Psychological Test and Assessment Modeling, Volume 55, 2013 (2), 181-206

Linking PISA 2000 and PISA 2009: Implications of instrument design on measurement invariance

Eunike Wetzel[1] *& Claus H. Carstensen*[2]

Abstract

An important pre-requisite of trend analyses in large scale educational assessments is the measurement invariance of the testing instruments across cycles. This paper investigates the measurement invariance of the PISA 2000 and PISA 2009 reading instruments using Item Response Theory models. Links between the PISA 2000 and PISA 2009 instruments were analyzed using data from a sample tested in 2009 which took both the PISA 2000 and PISA 2009 instruments and additionally using part of the German PISA 2000 sample. Model fit comparisons showed that the instruments are not measurement invariant and that some link items show large differences in item difficulty. Position effects may explain some of these differences and may also influence the size of the link error.

Key words: PISA, measurement invariance, linking, link error, position effects

[1] *Correspondence concerning this article should be addressed to:* Eunike Wetzel, Dipl.-Psych., Department of Psychology and Methods of Educational Research, Otto-Friedrich-University Bamberg, 96045 Bamberg, Germany; email: eunike.wetzel@uni-bamberg.de

[2] Otto-Friedrich-University Bamberg, Germany

Introduction

The introduction is structured into three sections. First, we will give a brief overview of the goals and study design of the Programme for International Student Assessment (PISA). Second, we will describe the linking of scores from different PISA assessments and introduce the computation of the link error, and third, we will present the aims of our study and our research questions.

Goal and study design of PISA

Starting in the year 2000, the Organisation for Economic Cooperation and Development (OECD) has been conducting the Programme for International Student Assessment (PISA) which assesses 15-year-olds every three years in the domains of reading, mathematics, and science. The aim of PISA is to measure life skills that enable people to succeed in modern societies (e.g., OECD, 2009a). Accordingly, PISA requires students to evaluate material and apply it to new situations. The three domains are defined in terms of a literacy concept similar to the one developed by previous surveys, for example the International Adult Literacy Survey (IALS; e.g., OECD & Statistics Canada, 2000). Reading literacy is characterized by a person's capacity to "understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD, 2009a; p. 14). Mathematical literacy is defined as "an individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen." (OECD, 2009a; p. 14). Scientific literacy comprises "an individual's scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues, understanding of the characteristic features of science as a form of human knowledge and enquiry, awareness of how science and technology shape our material, intellectual, and cultural environments, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen." (OECD, 2009a; p. 14).

In PISA, the main focus of the study alternates. In 2000 it was reading, in 2003 mathematics, and in 2006 science. With the completion of the fourth PISA assessment in 2009, a new cycle has begun in which reading was once again the first major domain. The major domain is assigned more testing time than the minor domains. In general, items are nested in units (e.g., items that refer to the same text passage) and several units compose a cluster. The items in one cluster all assess the same domain. Each test booklet contains four clusters. The test booklets are randomly assigned to the students participating in PISA. Comparisons of student achievements in the three domains across the participating countries have been drawn from the first PISA study in 2000 and continue to give important information regarding the standing of students in one nation compared to the students in other nations. Another central goal of PISA which is increasingly taking priority is conducting trend analyses. Trend analyses aim at investigating how student

achievements develop within participating countries over assessment periods (OECD, 2010). Trend analyses (with regard to the whole population or subpopulations) carry critical implications as they can be used to monitor the success of reforms in educational systems. For instance, policy makers may be interested in whether the proportion of low-achieving students has decreased or whether the potential gender gap in achievement has narrowed or widened.

Linking and the link error

Conducting methodologically sound trend analyses is not an easy task. One pre-requisite for trend analyses is the measurement invariance of the instruments across assessments (Kolen & Brennan, 2004). In their review of the PISA test design, Mazzeo and von Davier (2009) list several criteria that need to be fulfilled to establish stable trends. These include that the same construct should be measured in all assessments and in all participating countries. Furthermore, the relationship between the items and the underlying latent trait should be unchanged across assessments for items that are used in several assessments. Also, item presentation should be standardized and comparable across countries and assessments.

