Psychological Test and Assessment Modeling, Volume 61, 2019 (1), 37-64

Longitudinal linking of Rasch-model-scaled competence tests in large-scale assessments: A comparison and evaluation of different linking methods and anchoring designs based on two tests on mathematical competence administered in grades 5 and 7

Luise Fischer[1,2]*, Timo Gnambs*[1,3]*, Theresa Rohm*[1,2] *&*
Claus H. Carstensen[2]

Abstract

Measuring growth in an item response theory framework requires aligning two tests on a common scale known as longitudinal linking. So far, no consensus exists regarding the appropriate method for the linking of longitudinal data scaled according to the Rasch model in large-scale assessments. Therefore, an empirical study was conducted within the German National Educational Panel Study to identify appropriate linking methods for the comparison of competencies across time. The study examined two anchoring designs based either on anchor-items or an anchor-group and three linking methods (mean/mean linking, fixed parameters calibration, and concurrent calibration). Two tests on mathematical competence were administered to a sample of n = 3,833 German students (48 % girls) in Grades 5 and 7. An independent link sample (n = 581, 53 % girls) drawn from the same population was administered both tests at the same time. The assumptions of unidimensionality were confirmed; differential item functioning was examined using effect-based hypotheses tests. Anchoring designs and linking methods were compared and evaluated using diverse criteria such as link error, mean growth rate estimation, and model fit. Overall, little differences among the linking methods and anchoring designs were found. However, mean growth was found to be significantly smaller in the anchor-group design.

Keywords: linking, item response theory, longitudinal, effect based hypotheses testing, competences

[1] *Correspondence concerning this article should be addressed to:* Luise Fischer, Educational Measurement, Leibniz Institute for Educational Trajectories, Wilhelmsplatz 3, 96047 Bamberg, Germany; email: luise.fischer@uni-bamberg.de

[2] Department of Psychology and Methods of Educational Research, University of Bamberg, Bamberg, Germany

[3] Institute for Education and Psychology, Johannes Kepler University Linz, Austria

Introduction

The measurement of an individual's growth in an item response theory (IRT) framework requires placing two tests on a common scale. This is referred to as longitudinal linking (A. von Davier, Carstensen, & M. von Davier, 2006). Therefore, linking data is an essential prerequisite for investigating educational trajectories. Most large-scale assessments (LSA) focus on differences between age cohorts such as the *Programme of International Student Assessment* (PISA), the *Trends in International Mathematics and Science Study* (TIMMS), the *Progress in International Reading Literacy Study* (PIRLS), or the American *National Assessment of Educational Progress* (NAEP). Only few LSAs allow for the study of an individual's change over time, for example, the German *National Educational Panel Study* (NEPS; Blossfeld, Roßbach, & von Maurice, 2011), the American *Early Childhood Longitudinal Program* (ECLS), or longitudinal extensions of PISA (e.g., Prenzel, Carstensen, Schöps, & Maurischat, 2006). Furthermore, unidimensional Rasch-type models as well as the more complex two-parametric and three-parametric logistic models (2PL and 3PL; Birnbaum, 1968) are used in the practice of vertical scaling of educational assessments (A. von Davier et al., 2006). However, the latter models that additionally include discrimination and guessing parameters are clearly more popular; other model parameterizations such as the difficulty-plus-guessing model (Kubinger & Draxler, 2006) have also been introduced but, as of yet, have not been frequently used in LSAs. So far, no consensus exists regarding the appropriate method for the linking of longitudinal data scaled according to the Rasch model in large-scale assessments. This study investigated whether certain linking methods, usually applied in 2PL or 3PL modeled cross-sectional data, can be transferred to Rasch-model-scaled longitudinal data. Moreover, these linking methods were compared and evaluated in different anchoring designs (anchor-items design and anchor-group design) using data from the NEPS.

Linking of Rasch-type models

In Rasch-type models it is assumed that the probability *P* of person *n* to correctly answer item *i* is conditioned on the interaction of two parameters (both of them being necessary and sufficient), that is, a person's ability β (which is not directly observable and, therefore, latent) and an item's difficulty parameter *δ*. In order to model ordered response categories in polytomous items, Masters (1982) developed a partial credit model (PCM):

$$P\left(X_{kni}=1|\beta_n,\delta_{ik}\right)=\frac{\exp\left(\beta_n-\delta_{ik}\right)}{1+\exp\left(\beta_n-\delta_{ik}\right)}, \tag{1}$$

where δ_{ik} is the difficulty of the k^{th} step in item *i.* In the special case of dichotomous data, the PCM reduces to the well-known Rasch model (Rasch, 1980):

$$P\left(X_{ni}=1|\beta_n,\delta_i\right)=\frac{\exp\left(\beta_n-\delta_i\right)}{1+\exp\left(\beta_n-\delta_i\right)}. \tag{2}$$

In Rasch-type models the person ability parameter β_n and the item difficulty parameter δ_i are both localized on a common "ability" scale. As the zero in this ability scale is set arbitrarily (i.e., depending on the parameter constraints), any statement on the change of a person's ability over a period of time needs to be based on data that is longitudinally linked (van der Linden & Barrett, 2016).

Anchoring designs

Due to a time-lag between test administrations in longitudinal educational assessments accompanied by a corresponding ability development, ability distributions will most likely differ when assessing the same sample repeatedly. Therefore, participants are regarded as non-equivalent groups in repeated measurements (A. von Davier et al., 2006). Thus, the procedure of linking longitudinal data requires an overlap of information between the two tests (Pohl, Haberkorn, & Carstensen, 2015). This information overlap is either achieved by identical items (i.e., common items) administered at both measurement points (anchor-items design) or by persons who answer items from both tests at the same measurement point (anchor-group design; Vale, 1986; see Figure 1). If a common item in an anchor-items design meets several conditions (see below), it can serve as a link item. A. von Davier and colleagues (2006) refer to the same design in a cross-sectional context as a design for non-equivalent groups with anchor test (NEAT). Linking in the NEAT design is generally referred to as vertical linking (A. von Davier et al., 2006). Though longitudinal linking and vertical linking differ in name depending on the adopted data collection design (longitudinal versus cross-sectional), they do not differ conceptually, when the samples are non-equivalent groups. As such, both approaches are concerned with practical issues when it comes to linking (e.g., Seock-Ho Kim & A. Cohen, 1992; Kolen & Brennan, 2014). However, linking based on longitudinal designs

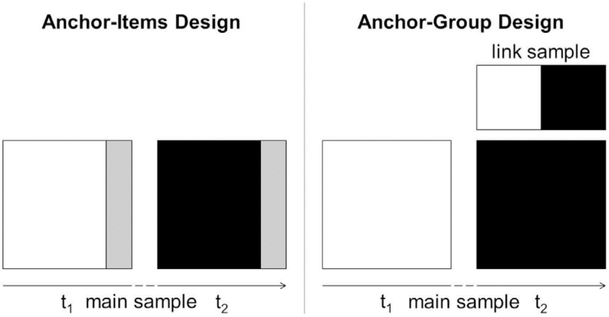

Figure 1:
Anchoring Designs for Longitudinal Linking. Each rectangle constitutes one measurement point. While white and black rectangles represent test specific items, grey rectangles represent common items between the tests. t1 = first measurement point; t2 = second measurement point

faces additional challenges due to participants' motivations and panel dropout that require reduced test lengths, potentially leading to a decreased accuracy in parameter estimation. As such, the absolute number in link items as well as the link items' estimation accuracy may (drastically) differ among longitudinal and vertical linking.

For the anchor-group design an overlap of information is achieved through an independent link sample. Participants of the link sample need to be sampled from the same population as the main sample (i.e., the sample a researcher is primarily interested in). The mean age of the link sample should correspond to the age of the main sample at t_1 or t_2 or should fall somewhere between the age groups of the two measurement occasions (Pohl et al., 2015). Thus, the link sample is administered both tests at the same measurement point. Therefore, the participant's answers on the items are not influenced by longitudinal ability development. Thus, the item difficulty parameters represent an unaltered relationship of the two tests. In this anchoring design, no common items are necessary.

With respect to our question concerning longitudinal designs where the same participants are assessed repeatedly, the choice of an anchoring design depends, amongst others, on test length, potential memory effects as well as repetition effects in link items and the expected amount of change in the latent construct between two measurements affecting item difficulty. Domains using content-based items (e.g., reading literacy) are more prone to memory and repetition effects than domains using number- and operation-based items (e.g., mathematical literacy). Although an anchor-group design causes additional costs and the resulting link information is afflicted with an additional sampling error, it may still be the preferable choice depending on the measured construct. In terms of the validity of a link (i.e., the extent to which the link information represents the various facets of the underlying construct) the anchor-group design is more comprehensive than the anchor-items design since all items (i.e., both tests completely) contribute to the link information. Also, when test length is limited and change in the latent ability is expected to be large, an anchor-group design may be preferable regarding the reliability of each measurement point. As reliability is increasing the more closely the test difficulty and the person ability distribution match, no item position is occupied by a common item that potentially provides only little information due to fitting the ability of the sample poorly at the second measurement point.

