Chapter 12
Linking PISA Competencies over Three Cycles – Results from Germany

Claus H. Carstensen

Abstract Since the publication of the PISA 2006 study results the question of reporting trends over the PISA cycles has received a lot of interest. This chapter discusses the possibilities and limitations of trend analyses based on data from this international comparative study and using complex test designs. The chapter succeeds trend analyses which were carried out with the German data from the first three PISA studies in 2000, 2003 and 2006 (Carstensen CH, Prenzel M, Baumert J, Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt? [Trend analyses in PISA: how did competencies in Germany develop between PISA 2000 and PISA 2006?] Zeitschrift für Erziehungswissenschaften, Sonderheft 10:11–34, 2008; Prenzel M, Artelt C, Baumert J, Blum W, Hammann M, Klieme E et al (eds), PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie [PISA 2006. Results of the third international comparison]. Waxmann, Münster, 2007).

The choice of a scaling and trend analysis model depends on the focus of the analysis and on the assessment design. With respect to international comparisons, very strict assumptions on the uni-dimensionality of the test instruments used have to be made to allow for trend analyses. What if these conditions are not met across all participating countries for all assessment cycles? This paper presents an alternative model for trend analyses, assuming uni-dimensionality only within a particular country but not across all participating countries. Trend results with this model can only be interpreted within the particular country and are not intended for use in international comparisons.

C.H. Carstensen (✉)
Institute of Psychology, University of Bamberg, Bamberg, Germany
e-mail: claus.carstensen@uni-bamberg.de

M. Prenzel et al. (eds.), *Research on PISA: Research Outcomes of the PISA Research Conference 2009*, DOI 10.1007/978-94-007-4458-5_12,

To establish the validity of the presented trend model, an empirical analysis of the different tests and subscales used in different assessment cycles was performed. As far as different versions of the instruments were administered within cycles, the correlations of these test forms give an empirical indication of the uni-dimensionality of the underlying constructs. Monte Carlo simulations were performed to analyse whether the correlations of these test forms indicate a uni-dimensional construct being measured over time. Having analyzed the correlations of the tests, a fit analysis at the item level followed. Further assumptions refer to the stability of item difficulties over time. This is addressed by estimating item by time interaction parameters, allowing for a descriptive analysis of items changing their difficulty over time and a model fit comparison to check whether item drift has an impact on their difficulties.

Results show that trends might be reported for the German data, using the short test for reading and using all pair wise link items in Mathematics and Science. In the conclusion, the results and some implications for the design of future PISA assessments will be discussed.

Keywords Trends • Scaling models • Measurement invariance

12.1 Issue: Trends from PISA Data

From a methodological perspective the question of trends is clearly distinct from the question PISA data have to answer in the first place: each PISA cycle compares student performance with the purpose of country comparisons. For country comparisons within a PISA cycle, the first optimization criterion for study design, scaling and analysis procedures would be the comparability across countries. The competencies measured need to have the same meaning and be interpreted identically within each country. In contrast, for trend analyses, the highest priority in optimizing study design, scaling and analysis procedures would be comparability across cycles; the competencies found in each cycle need to have the same meaning and be interpreted identically across cycles. In case researchers find that the proficiency distributions from a trend study cannot be analyzed under both perspectives, they will have to decide according to which perspective the data analysis procedures shall be optimized.

The rationale of trend analyses is to keep the instrument and assessment conditions the same across different studies and then to assume that any change in the item response frequencies is due to a change of proficiencies of the sampled populations. Hence, a prerequisite of trend analyses is to prove the equal characteristics of the common instrument or measurement invariance across studies (Kolen & Brennan, 2004). In the remainder of this section the invariance of the measurements within the first three PISA cycles will be discussed.

In order to provide differential information on student's competencies, the focus of the competence assessments varies across the cycles: in PISA 2000, reading literacy was the major domain of the study, which can be seen from the number of items among other criteria. In PISA 2003, mathematical literacy was the major

Table 12.1 Domains, numbers of items and framework development for the first three PISA cycles

	PISA 2000	PISA 2003	PISA 2006
Reading	**Major** 129 items full framework	Minor 28 link items 00/03/06	Minor 28 link items 00/03/06
Mathematics	Minor 20 link items 00/03	**Major** 84 items full framework	Minor 48 link items 03/06
Science	Minor 22 link items 00/03	Minor 25 link items 03/06	**Major** 108 items full framework

domain, and scientific literacy in PISA 2006. Thus, different test instruments were used in different cycles: instruments of the major domain comprise a large number of items; these tests are called long tests in this chapter. If a domain was assessed not as the major domain, a smaller number of items was administered in the assessment; these tests are called short tests in this chapter.