To ensure the comparability of scores from different assessments, link items, which are common across assessments, are used. For example, 28 of the 129 reading items used in PISA 2000 were included in PISA 2003, 2006, and 2009. Changes in the difficulty of these link items determine the transformation used to equate scores from one assessment with scores from a previous assessment (OECD, 2012). Since the chosen link items are a sample of all possible link items, a different transformation would result if an alternative set of link items had been chosen. Thus, uncertainty is introduced to the process of equating scores across data collections. The precision with which scores from different assessments are aligned on one performance scale is captured by the link error (or equating error). The computation of the PISA 2003 link error was shown to be inadequate by Monseur and Berezner (2007), so it was modified to take into account that items are organized in units and that partial credit items have a greater influence on scores than dichotomous items. The improved link error estimate has been used to link PISA 2009 and PISA 2006 data to previous data collections and is described in the PISA technical reports (OECD 2009b, 2012). First, the difference in item difficulty $\hat{\delta}_{ij}$ between two assessments (e.g., PISA 2009 and PISA 2006) is computed $c_{ij} = \hat{\delta}_{ij}^{2009} - \hat{\delta}_{ij}^{2006}$ with *i* items in a unit and $j = 1,...,K$ units. The mean number of score points is $\bar{m} = \frac{1}{K}\sum_{j=1}^{K} m_j$. Further it is defined that $c_{\bullet j} = \frac{1}{m_j}\sum_{j=1}^{m_j} c_{ij}$ and $\bar{c} = \frac{1}{N}\sum_{j=1}^{K}\sum_{i=1}^{m_j} c_{ij}$. Then the link error can be computed as

$$error_{2009,2006} = \sqrt{\frac{\sum_{j=1}^{K} m_j^2 (c_{\bullet j} - \bar{c})^2}{K(K-1)\bar{m}^2}} \quad (1)$$

PISA reported the link error to be 4.07 for the reading scale 2006 to 2009 and 4.94 for the reading scale 2000 to 2009 (OECD, 2012). Thus, when taking only the link error into account, the 95% confidence interval of the difference in mean scores is about 20 score points wide (Wu, 2010). Monseur and Berezner (2007) also argued that the link error may be larger than the sampling error and the measurement error. The link error influences trend results and conclusions drawn from trend analyses and as such has an effect on actions taken by policy-makers. Gebhardt and Adams (2007) demonstrated that trend results differed depending on whether international item parameters were used or whether national item parameters were used in computation. Since link errors threaten trend analyses, both Mazzeo and von Davier (2009) and Wu (2010) recommend increasing the number of link items to reduce the link error.

Aims and research questions of this study

As linking is such an important aspect of trend analyses, this study investigates the linking of PISA 2000 and PISA 2009 reading and science items for two German samples. In 2009, the German PISA consortium conducted study in addition to the regular PISA 2009 assessment in which the PISA 2009 booklets as well as five selected booklets from the PISA 2000 assessment were administered to students at 59 German high schools. These 59 high schools had already participated in PISA 2000-E as part of an extended sample for state comparisons (Baumert et al., 2002). Thus, data were available from the same 59 high schools for two different time points, 2000 and 2009, as well as items from two different PISA instruments, namely the PISA 2000 and the PISA 2009 test booklets. This design allowed the measurement invariance of the PISA 2000 and PISA 2009 reading instruments to be investigated within one sample (the sample from 2009) as well as between samples within one instrument (PISA 2000). The five booklets from PISA 2000 applied in 2009 originally contained mathematics and science items at the last cluster position. The last clusters in these five booklets were replaced with science clusters from the PISA 2006 assessment, enabling us to analyze the measurement invariance of the PISA 2006 and PISA 2009 science instruments for 44 out of 53 science link items as well.