IRT linking methods

A linking method translates the link information in order to put the parameters from two (or more) tests on a common scale (Vale, 1986). The selection of a linking method is codetermined by the anchoring design: While an anchor-group design is less common in practice and thus, only a small number of corresponding linking methods have been suggested, a wide selection of established linking methods is available for the anchor-items design / NEAT design (see M. von Davier and Carstensen, 2006 for an overview and Kolen & Brennan, 2014, for an elaborated introduction). Some of the most popular IRT linking methods are the mean/mean method (Loyd & Hoover, 1980), mean/sigma method (Marco, 1977), the characteristic curve methods (Haebara, 1980; Stocking & Lord, 1983), fixed parameters calibration, and concurrent calibration. Also hybrid approaches such as combining concurrent calibration with fixed parameters calibration (e.g., PISA cycle of 2015; Organisation for Economic Co-operation and Development (OECD), 2017) or characteristic curve methods (Briggs & Weeks, 2009) have been used in LSAs. All but the concurrent calibration use separate calibrations in each sample before transforming the item and person parameters. Thus, an already established scale (e.g., the scale from the first measurement point) serves as a reference scale. This may be attractive for example in longitudinal designs where the focus lies on measuring change and a reference scale has already been implemented due to sequentially published data. The following section describes three IRT linking methods: mean/mean linking, fixed parameters calibration, and concurrent calibration that are applicable in Rasch-type models (i.e., discrimination parameters retain their fixed value of one) and thus, were examined in this study in more detail.

Mean/mean method based on the anchor-items design (m/m$_{\text{AID}}$)

In this method the item difficulty parameter estimates δ_i (Rasch, 1980) of the link items are used for computing two scaling constants, slope *A* and intercept *B,* to shift scale *Y* to the reference scale *X* (Loyd & Hoover, 1980):

$$\delta_Y^* = \mathrm{A}\delta_Y + \mathrm{B} \tag{3}$$

The scaling constants based on the anchor-items design (AID) are computed from the link items as

$$A_{\text{AID}} = M\left(\alpha_{Y\text{link}}\right)/M\left(\alpha_{X\text{link}}\right) \tag{4}$$

with $M(\alpha_{Y\text{link}})$ and $M(\alpha_{X\text{link}})$ being the means of the link item discrimination parameters from scales *Y* and *X* and as

$$B_{\text{AID}} = M\left(\delta_{Xlink}\right) - A^*M\left(\delta_{Y\text{link}}\right) \tag{5}$$

with $\delta_{X\,link}$ = difficulty estimates of the link items of scale *X* and $\delta_{Y\,link}$ = difficulty estimates of the link items of scale *Y*. The discrimination parameter of the linked scale is then obtained by $\alpha_{\mathrm{Y}}^* = \alpha_{\mathrm{Y}} / A$. In Rasch models, mean/mean linking always results in $A = 1$ and therefore, the scale is shifted without changing the distribution of the item difficulty estimates[4]. Therefore, (5) reduces to

$$B_{\text{AID}} = M\left(\delta_{X\text{link}}\right) - M\left(\delta_{Y\text{link}}\right). \tag{6}$$

[4] In the mean/sigma method the slope of the linear scale transformation is computed as $A = SD(\xi_{Ylink}) / SD(\xi_{Xlink})$ using the standard deviations of the link item difficulty parameters from scales *Y* and *X,* typically resulting in $A \neq 1$. Consequently, the discrimination parameter $\alpha_{\mathrm{Y}}^* = \alpha_{\mathrm{Y}} / A$ is changed which would violate the basic assumption of the Rasch model that assumes constant discriminations of 1. Consequently, this linking method was excluded from further investigation in the present study.

To establish the link, B is added to each item difficulty parameter of the scale intended to link. In doing so, the item difficulty parameters of scale Y are shifted on the logit scale in such a way that the mean difficulty of the link items of both scales are equal. This procedure has no influence on the relation of item difficulty parameters within scale Y.

As the characteristic curve methods only differ with regard to the estimation of the scaling constants A and B the basic concept of the linking approach is identical to the moments approach (i.e., mean/mean linking). Previous studies in cross-sectional contexts found rather small differences in parameter accuracy for the two estimation approaches (e.g., Seonghoon Kim & Kolen, 2006). Consequently, the characteristic curve methods were excluded from further investigation in the present study.

Mean/mean method based on the anchor-group design (m/m$_{\text{AGD}}$)

To link scale Y to the reference scale X using an anchor-group design (AGD), the principle of the mean/mean method using anchor-items can be adapted by including the information of the link sample, which provides the unaltered relationship of the scales X and Y (as was described in the previous section). In contrast to the anchor-items design the link information is based on the entire test including all items. As in the anchor-items design the computation of $A_{\text{AGD}} = M(\alpha_Y) / M(\alpha_X)$ always results in 1. Adapting (6) for the anchor-group design results in

$$B_{\text{AGD}} = M\left(\delta_X\right) - M\left(\delta_Y\right) + \left(M\left(\delta_{Y,LS}\right) - M\left(\delta_{X,LS}\right)\right) \tag{7}$$

with $M(\delta_X)$ and $M(\delta_Y)$ being the mean item difficulties of the scales X and Y for the main sample and $M(\delta_{X,LS})$ and $M(\delta_{Y,LS})$ being the respective means of the link sample.

To establish the link, B is added to each item difficulty parameter of the scale intended to be linked. As in the anchor-items design, this procedure has no influence on the relation of item difficulty parameters within scale Y.

Because the mean/mean method (regardless of the underlying anchoring design) is based on a linear transformation, all difficulty estimates are shifted equally on the logit scale. Strictly speaking, the mean/mean method has no additional constraints, because the logit scale is invariant to linear transformations (Rasch, 1980). Thus, model fit is not influenced by the mean/mean method. As van der Linden and Barrett (2016) correctly point out, this shifting constant is always based on an arbitrarily chosen constraint that may (or may not) approximate the true parameters. In any case, no verification of this constraint is possible when using empirical data.

Fixed parameters calibration (FPC) based on the anchor-items design

The item difficulty parameters of the link items from the reference scale are fixed in the separate calibration of the scale intended to be linked. Thus, identical difficulty parameters of link items (anchored to the reference scale) result in both scales. This strict constraint may lead to a decrease in model fit in the linked scale due to possible differential item functioning (DIF) and due to sampling error whereas the model fit of the reference scale is not affected. A. von Davier and colleagues (2006) point out that this procedure is not advisable if two populations significantly differ in ability when taking two test forms. Transferring this thought to a longitudinal measurement would lead to the conclusion that the method of fixed parameters would not be advisable when large cognitive development is expected.

Concurrent calibration (CC) based on the anchor-items design

Both tests are scaled jointly in a concurrent analysis where each measurement loads on a single dimension. Items included in both tests are constrained to have identical item parameters in both samples. This quite strict one-step procedure strives for the golden mean between the two tests. Compared to calibrating both tests separately, some limitations in model fit have to be accepted on both tests due to possible DIF. Still, calibrating both tests concurrently seems a promising approach with regard to estimation efficiency (Jodoin, Keller, & Swaminathan, 2003) as well as the reduction of sampling error (Hanson & Beguin, 2002).

Previous findings on vertical linking

Though a lot of research has been done comparing IRT linking methods in the field of vertical linking the findings provide only little clarity on the suitability of linking methods (Arai & Mayekawa, 2011; Jodoin et al., 2003; Seock-Ho Kim & A. Cohen, 1992, 1992; Lei & Zhao, 2012; Tong & Kolen, 2007). This may be due to the vast variety of manipulated factors examined in studies that potentially influence the link outcome such as a) linking methods, b) anchoring design, c) type of data (empirical versus simulated), d) characteristics of common items (proportion within a test, range of item difficulties, dichotomous or polytomous items, DIF), e) test length, f) sample characteristics (size, motivation to participate such as high- versus low-stakes tests), g) test targeting, h) number of measurement points i) time gap/developmental progress between measurements, j) underlying IRT models, and k) violation of model assumptions (e.g., unidimensionality assumption). However, with the huge number of experimental conditions and comparisons drawn from different evaluation criteria, it seems rather difficult to disentangle the prior findings and to rank order the linking methods. Nevertheless, one effect that is consistently reported in literature is that increasing the sample size and the number of anchor items improves the link performance – regardless of the linking method. Some authors (Hanson & Beguin, 2002; Lei & Zhao,

2012) found that concurrent calibration resulted in smaller error than separate calibration. This effect was explained by Hanson and Beguin (2002) with the increased sample size in concurrent calibration compared to separate calibration. While Arai and Mayekawa (2011) suggested that the ratio of common items should exceed 10 % in order to not worsen the performance of concurrent calibration and fixed parameters calibration, Kolen and Brennan (2014) recommended a share of (at least) 20 %. Using empirical data Jodoin et al. (2003) found that separate calibration (mean/sigma method) resulted in less mean growth than concurrent calibration and fixed parameters calibration.