With each of the three domains being a major domain over the course of the 2000, 2003 and 2006 PISA cycles, the fully elaborated assessment frameworks were worked out successively in parallel with the study focus: the fully detailed framework for mathematical literacy was presented 2003 (OECD, 2003) and the fully detailed framework for science was presented in 2006 (OECD, 2006). As a consequence, the short tests in Mathematics and Science administered in the studies before these domains were the major domain are not necessarily subsets of items of the respective long tests. Moreover, the selection of items for the short tests could not be related to the respective long tests and the number of common items between cycles is smaller than necessary for a stable link. In contrast, the short test in reading was designed based on the results of the long test.

Table 12.1 gives an overview over the three domains and the three cycles, as well as over the framework development, the number of items in the long tests, and the number of common items.

The same short test for Reading was administered in PISA 2003 and PISA 2006. It consists of 28 items selected from the PISA 2000 reading assessment. For mathematics, 84 items were administered in the long test and 48 of these items were selected for the PISA 2006 assessment. However, only 20 items of the 34 items of the PISA 2000 Mathematics assessment also appear in the long test. Only eight items appear in both the PISA 2000 and the PISA 2006 Mathematics assessments. The long test in Science includes 108 items, the PISA 2003 short test consists of 25 items which also appear in the PISA 2006 assessment and nine unique items. The PISA 2000 science assessment has 22 items in common with PISA 2003 and 12 items in common with PISA 2006 and no unique items.

Given the assessment design of the three PISA studies, the OECD (2007) reported trends over the first three cycles for reading only, trends for Mathematics were reported from PISA 2003 to PISA 2006 and for Science no trends were reported. Gebhardt and Adams (2007) investigated the impact of different scaling and linking methodology on trend results. They found that using different instruments in different

Table 12.2 Mean difficulties for the common and unique items from PISA 2000

Reading assessment mean difficulties	OECD	Germany	Sweden	Mexico
Link items 2000/2003	−0.03	−0.05	−0.20	0.18
Unique items 2000	0.01	0.01	0.06	−0.05
Relative difficulty link items	0.04	0.06	0.26	−0.23

cycles for the same competence, like the long reading test and the short test in PISA 2003 and PISA 2006 may have led to biased results.

Gebhardt and Adams (2007) investigated item difficulties across countries for the first two assessment cycles (PISA 2000, 2003). They compared the mean difficulty of the common and the unique parts of the reading and science assessments across countries. Within all OECD countries, the difference between the common and the unique items in Reading in PISA 2000 is 0.04 logits, so the common items are slightly easier than the unique items. In the PISA 2003 reading assessment, only the common items were administered again. Therefore, students participating in PISA 2003 have a slight advantage in solving the items compared to the PISA 2000 participants. Through appropriate scaling and linking (OECD, 2005, 2009), however, the scale scores are in the same metric. If we look at a particular country, though, the difference between the common and the unique items appears different. For example, the mean item difficulties for Germany, Sweden and Mexico are presented in Table 12.2. In Sweden, the common items presented in PISA 2003 are 0.26 logits easier than the unique items, which in turn is 0.21 easier than in the OECD. Consequently, the Swedish students gain an advantage from switching to the short test. For Mexico, the short test is harder than the long test by 0.23 logits and consequently the disadvantage of the Mexican students in PISA 2003 is 0.27 logits (Gebhardt & Adams, 2007). For Germany, the difference to the OECD is only 0.02 logits and hence there is hardly any advantage or disadvantage gained from switching to the short test.

Gebhardt and Adams investigated the impact of these advantages and disadvantages on the trend estimates. They compared trend results from three different methods: the original scaling reported in OECD publications and two further scaling methods. Both latter methods (which will be further illustrated below) include the rescaling of the data for each country, so that the mean difference between the common and the unique items is modeled for each country individually and does not reflect the average OECD value of this difference. The main results of their study are that trend results are significantly different between the original scaling (with reference to the OECD value of the mean difference) and the alternative methods with country-specific mean difference treatment in 6 out of 28 countries. For the science assessments from PISA 2000 and PISA 2003, Gebhardt and Adams found significant differences in trend results for 2 out of 25 countries. These results may indicate the extent to which trend results in the PISA studies are variable conditionally on the scaling method.