The aim of this paper is to test the measurement invariance of the reading items from PISA 2000 and 2009 regarding the common items and link items and the science items from PISA 2006 and 2009 regarding a subset of the link items. Our goal is to examine whether it is possible to establish a link and if so, which items are adequate for establishing a stable link. Furthermore, trend results will be reported and factors that influence linkability will be discussed in terms of how they affect the size of the link error. One conceivable influence on linking are position effects, i.e., the phenomenon that items have different difficulties, depending on their position in the test. For PISA 2000, Adams and Carstensen (2002) showed that differences in item difficulties between positions occurred for each of the nine reading clusters. Position effects are possible in PISA be-

cause clusters contain different units of items between assessments, as some items are replaced and as changes in testing time need to be accommodated when the major domain alternates. Thus, it will be analyzed whether differences in position may account for differences in item difficulties across assessments and instruments.

The samples used in this study allow the assessment of measurement invariance from two perspectives, first concerning the link and common items in the PISA 2000 and PISA 2009 instruments and second concerning the link and common items in the PISA 2000 reading instrument for which data was collected in 2000 and 2009. Thus, in sum, our two main research questions are 1) whether the instruments from PISA 2000 and PISA 2009 are invariant regarding the reading link and common items and whether the instruments from PISA 2006 and PISA 2009 are invariant regarding the science link items for the same study undertaken in 2009 and 2) whether the instrument from PISA 2000 is invariant between different studies (2000 vs. 2009) regarding the reading link and common items.

Method

Instrument

Reading clusters from PISA 2000 and PISA 2009 as well as science clusters from PISA 2006 and PISA 2009 were used. Table 1 lists the number of items linking the assessments. As the number of common reading items between 2000 and 2009 (39 items) is larger than the number of link items (28), analyses will be conducted (a) with the common items and (b) with the link items. Since only items being used repeatedly between assessments were analyzed, subscales for the different domains were not taken into account. A list of all the items included in our analyses as well as the item parameter estimates obtained from separate partial credit models in each of the subsamples can be found in the Appendix.

Table 1:
PISA Link Items across Assessments for the Three Domains

		Instrument			
		PISA 2000	PISA 2003	PISA 2006	PISA 2009
Domain	Reading	**129 items**	28 link items 00/03/06	28 link items 00/03/06	**39 common items 00/09, 28 link items 00/03/06/09**
	Mathematics	20 link items 00/03	**84 items**	48 link items 03/06	35 link items 03/06/09
	Science	25 link items 00/03	22 link items 03/06	**108 items**	53 link items 06/09

Note. Major domains are depicted in boldface and the absolute number of items is reported.

Sample

Two datasets were combined to obtain the dataset analyzed here. Both datasets were collected from 9th graders at the same 59 German high schools, though during different assessments. The first dataset ("study 2000") consisted of 1487 students (54.2 % female) who were regular participants of the PISA 2000-E (Baumert et al., 2002) assessment in Germany. The booklet design of the PISA 2000 study is depicted in Table 2. The second sample ("study 2009"; N = 1948, 53.6% female) formed an additional sample to the German PISA 2009 sample. For this second sample, both the 13 new PISA 2009 booklets (with regular difficulty; OECD, 2012) as well as five additional booklets (OECD, 2002) were applied (see Table 3). These 18 booklets were randomly distributed, resulting in a subsample of 1394 students who filled out the PISA 2009 booklets (booklets 1 - 13) and a subsample of 554 students who filled out booklets 14 to 18. Booklets 14 to 18 contained reading clusters from PISA 2000 at cluster positions one to three, regarding these three clusters they were identical to booklets 1 to 5 in the original PISA 2000 assessment (see Tables 2 and 3). The last cluster in the PISA 2000 booklets was originally used for mathematics and science items; for our study this cluster was replaced by a science cluster from the PISA 2006 assessment. To differentiate between the different item sets, each instrument will be referred to by its domain (reading or science) and PISA study year that the items originated from, e.g., "reading 2000" refers to the reading items from the PISA 2000 instrument. Thus, booklets 14 to 18 are a combination of reading 2000 (cluster positions 1 – 3) and science 2006 (cluster position 4).