However, empirical studies comparing IRT linking methods on longitudinal data scaled with the Rasch model in LSAs are still missing. Therefore, the present empirical study aims at comparing and evaluating linking methods that fit the assumptions of Rasch-type models (i.e., mean/mean, fixed parameters calibration, concurrent calibration). Moreover, it aims at comparing linking methods based on two different anchoring designs (i.e., anchor-items design, anchor-group design) on the same data. As such, conclusions on the comparability of the link process and link outcome of linking methods based on the two anchoring designs could be drawn. In particular, the results of the present study will extend previous findings on vertical linking to the longitudinal context and complement research on two-parametric and three-parametric models by focusing on Rasch-type measurement models.

Method

Sample

We selected a panel sample (i.e., main sample) from the NEPS (Blossfeld, Roßbach, & von Maurice, 2011), which is a LSA based on a longitudinal design conducted in Germany. In the NEPS, participants from different age cohorts are followed up and are periodically administered low-stakes competence tests in various domains in order to measure competence development over the life span. In the present study, a total of n = 3,833 participants (48 % girls, 95 % born in Germany, 51 % attending high school), sampled representatively from schools across all 16 federal states, received a mathematics competence test in Grade 5 (age: M = 10.91, SD = .52) and Grade 7 (age: M = 12.91, SD = 0.52)[5]. Moreover, from the same population (but sampled from different schools) as the panel sample n = 581 participants (53 % girls, 93 % born in Germany, 44 % attending high school) attending Grade 7 (age: M = 13.08, SD = 0.59) were additionally sampled as independent link sample.

The study was approved by the Federal Ministries of Education in Germany and the data protection board of the National Educational Panel Study. Informed consent was given by parents, students, and educational institutions to take part in the study. Data from the

[5] Note that 1,360 of the initially 5,193 participants in Grade 5 did not take part in the measurement in Grade 7 and were thus, excluded from the analyses in the present study.

panel sample are available from http://www.neps-data.de for researchers who meet the criteria for access to confidential data. Data from the independent link sample are not accessible due to legal reasons.

Instruments

The conceptual framework underlying the mathematics tests administered in the NEPS is described in Neumann et al. (2013). Prior to test administration several pilot studies were conducted to guarantee that the final test form reflected the intended conceptual framework. As such, test development for the respective math tests in Grades 5 and 7 was theory driven, based on a Rasch-model-conforming unidimensional mathematical literacy concept. Additionally, the psychometric quality and fit to the Rasch model were empirically checked throughout test construction as well as for the final test forms, which were administered to the panel sample and the independent link sample.

The mathematics tests administered in Grades 5 and 7 included 24 items (marginal reliability = .80; Adams, 2005) and 23 items (marginal reliability = .76), respectively. In each test one item was polytomous, whereas the rest were dichotomous. Six dichotomous items were common to both tests and served as potential link items. As such, the number of common items corresponded to the recommended share of 20 % (Kolen & Brennan, 2014) in the literature. These common items were selected by educational experts on mathematics for broadly covering the underlying conceptual framework. Furthermore, these six items were expected to fit the anticipated change in ability between Grades 5 and 7 well. In order to prevent position effects (e.g., Hohensinn, Kubinger, Reif, Holocher-Ertl, Khorramdel, & Frebort, 2008; Trendtel & Robitzsch, 2018), all six common items retained their original position (see Tables 2 and 3) within each test from Grade 5 to Grade 7. Additionally, violation of local independence was checked to detect possible interaction effects with measurement point-unique items. To minimize the risk of memory effects the items reflected typical tasks administered in math classes at school. Thus, it was unlikely that students were able to remember correct solutions for these items across a time span of two years.

As the mathematics tests were not administered in a high-stakes setting, missing values were not handled as incorrect responses (Pohl & Carstensen, 2013). Consequently, if a participant gave no response, the answer was treated as missing (and not as incorrect). On average, participants had M = 1.8 (SD = 2.4) missing values in Grade 5 and M = 0.7 (SD = 1.4) missing values in Grade 7. The participants were tested at school in a group setting with a limited test time of 30 minutes per measurement occasion. For a detailed description of the scaling results see Duchhardt and Gerdes (2012) as well as Schnittjer and Gerken (2017).

Study Design

Both, anchor-items design and anchor-group design (see Figure 1) were combined in this study: While participants from the panel sample took the two mathematics tests with a time-lag of two years between Grades 5 and 7, participants of the link sample took both tests at one measurement point in Grade 7. To avoid memory and other effects in the link sample the six common items were included only in the Grade 7 test. In order to account for item position and test length effects the common items were replaced by new items of similar content and difficulty in Grade 5.

Statistical Analyses

All data were scaled using the PCM (Masters, 1982), which is an extension of the Rasch model to polytomous items applying item-specific rating scales. For linking methods based on separate calibration (i.e., mean/mean linking based on anchor-items and anchor-group design as well as FPC) each measurement occasion was scaled separately constraining the mean ability to zero while the linking was conducted afterwards. Applying the concurrent calibration we modelled our data using a two-dimensional PCM, setting the mean ability to zero at Grade 5 (dimension 1) and estimating the mean ability of Grade 7 (dimension 2). In line with Andersen (1985), we assumed that the difference in mean ability between Grades 5 and 7 represented the change of ability in the longitudinal panel sample. The software used was ACER ConQuest 4 (Adams, Wu, & Wilson, 2016) based on a marginal maximum likelihood estimation (Bock & Aitkin, 1981), in order to accommodate the partially missing responses. Note, that contemporary IRT software is unable to handle the present data when based on a conditional maximum likelihood estimation (Fischer & Molenaar, 2012).

As any empirical data can never fully meet the strict assumptions of a theoretical model such as the Rasch model, statistical tests will always discard a model if only the sample size is big enough. As a consequence, we assessed model fit using the weighted mean square (WMNSQ; Wright & Masters, 1982), its respective *t*-value and the corrected item-total correlation. The WMNSQ is a quantitative measure of fit discrepancy. It is based on the weighted deviation of an actual person's response from Rasch model expectation. Being distributed as mean squares, the expected value is 1 (Bond & Fox, 2015). In assessing model fit we adopted rules of thumb proposed in the literature (Pohl & Carstensen, 2012) and viewed a WMNSQ > 1.2 and a respective *t*-value > |8| as considerable item misfit. Note, that a well item fit according to the WMNSQs indicates that items of a test discriminate sufficiently at the various person ability levels, thus, meeting the respective specification in the Rasch model. For the corrected item-total correlation a value greater than .2 was deemed acceptable. Local independence on the item level was evaluated based on Yen's Q_3 (1993) statistic, indicating no substantial violation for values < |.20|. Moreover, visual comparisons of the observed and model-implied item characteristic curves were conducted to identify potentially misfitting items.

Examination of assumptions for longitudinal linking

In order to link adjacent measurement points, several assumptions have to be met. Tests and link items have to meet the assumptions of unidimensionality and must not show DIF (Pohl et al., 2015; A. von Davier et al., 2006). Common items that do not meet these assumptions should not be used as link items and may be modelled as group specific (unique) item parameters (e.g., Oliveri & M. von Davier, 2011).

Unidimensionality

To measure competence development within a domain over a period of time, the underlying theoretical construct must not change between time points. The unidimensionality assumption was examined twofold. First, a test can be considered essentially unidimensional when the standardized residuals of a one-dimensional model exhibit approximately zero-order correlations. While in case of an anchor-items design the residuals were derived from a one-dimensional model of the two separately scaled tests, in case of an anchor-group design the residuals were derived from a one-dimensional model of the two concurrently scaled tests that were administered in the link sample. Second, further evidence of a unidimensional scale is given if the ratio of the first two eigenvalues derived from the standardized residuals does not exceed 1.5 (Smith Jr, 2002).