In addition, other factors might have an impact on trend results as well. Gebhardt and Adams investigated the influence of different sample characteristics (such as the number of public and private schools or the distribution over socio-economic backgrounds) between cycles on trend results. Another factor might be seen in item-by-study-by-country interactions in the item difficulties, like item drift over time. Further assumptions for comparability across cycles have to be made with respect to the booklet design, i.e. the rotation of clusters within booklets and the positions of items within a cluster have to be the same for the common items.

As said before, trend analyses of the data collected in the PISA cycles require measurement invariance. Measurement invariance in IRT models can be assumed if the same item response model holds for all measurement occasions, i.e. for all studies in all countries, which can be assumed if the item difficulty parameters are the same for each item across studies and across countries. As shown by Gebhardt and Adams, this assumption does not hold with respect to the mean item difficulties between common and specific items for all countries. This chapter will investigate whether a model for trend analyses without assuming item parameter equality within and across cycles will allow trend analyses within a country. The research question in this chapter is whether an appropriate scaling method (modeling the mean differences mentioned above for each country) will prove to be a reliable basis for trend analyses for the German PISA data across PISA 2000, PISA 2003 and PISA 2006.

12.2 A Model for National Trend Analyses

12.2.1 IRT Scaling Model

The trend model to be investigated in this chapter applies a concurrent scaling to the data of the three PISA cycles from German students only. This model was introduced as the marginal trends model by Gebhardt and Adams (2007). In the marginal trends model, the response data from the three PISA cycles of interest, that is common and unique items, are calibrated concurrently, with the item difficulties of common items being assumed to be equal across cycles.

The proficiency distributions were estimated in two steps. In a first model (model 1) item parameters were estimated from a dataset with student responses from three cycles for each domain using a Rasch type model for dichotomous responses and a partial credit model (Masters, 1982) for items with three- or four-point scores using the ConQuest 2.0 software (Wu, Adams, Wilson, & Haldane, 2007). For PISA 2000 data, booklet effects were estimated as well, since in PISA 2000 item difficulties are confounded with booklets (Adams & Carstensen, 2002). For PISA 2003 and PISA 2006, no booklet effects were estimated, since in these studies they have no impact on item parameter estimates and proficiency distributions (OECD, 2005, S. 198, 2007).

In a second step (model 2), item parameters were kept fixed and student proficiency distributions were estimated conditionally on study, type of school and

their interaction effects. Plausible values (Adams & Wu, 2002) were derived and transformed into a metric for reporting trends. A metric with a mean of 100 and a standard deviation of 30 was chosen to make it obvious that these estimates were not obtained from the OECD scaling model.

This trend model differs from the model of Gebhardt and Adams (2007) in two respects: their conditioning model includes the students' age and gender, the socio-economic status of their parents, migration background and dummy variables for missing responses. Secondly, in this chapter three uni-dimensional models were used for reading, mathematics and science instead of the three-dimensional model of Gebhardt and Adams. Since trends will only be reported in terms of means and variances, no substantial differences are expected due to these minor differences between the trend models.

The proposed model analyzes data from all cycles concurrently. If results from earlier cycles have already been published, it might be impossible to report numerically different results for that wave from a new calibration. Xu and von Davier (2008) discuss different models for item parameter estimation in a comparable setup. If item parameters are fixed to their values from the first assessment, the linked item parameters for later cycles as well as the ability distributions and their changes differ significantly from the respective values if item parameters are estimated for all assessments concurrently. To extrapolate the trends proposed herein for the time after PISA 2006, one has to decide whether further concurrent calibrations may be performed with the data collected from further cycles, which may change results already published from previous cycles or if other linking models have to be adopted for continuing trend reports.

12.2.2 Trend Model Validity and Fit

The following section discusses whether crucial assumptions made in the trend model described above hold for the German data. This includes a discussion of the compatibility of the assessment frameworks across studies, an analysis investigating whether the common and unique items form uni-dimensional scales within each cycle and an analysis of item fit (both questions of construct validity) and an analysis whether item-by-study interactions are negligible (a question of trend model fit).