Analyses

Measurement invariance was assessed from two perspectives. The first research question asked whether the instruments from PISA 2000 and PISA 2009 were measurement invariant regarding the reading items from the same study in 2009. For the science items, this question pertained to the instruments from PISA 2006 and PISA 2009. The second research question was whether the instrument from PISA 2000 was measurement invariant for different studies (study 2000 vs. study 2009). This question was analyzed using the reading items.

To answer these research questions, random coefficients multinomial logit models (RCMLM; Adams & Wilson, 1996) were estimated using ConQuest (Wu, Adams, Wilson, & Haldane, 2007). The RCMLM is a flexible generalization of the Rasch model (Rasch, 1960) which integrates other Rasch-type models such as the rating scale model (Andrich, 1978), the partial credit model (PCM; Masters, 1982), multifaceted models (Linacre, 1994), and the linear logistic test model (LLTM; Fischer, 1973). Thus, the RCMLM allows group differences (e.g., between study 2000 and study 2009) to be incorporated into the model as well as item by group interactions (differential item functioning).

The basic model for items with dichotomous response formats was the Rasch model (Rasch, 1960)[3] which models the probability that person *v* with person parameter θ_v will give a correct response to item *i* with difficulty δ_i:

$$p(X_{vi}=1 \mid \theta_v, \delta_i) = \frac{\exp(\theta_v - \delta_i)}{1+\exp(\theta_v - \delta_i)}. \quad (2)$$

Equation 2 can also be expressed in logit form:

$$\text{logit} = \ln \frac{p(X_{vi}=1)}{1-p(X_{vi}=1)} = \theta_v - \delta_i. \quad (3)$$

The item difficulty δ_i can further be parameterized to account for properties that certain items share (e.g., cognitive operations involved in solving them). In this case, the LLTM (Fischer, 1973) results: $\text{logit} = \theta_v - \sum_{k=0}^{K} \eta_k \omega_{ik}$ where η_k is a difficulty parameter for item property *k* and ω_{ik} represents an indicator weight of item *i* on item property *k* which takes the value 1 if item *i* belongs to property *k* and 0 otherwise. Two extensions of this model were compared to test measurement invariance. Model 1 consisted of a Rasch Model and a unique mean parameter β_g with *g = 1, ..., G* for the student performance distribution in the respective study or instrument:

Model 1: $$\text{logit} = \theta_v - \delta_i + \beta_g. \quad (4)$$

Model 2 additionally modeled the interaction between study or instrument and the difficulty of the item:

Model 2: $$\text{logit} = \theta_v - \delta_{ig} + \beta_g. \quad (5)$$

That is, in Model 2, differences in item difficulties (differential item functioning) between the studies or instruments were also estimated[4]. To evaluate the magnitude of these differences, the classification system for differential item functioning (DIF) developed by Educational Testing Service (ETS) was applied. In this classification system, items with DIF values below .25 contain negligible DIF, items with DIF values between .25 and .37 contain slight to moderate DIF, and items with DIF values equal to or above .38 contain moderate to large DIF (cut-off values were transformed from the delta scale used by ETS; Zieky, 1993). For reading, the two models were computed once with the link items and a second time using the common items.

ConQuest applies marginal maximum likelihood estimation using an EM algorithm (Bock & Aitkin, 1981) to estimate the item parameters and a normally distributed ability density. For the model comparisons, the mean of the item parameters was constrained to

[3] For partial credit items the partial credit model (Masters, 1982) was used.