Differential item functioning

The localization of the linked scale is determined by the resulting link information. Consequently, the person ability estimation (and as such the magnitude of the participant's ability change between two measurement points) is influenced by the link information. In order to not mix up change in person ability and drift in item difficulty, the link item parameters $\hat{\delta}_{Xi\text{link}}$ and $\hat{\delta}_{Yi\text{link}}$ must not change (i.e., retain their relative position on the logit scale) between two test administrations. DIF was examined using a Wald test that compares the estimated item difficulties resulting from a maximum likelihood estimation (Draba, 1977):

$$t_{XY} = \frac{\hat{\delta}_{Xi\text{link}} - \hat{\delta}_{Yi\text{link}}}{\sqrt{SE\left(\hat{\delta}_{Xi\text{link}}\right)2 + SE\left(\hat{\delta}_{Yi\text{link}}\right)2}}. \tag{8}$$

The resulting test statistic is *t* distributed. LSAs often have to deal with excessive test power due to large sample sizes. Consequently, the result of statistical tests becomes less meaningful. Instead of a classical null hypothesis (Cohen, 1994), Murphy and Myors (1999) suggested using a minimum effect hypothesis. Here, the critical value is not defined by an assumed difference of zero but by a proportion of variance accounted for. We followed the Educational Testing Service determining the critical value (Zieky,

1993) as 1.54 % variance accounted for to identify relevant deviations of item difficulty parameters between two groups.

When using an anchor-group design, DIF is examined among the two groups of main sample and link sample, applying the same procedure as for the anchor-items design.

Evaluation of linking methods

The three linking methods and two anchoring designs were evaluated with regard to three criteria:

Link error

The link error becomes relevant when comparisons are made between ability estimates of different measurement points. It is conceptualized as standard error (*SE*) of differences between the separately scaled and linked item difficulty parameters of the link items from the test intended to be linked:

$$SE = SD_{Y,Y^*} / \sqrt{k} \tag{9}$$

with $SD_{Y,Y}*$ = standard deviation of the link item parameter differences from the separately scaled scale *Y* and the linked scale *Y**, and *k* = the number of link items (adapted from PISA 2009 Technical Report; OECD, 2012 and PISA 2012 Technical Report; OECD, 2014). When an anchor-group design is used all items are handled as link items. The standard error of differences is then calculated as standard error of differences between the main sample and the link sample for each test and is pooled afterwards. An adapted approach is necessary for the computation of the link error emerging from a CC based on an anchor-items design. In contrast to m/m and FPC where the link item estimates are only changed in the latter measurement point, the link item estimates are changed in both measurement points when using a concurrent calibration. Therefore, the amount of change in link item estimates is split among the two measurement points by leaving the number of link items unchanged. In order to avoid counting the number of link items double, *k* was halved. To account for the standard deviation of differences in link items twice (once for each measurement point X and Y), the link error had to be pooled. For the concurrent calibration the link error was then computed as

$$SE = \sqrt{\left(\frac{SD_{X,X^*}}{\sqrt{\frac{k}{2}}}\right)^2 + \left(\frac{SD_{Y,Y^*}}{\sqrt{\frac{k}{2}}}\right)^2} \tag{10}$$

As such, when analysing mean differences of a group including at least two time points, the link error has to be considered by including it into the pooled *SE* (for further details see Organisation for Economic Co-operation and Development, 2014). Consequently, a larger link error contributes to a reduced test power. Furthermore, the link error can be understood as bias, concerning every participant equally. Therefore, the standard deviation of ability scores is not affected by the link error.

Mean growth rate estimation

Since the linking methods are based on different link information, they vary in their estimation of the mean growth rate which reflects the estimated mean change in participant's ability between the two test administrations. However, because our research was based on empirical data where the true change is unknown, a potential bias in the link results cannot be further investigated. For the separately scaled models (based on m/m$_{AID}$, m/m$_{AGD}$ and FPC) the mean growth rate estimate was obtained by a "post hoc" two-dimensional analysis where each test administration (i.e., Grades 5 and 7) loaded on a single dimension. The mean ability of the first test administration served as a reference category (i.e., it was fixed to zero). Due to the preceding link procedure each difficulty parameter was estimated in prior analyses and, thus, fixed to these values. The mean growth rate was computed as the mean change in the weighted maximum likelihood ability estimate (WLE; Warm, 1989) using the examinee response vector and the item parameters.

Model fit

After linking the two measurement points, we fitted a two-dimensional model for each of the linked data that constrained the item parameters to the previously derived and linked values (see above). This intermediate step was necessary to make the model fits and information criteria (Akaike information criterion (AIC); Akaike, 1974 and Bayesian information criterion (BIC); Schwarz, 1978) of the separately scaled models (based on m/m$_{AID}$, m/m$_{AGD}$ and FPC) and the concurrently scaled model (using concurrent calibration) comparable in order to evaluate how the different restrictions inherent to the different linking methods effected the model fit.

Results

A PCM was used to analyze the panel sample and link sample. Model identification was obtained by constraining the mean ability to zero. For the panel sample the mean item difficulty estimates of the separately scaled mathematics tests applied in Grades 5 and 7 were M = -0.63 (SD = 1.11, Min = -2.74, Max = 1.44) and M = -0.58 (SD = 1.01, Min = -3.13, Max = 1.19), respectively. The latent correlation of the Mathematical competences was r = .93 (p = .00) across the two measurement points. For the concurrently scaled link sample the mean item difficulty estimates of Grades 5 and 7 were M = -1.16 (SD = 0.90, Min = -2.58, Max = 1.03) and M = -0.35 (SD = 0.99, Min = -2.93, Max = 1.50). Overall,

Table 1:
Item Fit of Separately Scaled Mathematics Tests in Grades 5 and 7 for Panel Sample and Link Sample

		Percentage correct	**Item difficulty**	***SE***	**WMNSQ**	***t***	r_{it}	**Yen's Q_3**
panel sample grade 5	*M (SD)*	60.66 (19.11)	-0.63 (1.11)	0.05 (0.01)	1.00 (0.05)	-0.07 (2.90)	0.37 (0.07)	0.00 (0.03)
	Min/Max	23.34/91.30	-2.74/1.44	0.04/0.07	0.92/1.14	-5.80/9.30	0.25/0.47	-0.06/0.37
panel sample grade 7	*M (SD)*	59.39 (16.94)	-0.58 (1.01)	0.05 (0.01)	1.00 (0.07)	-0.06 (4.17)	0.40 (0.08)	0.00 (0.02)
	Min/Max	27.68/93.25	-3.13/1.19	0.04/0.07	0.88/1.16	-8.4/10.10	0.24/0.55	-0.06/0.12
link sample grade 5	*M (SD)*	70.54 (15.61)	-1.16 (0.90)	0.12 (0.01)	1.01 (0.07)	0.20 (1.42)	0.37 (0.08)	0.00 (0.05)
	Min/Max	30.89/89.86	-2.58/1.03	0.11/0.15	0.89/1.14	-2.10/3.30	0.24/0.53	-0.14/0.33
link sample grade 7	*M (SD)*	55.62 (17.17)	-0.35 (0.99)	0.11 (0.02)	0.99 (0.07)	-0.14 (1.60)	0.40 (0.07)	0.00 (0.05)
	Min/Max	22.80/92.39	-2.93/1.50	0.11/0.17	0.88/1.15	-2.60/3.60	0.26/0.51	-0.14/0.14

Note. n = 3,833 (panel sample) and n = 581 (link sample); item difficulty = location parameter; *SE* = Standard error of difficulty / location parameter; WMNSQ = Weighted mean square; t = t-value for WMNSQ: comparing *Min/Max* among panel sample and link sample leads to the conclusion that mere sample size rather than actual item misfit was responsible for the difference in t-values between the two samples (Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008); r_{it} = Corrected item-total correlation; Yen's Q_3 = Yen's (1993) corrected Q_3: statistic tests for violation of local independence assumption if Q_3 values > |.20|;

Table 2:
Difficulty Estimates of the Separately, Concurrently and Linked Scaled Grade 5-Test