With the trend model presented, a uni-dimensional scale based on items across studies for each domain within a single country will be established. Note that with the cross-sectional comparisons reported, the assumption of uni-dimensionality has been assessed within cycles across countries. To validate a trend model, the definition of the domains across cycles needs to be consistent. Since the assessment frameworks have been developed over the studies, this consistency will be discussed for each domain. Reading was the focus in PISA 2000. The assessment instrument comprises 129 items and can be analyzed with respect to different reading subscales. In the 2003 and 2006 PISA assessments, the same short test was administered,

consisting of 28 items[1] which were selected to represent the reading scale as a whole. Thus, the reading proficiency scales of the three PISA studies might be linked using the short test from all three cycles. In order to use all items available, the long test from PISA 2000 might be linked with the short test from the following two cycles. Whether the short test and the long test measure the same construct empirically can be analyzed from the PISA 2000 data. Results of this analysis will be reported in the following section.

The mathematics assessments used three different tests, two different short forms and a long test in PISA 2003. The long test included 84 items and differentiated four subscales. The short test for PISA 2006 consisted of 48 items balanced over these subscales. However, the PISA 2000 short test consisted of 31 items and basically includes two of the subscales. It is not balanced with respect to the mathematical content areas defining the subscales. Furthermore, it shares 20 items with the long test and only 8 items with the PISA 2006 short test. The OECD reports on trends in mathematics refer to the two subscales included in the 2000 short test. In contrast, the analyses presented here link both short tests to the PISA 2003 assessment. Whether these tests measure the same construct of mathematical literacy will be investigated as an empirical question in the following section.

The full framework for scientific literacy was developed for the PISA 2006 study. Nevertheless, it is largely consistent with the prior frameworks (OECD, 2006, p. 25), and the science tests from the three studies are thus constructed rather consistently with respect to the combined science score. The 2000 and 2003 short tests share 25 items. The long test shares 14 items with the PISA 2000 short test and 22 items with the PISA 2003 short test. Just as for reading and mathematics, results from an empirical analysis of the factorial validity will be presented for science in the following section as well.

12.3 Results from German Data

For using the trend model presented, items from different tests are selected and analyzed together. Hence, strictly speaking, new tests are constructed. For the domains of mathematics and science it is furthermore assumed that these new combinations of items measure a common construct within each domain. In order to investigate whether these assumptions hold, the results of empirical analyses of the factorial validity of the new tests and of item fit analyses are presented. Moreover, an analysis of whether the items from the three assessments in each domain form a common scale using item fit statistics will be undertaken. Finally, the trend results from German data will be presented.

[1] Due to deletion of one reading item for the German data set, the short test in the following analysis includes 27 items.

Table 12.3 Estimated correlations and bootstrap results for five links in the trend model: data source, contrast, observed correlations, smallest correlation from r = 100 bootstrap samples and number of unique and link items in the booklet design

Data	Contrast	Observed correlations	Smallest corr. from bootstrap	No of items: (unique/link)
PISA 2000	Reading unique vs. link 00/03/06	0.926	0.942	101/27
PISA 2003	Mathematics unique vs. link 00/03	0.944	0.948	64/20
PISA 2003	Mathematics unique vs. link 03/06	0.970	0.970	36/48
PISA 2006	Science unique vs. link 03/06	0.955	0.934	81/22
PISA 2006	Science unique vs. link 00/03/06	0.960	0.908	89/14

12.3.1 Empirical Analysis of the Factorial Validity

To assess empirically whether the common and the unique items from the tests linked in the trend model form uni-dimensional scales, the following analyses were performed: latent correlations were computed for a two-dimensional model which contrasts the unique items and the common items for each test in the trend model. The correlations between the common items and the unique items have a sampling variance due to sampling of responses and due to measurement error, especially for links based on small numbers of common items. If both sets of items measure the same construct, those correlations should be maximal, that is not significantly different from r = 1. To obtain confidence intervals for these correlations, a bootstrap procedure was applied. The bootstrap model had a simplified set-up without creating a design for non-administered items from a multi-matrix design. The number of items for each bootstrap model was chosen to correspond to the number of unique and link items for each link evaluated. Particularly, the bootstrap procedure was based on the average number of link and unique items administered to students through all booklets for a particular domain. The sample size for each domain reflected the sample sizes of the respective PISA assessments as defined by their booklet design. The data sets were generated according to a two-dimensional Rasch model in which true values for the item parameter and ability distributions were generated for each replication, thus implementing a non-parametric set up using the "simulate" option of the Conquest software. Standard PISA analyses of correlations do not reflect dependencies of item responses due to unit design, so neither does the bootstrap design. However, not reflecting item dependencies and, possibly, fatigue effects and others in the bootstrap design might result in higher correlations and thus might suggest too liberal decisions in detecting non-equivalence of link and unique items.