[4] ConQuest model statements for the two models are: Model 1: item + item*step + instrument; Model 2: item + item*step + instrument + item*instrument

be zero for model identification purposes. Note that for the PCMs reported in the Appendix the model identification constraint was placed on the cases, yielding a mean latent variable of zero. Missing values were treated according to the PISA procedure (e.g., OECD, 2012). That is, responses to items that the student had reached and were missing or invalid were recoded as incorrect while items that the student had not reached were treated as not administered. Comparisons of model fit between the models test the assumption that differences in item difficulties between assessments are negligible and that joint scaling can therefore be conducted across assessment periods. The difference in the deviance (-2 x log-likelihood) of the two models was tested for significance using a χ^2-test. Furthermore, Akaike's Information Criterion (AIC; Akaike, 1973), the Bayesian Information Criterion (BIC; Schwarz, 1978), and the consistent Akaike's Information Criterion (CAIC; Bozdogan, 1987) were consulted for finding the better-fitting model. Note that standard errors reported for estimates on group differences and estimates of the interaction between study or instrument and the item do not take into account the link error and neither the sampling error but only represent the statistical uncertainty due to parameter estimation.

Furthermore, the link error (see introduction) was investigated. The link error was computed for the reading link items, the common reading items, and the science link items for each of the different combinations between study and instrument. The link errors for the common reading items and the science link items were compared to the ones reported in the PISA 2009 Technical Report (OECD, 2012). Additionally, the link error 2000/2009 was computed separately by cluster position for the common reading items to investigate whether there were differences in the size of the link error depending on the items position in the test booklet. As the common reading items were only at positions one to three in PISA 2000 (see Table 2), a separate link error for cluster position four could not be computed. This analysis was conducted using data from the 28 OECD countries that had taken part in PISA 2000 and PISA 2009[5]. For each of the cluster positions, a random sample of about 500 students per country was drawn. Selection probabilities in the random sample should be equal to those in the complete sample. To achieve this, we multiplied the final student weights (which reflect the variation in selection probabilities) with random numbers from a uniform distribution to draw the random sample. For the computation of the link error across all cluster positions, both a random sample of about 500 students per country as well as a random sample of about 2000 students per country were drawn.

We analyzed whether differences in the position of items might be explanative for differences in the difficulty between studies and instruments. The PISA test design (from PISA 2003 on) has been balanced regarding the item clusters; that is, each cluster as a whole appears at each of the four cluster positions in one of the test booklets. However, the position of each item unit within its respective cluster is fixed. Thus, the test design is not balanced regarding the position of the item units within clusters. Between assessments the allocation of item units to clusters can change for example, due to differing

[5] Public use data from the PISA assessments is available online at http://www.oecd.org/pages/0,3417,en_32252351_32235731_1_1_1_1_1,00.html

amounts of testing time. In consequence, it is possible for an item to differ in its position in the cluster between two PISA assessments. For example, item R055Q01 was at position 9 in cluster R2 in PISA 2009 while it was at position 3 in cluster R5 in PISA 2000. This means that item difficulties are by design always confounded with the position of items in clusters and the positions are not perfectly controlled for due to constraints in test assembly.

To test directly using a model-based approach whether there is an interaction between item position and instrument would be an interesting prospect. However, this would require a balanced design with regard to the item position which provides data for each possible combination of an item with a position in the instrument. Since this is not the case in the design of the presented national add-on study, this model cannot be estimated. For an application of this LLTM to a large-scale assessment where the item position is balanced see Hohensinn et al. (2008). To approximate this model we instead extended Model 1 (Equation 4) to include an interaction between item, cluster, and instrument. To test whether there was a meaningful interaction between these three components, we compared Model 1 to a model including this three-way interaction[6] where the cluster position is *c = 1, ..., C:*

Model 3: $$\text{logit} = \theta_v - \delta_{igc} + \beta_g. \qquad (6)$$

From PISA 2003 on the test design has been balanced regarding the cluster positions. However, in PISA 2000 this was not yet the case so the model comparison concerning instrument 2000 and instrument 2009 had to be conducted with the set of items that were positioned at all cluster positions in instrument 2000 (17 of the common reading items).

