.54
reg70220_c	0.031	0.144	0.05
reg7023s_c	0.128	0.170	0.57
reg7024s_c	0.348	0.144	5.80
reg70250_c	-0.106	0.122	0.76
reg7026s_c	-0.025	0.167	0.02
reg70310_c	0.416	0.199	4.38
reg70320_c	0.053	0.137	0.15
reg7033s_c	-0.088	0.138	0.41
reg70340_c	-0.006	0.135	0.00
reg70350_c	-0.078	0.150	0.27
reg70360_c	-0.186	0.128	2.09
reg70410_c	0.091	0.169	0.29
reg70420_c	-0.063	0.149	0.18
reg70430_c	0.114	0.178	0.41
reg70440_c	0.021	0.157	0.02
reg7045s_c	0.321	0.134	5.79
reg70460_c	-0.060	0.131	0.21
reg7051s_c	0.832	0.313	7.07
reg70520_c	-0.046	0.199	0.05
reg7053s_c	0.019	0.220	0.01
reg7055s_c	0.350	0.173	4.08
reg70560_c	-0.701	0.186	14.25
reg70610_c	-0.762	0.228	11.15
reg70620_c	0.294	0.159	3.42
reg7063s_c	-0.440	0.240	3.36

Item	Δσ	$SE_{\Delta\sigma}$	*F*
reg70640_c	0.145	0.153	0.89
reg70650_c	-0.042	0.153	0.08
reg7066s_c	-0.265	0.153	3.02
reg7071s_c	0.540	0.256	4.43
reg70720_c	0.228	0.196	1.36
reg70730_c	-0.019	0.198	0.01
reg70740_c	-0.351	0.202	3.01
reg7075s_c	0.463	0.197	5.50

Note. Δσ = Difference in item difficulty parameters between the longitudinal subsample in Grade 7 and the link sample (positive values indicate easier items in the link sample); SEΔσ = Pooled standard error; F = Test statistic for the minimum effects hypothesis test based on(5). The critical value for the minimum effects hypothesis test using an α of .05 is F0154 (1, 4,435) = 99.7. A non-significant test indicates measurement invariance.