No.	Item	Panel sample	Link sample	Linked Estimates			
				m/m		FPC	CC
				m/m$_{\text{AGD}}$	m/m$_{\text{AID}}$		
1	Item 1	-0.51	-1.06	≙ PS	≙ PS	≙ PS	-0.51
2	Item 2	-1.15	-1.32	≙ PS	≙ PS	≙ PS	-1.15
3	Item 3	-0.92	-1.42	≙ PS	≙ PS	≙ PS	-0.92
4	Item 4	0.86	0.24	≙ PS	≙ PS	≙ PS	0.86
5	Item 5	-0.17	-1.74	≙ PS	≙ PS	≙ PS	-0.17
6	Item 6	0.38	-0.03	≙ PS	≙ PS	≙ PS	0.38
7[a]	Item 7	0.49	-	≙ PS	≙ PS	≙ PS	0.49*
8	Item 8	-1.98	-1.58	≙ PS	≙ PS	≙ PS	-1.98
9[a]	Item 9	-2.72	-	≙ PS	≙ PS	≙ PS	-2.58*
10[a]	Item 10	-0.69	-	≙ PS	≙ PS	≙ PS	-0.85*
11	Item 11	-0.86	-1.42	≙ PS	≙ PS	≙ PS	-0.86
12	Item 12	1.44	1.03	≙ PS	≙ PS	≙ PS	1.44
13	Item 13	-0.22	-0.78	≙ PS	≙ PS	≙ PS	-0.22
14	Item 14	-1.33	-2.10	≙ PS	≙ PS	≙ PS	-1.33
15	Item 15	-1.55	-1.40	≙ PS	≙ PS	≙ PS	-1.55
16	Item 16	-2.19	-2.29	≙ PS	≙ PS	≙ PS	-2.19
17	Item 17	-0.53	-1.28	≙ PS	≙ PS	≙ PS	-0.53
18	Item 18	-0.23	-1.11	≙ PS	≙ PS	≙ PS	-0.23
19[a]	Item 19	0.11	-	≙ PS	≙ PS	≙ PS	0.18*
20	Item 20	-1.27	-1.65	≙ PS	≙ PS	≙ PS	-1.27
21[a]	Item 21	0.92	-	≙ PS	≙ PS	≙ PS	0.98*
22	Item 22	-2.74	-2.58	≙ PS	≙ PS	≙ PS	-2.74
23[a]	Item 23	-0.41	-	≙ PS	≙ PS	≙ PS	-0.43*
24	Item 24	0.22	-0.46	≙ PS	≙ PS	≙ PS	0.21
$M_{\text{all items}}$		-0.63 (1.11)	-	≙ PS	≙ PS	≙ PS	-0.62 (1.10)
$\text{M}_{\text{link items excluded}}$		-0.71 (1.07)	-1.16 (0.90)	≙ PS	≙ PS	≙ PS	-0.71 (1.07)
$\text{M}_{\text{link items}}$		-0.38 (1.28)	-	≙ PS	≙ PS	≙ PS	-0.37 (1.26)

Note. n = 3833 (panel sample) and n = 581 (link sample). The difficulty estimates for the linking methods m/m (based on anchor-items as well as on an anchor-group) and FPC match the independently scaled Grade 5 test, because they do not change the reference scale. The CC changes the reference scale and therefore, the resulting difficulty estimates differ from those of Grade 5. The six common items in the link sample were replaced by new items which were excluded from analyses. The mean difficulty estimates include the respective standard deviation in parentheses *M* (*SD*). Grades 5 and 7 were modelled unidimensional in the link sample. No. = item order in test administration; PS = panel sample; m/m$_{\text{AID}}$ = mean/mean method based on anchor-items design; m/m$_{\text{AGD}}$ = mean/mean method based on anchor-group design; FPC = fixed parameters calibration; CC = concurrent calibration.

[a]link item; *parameter was constrained

Table 3:
Difficulty Estimates of the Separately, Concurrently and Linked Scaled Grade 7-Test

				Linked Estimates			
				m/m			
No.	**Item**	**Panel sample**	**Link sample**	**m/m$_{\text{AGD}}$**	**m/m$_{\text{AID}}$**	**FPC**	**CC**
1	Item 25	-0.36	-0.07	0.32	0.36	0.39	0.39
2	Item 26	0.50	0.62	1.19	1.22	1.25	1.25
3	Item 27	0.21	0.36	0.89	0.93	0.96	0.96
4	Item 28	0.29	0.37	0.97	1.01	1.04	1.04
5[a]	Item 7	-0.27	-0.04	0.42	0.46	0.49*	0.49*
6	Item 29	-1.36	-0.98	-0.67	-0.63	-0.61	-0.60
7[a]	Item 9	-3.13	-2.93	-2.44	-2.40	-2.72*	-2.58*
8[a]	Item 10	-1.83	-1.41	-1.14	-1.10	-0.69*	-0.85*
9	Item 30	-0.52	-0.46	0.16	0.20	0.23	0.23
10	Item 31	0.24	0.39	0.93	0.97	0.99	1.00
11	Item 32	-0.65	-0.50	0.04	0.08	0.10	0.11
12	Item 33	-1.83	-1.42	-1.14	-1.10	-1.08	-1.07
13	Item 34	-0.12	-0.03	0.57	0.61	0.63	0.63
14	Item 35	-1.82	-1.57	-1.13	-1.09	-1.07	-1.06
15	Item 36	-1.35	-1.29	-0.66	-0.62	-0.60	-0.59
16	Item 37	-0.15	0.26	0.54	0.58	0.60	0.60
17	Item 38	-0.39	-0.38	0.29	0.33	0.36	0.36
18[a]	Item 19	-0.50	-0.08	0.19	0.22	0.11*	0.18*
19	Item 39	0.66	1.05	1.35	1.39	1.41	1.41
20[a]	Item 21	0.28	0.38	0.97	1.01	0.92*	0.98*
21	Item 40	1.19	1.50	1.88	1.91	1.94	1.94
22[a]	Item 23	-1.22	-0.71	-0.53	-0.49	-0.41*	-0.43*
23	Item 41	-1.26	-1.09	-0.57	-0.53	-0.51	-0.50
$M_{\text{all items}}$		-0.58 (1.01)	-0.35 (0.99)	0.11 (1.01)	0.14 (1.01)	0.16 (1.03)	0.17 (1.02)
$M_{\text{link items}}$		-1.11 (1.24)	-0.80 (1.22)	-0.42 (1.24)	-0.38 (1.24)	-0.38 (1.28)	-0.37 (1.26)

Note. n = 3833 (panel sample) and *n* = 581 (link sample). Linking with m/m$_{\text{AGD}}$: constant $B_{\text{anchor-group}}$ = .681 (see (7)) is added to each parameter of PS. Linking with m/m$_{\text{AID}}$: constant $B_{\text{anchor-items}}$ = .726 (see (6)) is added to each parameter of PS. Using FPC or CC: constraints are set by anchoring or equalizing parameters (indicated by *). The mean difficulty estimates include the respective standard deviation in parentheses *M* (*SD*). Grades 5 and 7 were modelled unidimensional in the link sample. No. = item order in test administration; m/m$_{\text{AGD}}$ = mean/mean method based on anchor-group design; m/m$_{\text{AID}}$ = mean/mean method based on anchor-items design; FPC = fixed parameters calibration; CC = concurrent calibration.
[a]link item; *parameter was constrained

item fit was satisfactory (see Table 1; Pohl & Carstensen, 2012) as indicated by corrected item-total correlations (r_{it}) exceeding .23 and WMNSQ falling between 0.88 and 1.16 (Pohl & Carstensen, 2012). With Smith Jr's (2002) ratio test not exceeding 1.5 and the *Q*3 statistics falling between *Min* = -.14 and *Max* = .37, the unidimensionality assumption for both tests in the panel sample and link sample was supported and for all but two items[6] no violation of local independence was detected.

The item difficulty estimates of the panel sample and link sample of the separately, concurrently and linked scaled tests of Grades 5 and 7 are summarized in Tables 2 and 3.

Differential item functioning

DIF was examined between the common items (in the anchor-items design) and between the panel sample and link sample (in the anchor-group design). The difference in item difficulties between the six items administered at both measurement occasions and the results of the respective minimum effect hypotheses tests are summarized in Table 4.

Table 4:
Examination of Differential Item Functioning for the Six Common Items in the Anchor-Items Design

Link item	δ_{G5}	δ_{G7}	$\Delta\delta$	$SE_{\Delta\delta}$	t	F
Item 7	0.87	0.84	-0.03	0.06	-0.48	0.23
Item 9	-2.34	-2.02	0.32	0.09	3.33	11.09
Item 10	-0.30	-0.72	-0.42	0.06	-6.44	41.52
Item 19	0.49	0.61	0.12	0.06	2.03	4.12
Item 21	1.30	1.39	0.09	0.06	1.56	2.42
Item 23	-0.02	-0.11	-0.09	0.06	-1.39	1.93

Note. Item difficulty estimates (with their means set to zero) based on participants that took part in Grades 5 and 7 (N = 3,833). The t and F statistics resulted from a Wald test (see (8)). None of the common items exceeded the critical value of $F_{0154}(1, 3{,}831) = 88.3$ for a p = .05. Therefore, all items met the assumption of showing no substantial DIF and qualified as link items. $\Delta\delta$ = difference in item difficulty parameters between Grades 7 and 5 (positive values indicate easier items in Grade 5); $SE_{\Delta\delta}$ = pooled standard error; t = t statistic; F = F statistic.

[6] The residuals of the Grade 5 test-unique items 2 and 3 correlated at .37 in the panel sample and at .33 in the link sample. An inspection of the item content revealed that both items were somewhat similarly phrased (i.e., some words overlapped) and were presented one after another. However, visual checks of the respective item characteristic curves and an evaluation of the item level fit statistics for these items in the panel sample (WMNSQ: 0.98, 1.03; t-value: -0.9, 1.9; corrected item-total correlation: .43, .38) and the link sample (WMNSQ: 1.1, 1.02; t-value: 1.9, 0.4; corrected item-total correlation: .29, .38) did not identify a severe misfit. Therefore, both items were included in the final scaling procedure as intended by the test developers.