In Table 12.3, the estimated latent correlations and the results of the bootstrap procedure are shown with respect to five correlations: one for the link in reading

across all three cycles; for mathematics one for the link between the first and the second cycles and one for the link between PISA 2003 and PISA 2006; for science one for each link between PISA 2003 and PISA 2006 and one for all three cycles. All correlations were computed from data taken from the study where a domain was the major domain. In the model for reading the booklet coefficients were omitted since they would have had no impact on the dimensionality. The smallest correlation from 100 generated data sets for evaluating the correlation between both parts of the reading test of PISA 2000 was found to be 0.942. The observed latent correlation for reading from PISA 2000 data cannot be found within the range of the correlations from 100 replications. Hence, the observed correlation is statistically lower than $r = 1$ and the two dimensions, the set of common items and the set of unique items, are not the same. The assumption that both parts of the reading test can be seen as parts of the same instrument does not hold for the German data.

For mathematics, the picture is different: the observed latent correlations from both links are quite close to the values of the bootstrap analysis and it can be concluded that with a probability of around $p = 0.01$ the observed correlation is from within the range of observable correlations if the generated correlation is $r = 1$. For the trend model, the common and the unique items in mathematics are used for both links. Both correlations for science are well inside the range of correlations from the bootstrap and thus it can be concluded that both parts of the science tests measure the same construct in the German data. These finding are not completely in line with the expectations from reviewing the frameworks; for reading, a good connection between both parts of the assessment had been expected. The links in mathematics between PISA 2000 and PISA2003 and the links in science between PISA 2003 and PISA 2006 were expected to be a bit weaker. As a consequence of these analyses, the trend model for Germany in reading will be computed based on the common items only, i.e. the short test from all three cycles. For the trend model in mathematics and science, all common and unique items administered in one or more cycles will be used.

Assuming that the tests for the trend model are uni-dimensional, the fit of single items into the scales can be assessed empirically. The PISA consortium (Adams & Wu, 2002) uses, among others, the "weighted mean square residual fit index" (Wright & Masters, 1982), which basically evaluates the discrimination of each item. Inspecting these values for the items of the trend model under investigation, only a few items show indications of misfit: 2 out of 27 reading items, 5 of 95 mathematics items and 3 out of 124 science items show significantly low fit values. The total percentage of significant fit values is 4% which is less than expected assuming a conventional 5% error probability. Hence, no items have been removed from the trend models because of item fit.

With a trend analysis, the difficulties of item responses are evaluated over time. If all items become easier by the same degree, the change may easily be attributed to a change in a population's proficiency. If, however, items change differently in their difficulty, that is if there is an item by study interaction, the change cannot be attributed to a single dimension. In order to evaluate whether item by study interactions have an impact on the trend model, the following analysis was performed: based on

Table 12.4 Linking errors for three domains: domain, link, link error and number of link items

Linking errors			
		Link 2000 2003	Link 2003 2006
Reading	Error	4.73	4.67
	# link items	27	27
Mathematics	Error	4.43	2.17
	# link items	20	48
Science	Error	4.38	3.33
	# link items	24	22

Note: The link errors are in a metric with SD = 100 to be compared to OECD PISA values

the scaling (calibration) model for estimating the trends, two further models were estimated for each domain. For both models, the item parameters are held fixed at values from the calibration model and no conditioning model is specified. With the first model, the study is included as a conditioning variable for item difficulties to capture overall changes in the populations over time; this model does not assume an item by study interaction. With the second model, item by study parameters are introduced additionally. Comparing the fit of both models allows evaluation of whether item by study interactions have a significant impact on the item difficulties of the common items estimated from the German data.

Evaluating the item by study parameter estimates descriptively, one finds that a small number of these estimates are larger than 0.3 logits or smaller than −0.3 logits. For the science test, most of these estimates are even in the range from −0.2 to 0.2 logits. Linking errors were computed for the link between each of the two pairs of consecutive cycles and each domain. These link errors are displayed in Table 12.4, in a scale with SD = 100 to enable comparison with linking errors for PISA reported by the OECD. Linking errors for original trends in PISA (OECD, 2005, table 12.28) vary from 1.38 points (mathematics from 2003 to 2006) to 5.31 points (reading from 2000 to 2003) in their reporting scales. Monseur and Berezner (2007) compute linking errors using jack-knife techniques to reflect the item structure in units, item by country DIF and partial credit items. They report link errors for reading at the country level from about 6 to 12 points, for Germany they find an error of 9.54 points. The link errors of the linking model proposed here for the German data are in the same magnitude as the linking errors for the international trend model.