None of the resulting *F* statistics was significant (all *p*s > .05) and, as such, indicated no pronounced DIF qualifying the six common items as link items. Also, no relevant DIF was found between the panel sample and link sample (see Table 5). None of the *F* statistics indicated significantly ($p < .05$) different item parameters between the two samples.

Table 5:
Examination of Differential Item Functioning between Panel Sample and Link Sample in Grades 5 and 7

	Grade 5					**Grade 7**			
Test	**item**	$\Delta\delta$	$SE_{\Delta\delta}$	F	**Test**	**item**	$\Delta\delta$	$SE_{\Delta\delta}$	F
G5	Item 1	0.09	0.12	0.64	G7	Item 25	-0.06	0.11	0.28
G5	Item 2	-0.29	0.13	5.39	G7	Item 26	0.12	0.11	1.02
G5	Item 3	0.04	0.13	0.11	G7	Item 27	0.08	0.11	0.51
G5	Item 4	0.16	0.11	2.06	G7	Item 28	0.16	0.11	1.86
G5	Item 5	1.12	0.13	72.42	G7	Item 7	0.00	0.11	0.00
G5	Item 6	-0.04	0.12	0.12	G7	Item 29	-0.14	0.12	1.45
G5	Item 7	-	-	-	G7	Item 9	0.03	0.19	0.03
G5	Item 8	-0.86	0.13	41.36	G7	Item 10	-0.18	0.13	2.06
G5	Item 9	-	-	-	G7	Item 30	0.17	0.11	2.16
G5	Item 10	-	-	-	G7	Item 31	0.08	0.11	0.54
G5	Item 11	0.11	0.13	0.65	G7	Item 32	0.08	0.11	0.53
G5	Item 12	-0.04	0.12	0.13	G7	Item 33	-0.18	0.13	1.96
G5	Item 13	0.11	0.12	0.90	G7	Item 34	0.14	0.11	1.52
G5	Item 14	0.31	0.14	4.77	G7	Item 35	0.02	0.13	0.02
G5	Item 15	-0.60	0.13	21.94	G7	Item 36	0.17	0.12	1.96
G5	Item 16	-0.36	0.15	5.35	G7	Item 37	-0.18	0.11	2.49
G5	Item 17	0.30	0.12	5.82	G7	Item 38	0.22	0.11	3.71
G5	Item 18	0.42	0.12	11.74	G7	Item 19	-0.19	0.11	2.80
G5	Item 19	-	-	-	G7	Item 39	-0.16	0.12	1.78
G5	Item 20	-0.08	0.14	0.31	G7	Item 21	0.14	0.12	1.48
G5	Item 21	-	-	-	G7	Item 40	-0.07	0.13	0.33
G5	Item 22	-0.62	0.17	13.67	G7	Item 23	-0.27	0.12	5.28
G5	Item 23	-	-	-	G7	Item 41	0.07	0.17	0.17
G5	Item 24	0.22	0.11	3.59					

Note. Item difficulty estimates based on the panel sample and the link sample. The *F* statistics resulted from the squared *t*-value of a Wald test (see (8)). None of the items exceeded the critical value of F_{0154} (1, 4,412) = 99.2 for $p = .05$. Therefore, no DIF was found. G5 = mathematics test in Grade 5; G7 = mathematics test in Grade 7; $\Delta\delta$ = difference in item difficulty parameters between panel sample and link sample (positive values indicate easier items in the link sample); $SE_{\Delta\delta}$ = pooled standard error; $t = t$ statistic; $F = F$ statistic.

Evaluation of linking methods

The results of the evaluation criteria for the three linking methods mean/mean (either based on an anchor-group or an anchor-items design), FPC and CC are summarized in Table 6.

Table 6:
Results of Linking Method Evaluation

Linking Method	Link Error	$\Delta\beta$	$Var(\Delta\beta)$	AIC	BIC	Parameters
m/m$_{AGD}$	0.11	0.68 (0.78)	0.76	190,964	191,295	53
m/m$_{AID}$	0.10	0.72 (0.83)	0.76	190,964	191,295	53
FPC	0.10	0.74 (0.85)	0.76	191,079	191,373	47
CC	0.10	0.75 (0.86)	0.76	191,023	191,317	47

N = 3,833. The link error was calculated using (9) and (10). m/m = mean/mean method (based either on the AID or AGD). Δβ = mean growth estimation in person ability parameters between Grades 5 and 7 in logits: positive values indicate a gain of ability between the measurement points (in parentheses: Cohen's d for repeated measures ANOVA; Morris & DeShon, 2002); *Var*(Δβ) = Variance of Change between Grades 5 and 7; AIC = Akaike's information criterion; BIC = Bayesian information criterion; Parameters = number of estimated parameters during scaling; m/m$_{AGD}$ = mean/mean method based on anchor-group design; m/m$_{AID}$ = mean/mean method based on anchor-items design; FPC = fixed parameters calibration; CC = concurrent calibration.

Link error

While the link errors for the methods m/m$_{AID}$, FPC and CC were equivalent (i.e., 0.10), the link error of m/m$_{AGD}$ was slightly larger, i.e., 0.11. Note, that m/m$_{AID}$ and FPC always result in perfectly matching link errors due to the fact that the calculations were based on the same difficulty estimates.

Mean growth rate

For Grades 5 and 7 the *SD* of ability estimates was identical among all linking procedures (SD_{G5} = 1.16, SD_{G7} = 1.24). Taking into account the high latent correlation in ability ($r_{G5,G7}$ = .93) between the Grades 5 and 7 it was not surprising that no substantial differences were found. As such, no evidence for the phenomenon of scale shrinkage (i.e., a reduction of sample variance induced trough IRT linking methods; see Briggs & Weeks, 2009) was found among the linking methods. The differences in the mean growth rates (m/m$_{AGD}$: $\Delta\beta$ = 0.68, m/m$_{AID}$: $\Delta\beta$ = 0.72, FPC: $\Delta\beta$ = 0.74, CC: $\Delta\beta$ = 0.75) were analyzed using a one-way repeated-measures ANOVA. Mauchly's test indicated that the assumption of sphericity had been violated ($\chi^2(5)$ = 46,30, p = .00); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .46). The results showed that the amount of growth was significantly effected by the applied linking procedure *F*(1.37, 5,266) = 180,393, *p* < .001. As such, effect sizes for the differences in mean growth between Grades 5 and 7 (d_{RM}) among the linking procedures were calculated (see Morris & DeShon, 2002). Effect sizes resulted in m/m$_{AGD}$: d_{RM} = 0.78, m/m$_{AID}$: d_{RM} = 0.83, FPC: d_{RM} = 0.85 and CC: d_{RM} = 0.86. With the differences in effect sizes having a range of 0.08 (between m/m$_{AGD}$ and CC), the difference in mean growth among the linking procedures was considered rather small. Still, while m/m$_{AID}$, FPC and CC form a homogenous group, m/m$_{AGD}$ seemed a bit trailed off. For these differences directly trace back to the differences in difficulty estimates resulting from the different linking procedures, additional analyses were calculated. No significant difference between the mean difficulty estimates of the separately scaled Grade 5 test (*M* = 0.63, *SD* = 1.11; which equally represented the estimates of m/m$_{AGD}$, m/m$_{AID}$ as well as FPC) and the concurrent calibration (*M* = 0.62, *SD* = 1.10, *t*(23) = -0.34, *p* = .74, d_{RM} = 0.01) was found (see Figure 2). Furthermore, difficulty estimates of Grade 7 (see Fig 3) were analyzed using a one-way repeated-measures ANOVA. Again, Mauchly's test indicated that the assumption of sphericity had been violated (χ^2(5) = 220.21, *p* = .00); therefore, degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity (ε = .34). The results showed that the difficulty estimates were significantly effected by the linking procedure *F*(1.00, 22.09) = 5.69, *p* < .03, η_p^2 = .21. Post hoc tests using the Bonferroni correction revealed that m/m$_{AGD}$ (*M* = 0.11, *SD* = 0.21) was signifi-