Table 12.5 displays results for model fit comparisons for reading, mathematics and science. It lists the model estimated, the number of students, the difference in parameters between both models for each domain, the deviance (−2ln likelihood) and the CAIC information criterion (Bozdogan, 1987). The CAIC indicates a better fit of the model with the smaller index value and is computed with respect to the difference in likelihood and number of parameters of the models compared given a sample size of the analyzed data set. However, it does not make assumptions about the distribution of the index values and does not provide a test for the significance of differences.

Table 12.5 Model fit results for three domains: model, number of model parameters, deviance (=−2log*L*) and CAIC, the sample size is N = 14,624 for each model

Model	# of par.	−2 ln*L*	CAIC
Reading			
Item + cycle	4	322055	322096
Item + cycle + cycle x item	4 + 54	321765	322359
Mathematics			
Item + cycle	4	259540	259581
Item + cycle + cycle x item	4 + 68	259269	260013
Science			
Item + cycle	4	281626	281667
Item + cycle + cycle x item	4 + 46	281442	281954

For each domain in the first model, four parameters are estimated – a study mean, two differences between study means and a variance – while item parameters are fixed to their values from the calibration. With each second model, a parameter for each common item and occasion is estimated. In reading, 27 items of the long test were administered in PISA 2003 and in PISA 2006, resulting in 54 item by time parameters. For mathematics, 68 item x study parameters and for science, 46 item x study parameters were estimated. For all three domains, the CAIC values of the item by study interaction model are bigger than the index value from the non interaction model. This indicates a better fit of all three non-interaction models. Given these results and the comparison of the link errors of the proposed model with PISA original scaling link errors, the item by study interaction parameters are assumed to be negligible with respect to measuring trends and all common items are linked by restricting their difficulty to be the same over cycles.

12.3.2 Trend Results

In the following section, the trend results for German data estimated using the proposed trend model (Carstensen, Prenzel, & Baumert, 2008) are reported. Due to the estimation of country-specific item difficulty parameters, the proficiency scales reflect curricular and cultural characteristics of the German educational system. Therefore, the scale values are not to be directly compared to international PISA scale values. To remind the reader of this, the trend scale values are reported in a metric with a mean of 100 points for the reference study and a pooled standard deviation of 30 over all three cycles.

In Table 12.6, the trend results are printed for all three domains. According to our trend model, the mathematical competencies of the 15-year-olds in Germany have increased over cycles. The PISA 2003 mean is set to 100; the PISA 2000 mean of 93 points is significantly lower with standard errors of 0.8 and 0.7 for both means.

Table 12.6 Trend results for reading, mathematics and science in Germany: Mean, SE and SD for each study and a linear trend estimate

	PISA 2000			PISA 2003			PISA 2006			Trend	
	mean	*SE*	SD	mean	*SE*	SD	mean	*SE*	SD	mean	*SE*
Mathematics	93	*0.8*	26	100	*0.7*	32	101	*1.2*	31	4.4	*0.7*
Science	100	*0.7*	26	104	*1.1*	33	107	*1.1*	30	2.4	*0.8*
Reading	100	*0.8*	26	100	*1.0*	32	100	*1.3*	32	−0.0	*0.7*

Notes: Means constrained on M = 100 for one assessment cycle, standard deviation fixed to SD = 30 over three cycles; concurrent calibration for German data

The PISA 2006 mean at 101 points is numerically higher than the PISA 2003 mean, but this difference is not statistically significantly different from zero. A linear trend was estimated over the three cycles (based on a dataset with plausible values from all three cycles) as well. The linear increase between two cycles is 4.4 points, which equals d = 0.15 in terms of effect size. Given its standard error of 0.7 points, this increase is statistically significantly larger than zero: on average, there is a significant increase in mathematical literacy over cycles in Germany.

The OECD reported trends in mathematics on the overall scale between PISA 2000 and PISA 2003[2]; in these reports, Germany gained about one point (in the SD = 100 metric; see Prenzel et al., 2007) which converts to 0.24 points in the SD = 30 metric. Comparing the two trend estimates, we find positive values neither of which are significantly different from zero.