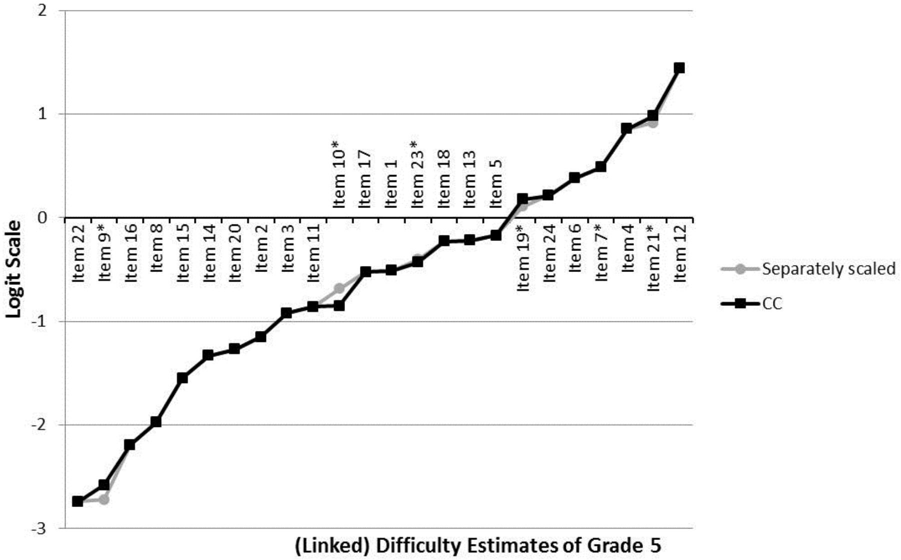

Figure 2:
Difficulty Estimates of Grade 5. Separately scaled = the separately scaled Grade 5 test equals the estimates of m/mAGD, m/mAID and FPC; CC = concurrent calibration; item difficulties are in ascending order; link items are denoted by *

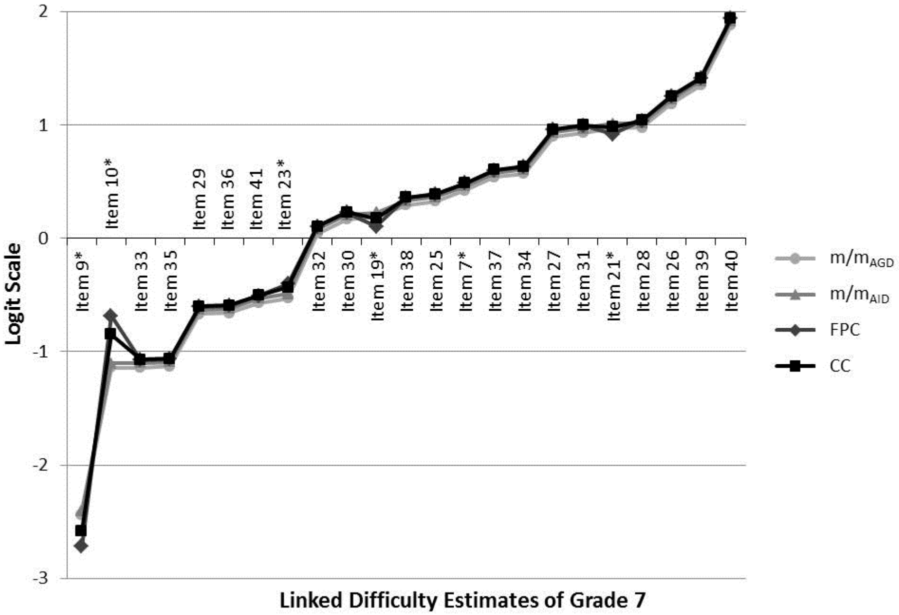

Figure 3:
Linked Difficulty Estimates of Grade 7. m/m(AGD) = mean/mean linking based on anchor-group design; m/m(AID) = mean/mean linking based on anchor-items design; FPC = fixed parameters calibration; CC = concurrent calibration; item difficulties are in ascending order; link items are denoted by *

cantly different from m/m$_{AID}$ (M = 0.14, SD = 0.21, p = .00, Cohen's d = -0.38) and CC (M = 0.17, SD = 0.21, p = .00, Cohen's d = -0.63) but not significantly different from FPC (M = 0.16, SD = 0.21, p = .18, Cohen's d = -0.55).

Model fit

The identical model fit of m/m$_{AGD}$ and m/m$_{AID}$ (Deviance = 190,857, number of parameters = 53, AIC = 190,964, BIC = 191,295) originated from their shared principle of linking methods. As mentioned above, model fit is not influenced when scaling using the mean/mean method (regardless of the anchoring design). Therefore, there is no point in comparing model fit among the mean/mean method and the other two linking methods in terms of building an evaluative rank order between them. Still, as the resulting model fit of the mean/mean method represents the cumulated model fit of the two separately scaled measurement points of Grades 5 and 7, it may serve as a general reference value for FPC (Deviance = 190,985, number of parameters = 47, AIC = 191,079, BIC = 191,373) and CC (Deviance = 190,928, number of parameters = 47, AIC = 191,023, BIC = 191,317).

However, it is no surprise that the information criteria both favored the CC over the FPC, given the method's constraints. Clearly, equalizing parameters leaves a model more space to fit the data than anchoring parameters to a fixed value.

Discussion

In this study we compared and evaluated the methods mean/mean linking (based on an anchor-items design (m/m_{AID}) as well as on an anchor-group design (m/m_{AGD}), fixed parameters calibration (FPC) and concurrent calibration (CC) on their performance to align two tests on a common scale. We applied the criteria of link error, mean growth rate estimation and model fit to evaluate the linking performance. The empirical data used in this study are based on participants that were administered two tests on mathematical literacy in Grades 5 and 7. In practice, the link information in LSA is typically either based on an anchor-group design or an anchor-items design. In contrast, the design of this study allowed a simultaneous comparison of both, anchoring designs as well as linking methods.

Overall, little differences among the linking methods were found. With the linking based either on a rather small absolute number of six link items (representing a proportion of 25 %) or a small link sample, measurement error and sampling error were less likely to cancel out. Of all evaluation criteria, differences among the linking methods and anchoring designs were most explicitly reflected by the mean growth. Though we found rather small differences in mean growth among the linking methods (in ascending order: m/m_{AID} < FPC < CC) this trend supported the findings of Jodoin et al. (2003) who reported less mean growth for methods based on a linear transformation (i.e., mean/sigma method) compared to FPC and CC using empirical data. Concluding from the findings of our more in-depths analysis it seems plausible to expect increasing differences among the linking methods the more subsequent measurement points are added. A bigger difference in mean growth was found between the anchoring designs. The significant differences in difficulty estimates of medium effect size between m/m_{AGD} and m/m_{AID} as well as m/m_{AGD} and CC probably resulted from the different sources of link information (i.e., either link sample or anchor items). Though the anchor-group design should result in a more valid link due to the bigger number of link items, it was also based on a smaller sample size and thus, more prone to sampling error. However, as little research exists in evaluating linking methods based on the anchor-group design more research is necessary to further investigate effects of age, sample size, characteristics of domain-specific development and the amount of time between measurement points.

Consistent with Hanson and Béguin (2002) and Lei and Zhao (2012) we found no difference in link error among the linking methods nor the two anchoring designs. Rather small differences in model fit criteria reflected the model's constraints as expected. Furthermore, the linking methods showed no substantial influence on the sample variance.

As no DIF in link items was found among Grades 5 and 7 in the panel sample, as well as among the panel sample and the link sample in Grade 7 we concluded that there was no substantial memory effect in repeatedly administered link items. As such, no effect was found on the response behavior of the students to answer a math item they had already worked on two years ago.

In contrast to the suggestion of A. von Davier and colleagues (2006) no evidence was found that FPC was not advisable when two populations significantly differed in (mean) ability when taking two test forms. However, as intraindividual change over time was very homogenous in our sample (with barely any change in rank order), there were only little differences in ability distributions among Grades 5 and 7.

Overall, the present case study found few differences between the examined linking methods. This suggests that the estimation of competence development is not profoundly effected by the methodological choices adopted for scaling the results. However, for the interpretation of these results one needs to keep in mind that they are based on a rather specific setting: longitudinal comparisons between Grades 5 and 7 for mathematical competencies among German students. It is unclear to what degree these findings extend to, for example, other populations, content domains, or age groups. Therefore, the generalizability of the presented results needs to be explored in further research that evaluates the robustness of linking methods applied to Rasch-model-scaled longitudinal data in different settings.

Limitations of the study

Since our analyses are based only on two measurement points, effects may accumulate over measurement points when adding subsequent measurements. This urges the necessity of sticking to an already applied linking method when linking data of more than two measurement occasions to avoid change in competence development being influenced by a change of linking methods. Furthermore, dropout rate is an issue in longitudinal designs (see Zinn & Gnambs, 2018). For various reasons participants drop out of the sample and cannot be reached anymore. Therefore, refreshing the sample periodically is necessary to perpetuate a proper sample size. Especially in the context of institutional education (e.g., school, university) the remaining sample after dropout typically represents a positive selection of participants. As a consequence, DIF in item parameters has to be examined between both groups of participants defined by taking part in one or two measurement points. In this study, 1,360 from 5,193 participants took only the test in Grade 5 and did not take part in Grade 7. The mean ability of these 1,360 participants is 0.47 logits lower than the mean of the participants that were to stay in the sample. Hence, future research is challenged with the question if and how linking methods differ in their ability estimation when applied in extended samples (i.e. samples including also ‘cross-sectional’ participants).