For scientific literacy the German trend estimates show an increase over cycles: with the mean value for PISA 2000 being fixed to 100, the means increase to 104 and 107 points respectively, and the linear trend is a 2.4 points increase between cycles. All mean differences between cycles as well as the trend are significantly different from zero. Gebhardt and Adams (2007) also report a significant increase between PISA 2000 and PISA 2003 for science in Germany; in the OECD scaling (OECD, 2004) we find a significant increase between these two study means as well. As far as any are available, the different trend estimates consistently show an increase in scientific literacy.

The reading proficiency of the fifteen-year-olds in Germany did not change significantly over cycles. The mean value from PISA 2000 is set to 100, the mean values from the other two cycles are both 100 points as well, so the linear trend is also zero. This result is somewhat in contrast to the trend computed from the OECD scaling, in which the study means for Germany are 484, 491 and 495 points, showing a numerical increase. However, neither of the study mean differences are statistically significant. Gebhardt and Adams found no increase in reading literacy from PISA 2000 to PISA 2003 in Germany, which is consistent with the national trend

[2] Earlier trend reports were restricted to two subscales which were assessed with sufficiently large item numbers.

model presented in this chapter. This inconsistency between national trend estimates and trends from OECD scaling values is due to the different trend models implemented. The national model re-estimates item parameters specifically for a country, which is in case of differences the model fitting more closely and thus more reliable. Trends from OECD scaling values suffer from item by country interactions in item difficulty and especially from combining the use of different test forms (long and short form). Making the PISA test comparable across countries reduces the stability of trend analyses if the item difficulties vary over countries. Again, such variation might be due to cultural and school system factors and could be the result of sampling variation of items as well. As Gebhardt and Adams showed, the OECD scaling trend results for Germany are biased and overestimate the performance of German students in PISA 2003 and PISA 2006.

12.4 Discussion

This chapter has addressed the question of an adequate trend model for German data from the PISA 2000, PISA 2003 and PISA 2006 studies. In general, trend analyses within a country are a task with different requirements in contrast to comparing countries using cross-sectional results from PISA studies. Different models for scaling trend data and analyzing trends have been discussed with respect to the PISA trend design; in the context of PISA, Gebhardt and Adams (2007) discuss the appropriateness of the original OECD scaling, a concurrent calibration and a conditional analysis of the concurrent calibration for each country in PISA 2000 and PISA 2003. Xu and von Davier (2008) elaborate on different linking models, Monseur and Berezner (2007) examine the effect of omitting link items and Mazzeo and von Davier (2008) discuss the assessment design and analysis models for trends in PISA in comparison to the National Assessment of Educational progress (NAEP) in the USA. They point out, that a very conservative test design with minimum changes over assessment cycles is a key for the stability of trends in NAEP since reporting trends requires more precision in the proficiency distribution estimates than reporting country comparisons.

In this chapter, a national trend model for reporting trends for Germany (Prenzel et al., 2007) and for the German federal states (Prenzel et al., 2008) has been presented and its fit to the German data has been investigated. The model estimates item parameters for German data concurrently for all three cycles and is based on the marginal trends by Gebhardt and Adams (2007). As a result of empirical analyses of factorial validity and item fit analyses, the model was estimated using the link items only for Reading and all available items, link and unique items, for Mathematics and Science. Furthermore, the difference in the underlying construct between the long and the short reading tests in German data became evident and thus the national trend model was restricted to the short test with identical items in all three cycles. This result had not been expected, since the same framework was used to construct both the long and the short tests. The reading assessment was constructed according

to a fully detailed framework with a balanced items distribution over subscales. However, the short test does not seem to measure the same construct as the long test in Germany. In contrast, the long versions for mathematics and science do measure the same construct respectively, according to the empirical analyzes, while the assessment instruments were in part constructed according to not yet fully detailed framework versions.

For mathematics and science, the trend results from the national trend model were rather consistent with results from Gebhardt and Adams and with results from the OECD scale values. Only for the domain of reading were the trend results from the OECD report different from the national calibration and Gebhardt and Adams´ marginal trends; these differences may be due to the different test instruments used in the different trend models.

With respect to implications for trend reports from PISA cycles, modeling competencies on the basis of different test instruments over cycles will be a fundamental challenge. Other challenges are obviously variations of item difficulties across countries within a study (country DIF) and across cycles within each country (item drift). An essential prerequisite for providing reliable trends seems to be a test design that administers as many link items as possible in exactly the same set-up within booklets over cycles. This issue is not at the focus of the present chapter and given all restrictions in constructing PISA assessments, this might be the hardest challenge to master. From the perspective of trend analyses, it seems to be of special importance to ensure a construction of test instruments that implement link clusters of items for each domain which are held constant over cycles. Ideally, even the assignment of link clusters to booklets might be kept constant, resulting in link booklets.

Depending on the booklet designs of consecutive PISA cycles, appropriate scaling models for trend analyses have to be developed. One way of thinking might be to accept different models for different questions and to report cross-sectional results on the international PISA scale, while trend results are reported on national scalings. This strategy would provide results with a high degree of fit between data and scaling results; however, implementing it would make it necessary to communicate the rationale for different scaling models to the public.

Another way of thinking might be to relate the cross-sectional scaling and the trend scaling as closely as possible. To address the major difference between national or marginal trend models and original trend or reports from OECD scalings, basing cross-sectional results on the same set of items as trend model results, the items in link clusters only, might be an option. However, a large number of items in the assessment of a major domain would then be omitted in the scaling of proficiency distributions for the combined scales. Instead, these items might then be constructed more independently to assess more distinct subscales or variations from the combined scale. Even if both models, for cross-sectional comparison and trends, were based on the same set of link items only, any country DIF would still be a threat to consistent results for both purposes. However, the rather consistent results for mathematics and science give an indication that country DIF might be a source of much smaller inconsistencies as different test forms (in reading) are. This is one of many questions for future research.

References

Adams, R. J., & Carstensen, C. H. (2002). Scaling outcomes. In R. J. Adams & M. Wu (Eds.), *PISA 2000 technical report*. Paris: OECD.

Adams, R. J., & Wu, M. L. (2002). *PISA 2000 technical report*. Paris: OECD.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*(3), 345–370.

Carstensen, C. H, Prenzel, M., & Baumert, J. (2008). *Trendanalysen in PISA: Wie haben sich die Kompetenzen in Deutschland zwischen PISA 2000 und PISA 2006 entwickelt?* [Trend analyses in PISA: How did competencies in Germany develop between PISA 2000 and PISA 2006?] Zeitschrift für Erziehungswissenschaften, Sonderheft 10/2008, pp. 11–34.

Gebhardt, E., & Adams, R. J. (2007). The influence of equating methodology on reported trends in PISA. *Journal of Applied Measurement, 8*(3), 305–322.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling and linking*. New York: Springer.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*. Princeton, NJ: ETS.

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

Monseur, C., & Berezner, A. (2007). The computation of equating errors in international surveys in education. *Journal of Applied Measurement, 8*(3), 323–335.

OECD. (2003). *The PISA 2003 assessment framework – Mathematics, reading, science and problem solving knowledge and skills*. Paris: OECD.

OECD. (2004). *Learning for tomorrow's world: First results from PISA 2003*. Paris: OECD.

OECD. (2005). *PISA 2003 technical report*. Paris: OECD.

OECD. (2006). *Assessing scientific, reading and mathematical literacy. A framework for PISA 2006*. Paris: OECD.

OECD. (2007). *PISA 2006: Science competencies for tomorrow's world* (Analysis, Vol. 1). Paris: OECD.

OECD. (2009). *PISA 2006 technical report*. Paris: OECD.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., et al. (Eds.). (2007). *PISA 2006. Die Ergebnisse der dritten internationalen Vergleichsstudie* [PISA 2006. Results of the third international comparison]. Münster, Germany: Waxmann.

Prenzel, M., Artelt, C., Baumert, J., Blum, W., Hammann, M., Klieme, E., et al. (Eds.). (2008). *PISA 2006. Die Kompetenzen der Jugendlichen im dritten Ländervergleich*. Münster, Germany: Waxmann.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: University Press.

Wu, M., Adams, R. J., Wilson, M., & Haldane, S. (2007). *Conquest 2.0*. Camberwell, Australia: ACER Press.

Xu, X., & von Davier, M. (2008). Linking for the general diagnostic model. In *IERI monograph series: Issues and methodologies in large-scale assessments* (Vol. 1, pp. 97–112). Princeton, NJ: IEA-ETS Research Institute.