Though various procedures are discussed in the literature to provide statistically guided help to identify reasonable link items (e.g., Bechger & Maris, 2015) the authors were not aware of a solution to overcome the more general issue of identification in Rasch-type models. We therefore opted for a construct-driven decision procedure for selecting suitable link items from the common items.

Conclusion

As memory effects in items become more likely with repeated administration the anchor-group design seems a conceivable alternative to the anchor-items design in longitudinal measurements. Despite the overall small effects found, we join Hanson and Béguin (2002) in their advice to compare various linking methods as this enables the researcher to examine differences in linking methods (albeit the true parameter is never known when analyzing empirical data) but also serves as a reminder that the link result is based on an arbitrarily chosen link information (e.g., van der Linden & Barrett, 2016).

Author Note

This research was supported in part by grants of the Deutsche Forschungsgemeinschaft awarded to Claus H. Carstensen.

This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 3–Grade 5 and Grade 7, NEPS:SC3:5.0.0. From 2008 to 2013, NEPS data were collected as part of the Framework Programme for the Promotion of Empirical Educational Research funded by the German Federal Ministry of Education and Research (BMBF; https://www.bmbf.de). As of 2014, the NEPS survey is carried out by the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg in cooperation with a nationwide network. The authors declare that they have no competing interests.

References

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in Educational Evaluation*, *31*, 162–172. https://doi.org/10.1016/j.stueduc.2005.05.008

Adams, R. J., Wu, M. L., & Wilson, M. R. (2016). *ConQuest*. Camberwell: ACER. Retrieved from https://www.acer.edu.au/conquest/acer-conquest1

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. https://doi.org/10.1109/TAC.1974.1100705

Andersen, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16. https://doi.org/10.1007/BF02294143

Arai, S., & Mayekawa, S.-i. (2011). A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, *38*, 1–16. https://doi.org/10.2333/bhmk.38.1

Bechger, T. M., & Maris, G. (2015). A statistical test for differential item pair functioning. *Psychometrika*, *80*, 317–340. https://doi.org/10.1007/s11336-014-9408-y

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability.In F. M. Lord, & M. R. Novick (Eds.), *Statistical Theories of Mental Test Scores* (pp. 397-479). Reading, England: Addison-Wesley.

Blossfeld, H.-P., Roßbach, H.-G., & von Maurice, J. (Eds.). (2011). *Zeitschrift für Erziehungswissenschaft Sonderheft: Vol. 14. Education as a lifelong process: The German National Educational Panel Study (NEPS)*. Wiesbaden, Germany. VS Verlag für Sozialwissenschaften.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443–459. https://doi.org/10.1007/BF02293801

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. Mahwah, NJ:Routledge.

Briggs, D. C., & Weeks, J. P. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practice*, *28*, 3–14. https://doi.org/10.1111/j.1745-3992.2009.00158.x

Cohen. (1994). The earth is round (p <.05). *American Psychologist*, *49*, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Draba, R. E. (1977). The identification and interpretation of item bias. *Research Memorandum*, *25*.

Duchhardt, C., & Gerdes, A. (2012). *NEPS Technical Report for mathematics – scaling results of Starting Cohort 3 in fifth grade* (NEPS Working Paper). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Fischer, G. H., & Molenaar, I. W. (2012). *Rasch models: Foundations, recent developments, and applications*: Springer.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144-149;. https://doi.org/10.4992/psycholres1954.22.144

Hanson, B. A., & Beguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3–24. https://doi.org/10.1177/01466 21602026001001

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, S., Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly*, *50*, 391.

Jodoin, M. G., Keller, L. A., & Swaminathan, H. (2003). A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education*, *71*, 229–250. https://doi.org/10.1080/00220970309602064

Kim, Seock-Ho, & Cohen, A. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*, 51–66.

Kim, Seonghoon, & Kolen, M. J. (2006). Robustness to format effects of IRT linking methods for mixed-format tests. *Applied Measurement in Education*, *19*, 357–381.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (Third Edition). *Statistics for Social and Behavioral Sciences*. New York, NY: Springer.

Kubinger, K. D., & Draxler, C. (2006). A comparison of the Rasch model and constrained item response theory models for pertinent psychological test data. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 295–312). New York, NY: Springer.

Lei, P.-W., & Zhao, Y. (2012). Effects of vertical scaling methods on linear growth estimation. *Applied Psychological Measurement*, *36*, 21–39. https://doi.org/10.1177/0146621611425171

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch model. *Journal of Educational Measurement*, *17*, 179–193. https://doi.org/10.1111/j.17453984.1980.tb00825.x

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement*, *14*, 139–160. https://doi.org/10.1111/j.1745-3984.1977.tb00033.x

Masters. (1982). A rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. https://doi.org/10.1007/BF02296272

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*, 105–125. https://doi.org/10.1037/1082-989X.7.1.105

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. https://doi.org/10.1037/0021-9010.84.2.234

Neumann, I., Duchhardt, Christoph, Grüßing, M., Heinze, A., Knopp, E., & Ehmke, T. (2013). Modeling and assessing mathematical competence over the lifespan. *Journal for Educational Research Online / Journal Für Bildungsforschung Online*, *5*, 80–109.

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*, 315.

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 Technical Report*. Paris, France: Author. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10595644

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 Technical Report*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2017). *PISA 2015 Technical Report*. Paris, France: Author.

Pohl, S., & Carstensen, C. H. (2012). *Scaling the data of the competence tests (NEPS technical report 14)*. Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Pohl, S., & Carstensen, C. H. (2013). Scaling of competence tests in the National Educational Panel Study-Many questions, some answers, and further challenges/Skalierung der Kompetenztests im Nationalen Bildungspanel-Viele Fragen, einige Antworten und weitere Herausforderungen. *Journal for Educational Research Online*, *5*, 189.

Pohl, S., Haberkorn, K., & 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 Sciences Research (pp. 281-308). Springer.

Prenzel, M., Carstensen, C. H., Schöps, K., & Maurischat, C. (2006). Die Anlage des Längsschnitts bei PISA 2003. In *PISA Konsortium Deutschland (Ed.). PISA 2003: Untersuchungen zur Kompetenzentwicklung im Verlauf eines Schuljahres* (pp. 29–62). Münster, Germany: Waxmann.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: Mesa Press.

Schnittjer, I., & Gerken, A.-L. (2017). *NEPS Technical Report for mathematics - scaling results of Starting Cohort 3 in seventh grade* (NEPS Survey Papers No. 16). Bamberg, Germany: University of Bamberg, National Educational Panel Study.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464. https://doi.org/10.1214/aos/1176344136

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*, 33. https://doi.org/10.1186/1471-2288-8-33

Smith Jr, E. V. (2002). Understanding Rasch measurement: Detecting and evaluating the impact of multidimenstionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3,* 205-231.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201–210. https://doi.org/10.1177/014662168300700208

Tong, Y., & Kolen, M. J. (2007). Comparisons of methodologies and results in vertical scaling for educational achievement tests. *Applied Measurement in Education*, *20*, 227–253.

Trendtel, M., & Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psychological Test and Assessment Modeling*, *60*, 241–263.

Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*, 333–344. https://doi.org/10.1177/014662168601000402

Van der Linden, W., & Barrett, M.-D. (2016). Linking item response model parameters. *Psychometrika*, *81*, 650–673. https://doi.org/10.1007/s11336-015-9469-6

von Davier, A., Carstensen, C. H., & von Davier, M. (2006). Linking competencies in educational settings and measuring growth. *ETS Research Report Series*, *2006*, 36. https://doi.org/10.1002/j.2333-8504.2006.tb02018.x

von Davier, M., & Carstensen, C. H. (Eds.). (2006). *Multivariate and mixture distribution Rasch models: Extensions and applications*. New York: Springer.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450. https://doi.org/10.1007/BF02294627

Wright, B. D., & Masters. (1982). *Rating scale analysis*. Chicago, Il: Mesa Press.

Yen, W. M. (1993). Scaling performance assessments: strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187–213. https://doi.org/10.1111/j.1745-3984.1993.tb00423.x

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P. W. H. H. Wainer (Ed.), *Differential item functioning* (pp. 337–347). Hillsdale, NJ: Lawrence Erlbaum.

Zinn, S., & Gnambs, T. (2018). Modeling competence development in the presence of selection bias. *Behavior Research Methods*, *50*, 2426–2441. https://doi.org/10.3758/s13428-018-1021-z