To further test whether position effects may have been responsible for differences in item difficulty, correlations were computed. An index for the cluster position was created which takes into account the number of items in each respective cluster, the position of the item within the cluster, and the position of the cluster in the respective booklet. The first value of the index identifies the item's cluster position (1, 2, or 3). The fraction consists of the position of the item within the cluster (counting from 0) divided by the number of items in the cluster: $Index = cluster\ position + \frac{(item\ number - 1)}{N\ items\ in\ cluster}$. For example, item R055Q01 was the ninth of 15 items in reading cluster R2 which was at position 1 in booklet 8. Thus, item R055Q01 received the index 1 (9-1)/15. In booklet 13, cluster R2 was at position 2 and item R055Q01 therefore received the index 2 8/15. Then, these indices were averaged to obtain the mean position of the items (see Appendix 1). The differences in this position index between instruments were correlated with the differences in item difficulty. If the mean position of items differs between instruments, potentially a bias in the item difficulties might be introduced which corresponds to the average position of the items. To quantify this potential bias, differences in item difficulty were regressed on differences in the position index. The potential bias then equals the predicted value in the item difficulty difference for the average difference in position.

[6] This corresponds to item*cluster*instrument in the ConQuest model statement

Table 2:
PISA 2000 Booklet Design

Booklet ID	Cluster			
	1	**2**	**3**	**4**
1	R1	R2	R4	M1 M2
2	R2	R3	R5	S1 S2
3	R3	R4	R6	M3 M4
4	R4	R5	R7	S3 S4
5	R5	R6	R1	M2 M3
6	R6	R7	R2	S2 S3
7	R7	R1	R3	R8
8	M4 M2	S1 S3	R8	R9
9	S4 S2	M1 M3	R9	R8

Note. R = reading, M = mathematics, S = science.

Table 3:
Study 2009 Booklet Design

	Booklet ID	Cluster			
		1	**2**	**3**	**4**
PISA 2009 booklets	1	M1	R1	R3A	M3
	2	R1	S1	R4A	R7
	3	S1	R3A	M2	S3
	4	R3A	R4A	S2	R2
	5	R4A	M2	R5	M1
	6	R5	R6	R7	R3A
	7	R6	M3	S3	R4A
	8	R2	M1	S1	R6
	9	M2	S2	R6	R1
	10	S2	R5	M3	S1
	11	M3	R7	R2	M2
	12	R7	S3	M1	S2
	13	S3	R2	R1	R5
Reading 2000, Science 2006	14	R1	R2	R4	S1-MS06
	15	R2	R3	R5	S4-MS06
	16	R3	R4	R6	S5-MS06
	17	R4	R5	R7	S6-MS06
	18	R5	R6	R1	S7-MS06

Note. R = reading, M = mathematics, S = science, MS = main study. Booklets 14 to 18 contain reading clusters from PISA 2000 and science clusters from PISA 2006. For some clusters there were two versions, a regular one (A) and an easier one (B).

Results

In the following, results will be reported for the analyses on measurement invariance, the link error, and position effects. The first section contains the results for reading and the second section contains the results for science.

Reading

Two of the reading items, R219Q01T and R219Q01E, were removed at the international level due to data entry errors as described in the PISA 2009 technical report (OECD, 2012). Thus, the reading link consisted of 26 items and there were 37 common reading items between reading 2000 and reading 2009. The group parameters included in Model 1 show for which subsample the items, taken as a whole, were easier. For study 2009, participants filling out reading 2009 were slightly better compared to participants filling out reading 2000 (-0.02 logits, SE = 0.03) regarding the 37 common items. Concerning the 26 link items, the difference in the group parameter was -0.08 logits (SE = 0.03) for study 2009, again favoring participants tested with the PISA 2009 instrument. For the PISA 2000 instrument, the 37 reading items were easier for participants tested in 2000 compared to participants tested in 2009 (-0.32 logits, SE = 0.03). The 26 reading link items were -0.33 logits (SE = 0.03) easier for students assessed with the PISA 2000 instrument in 2000 compared to students assessed with the same instrument in 2009. Bischof, Hochweber, Hartig, and Klieme (in press) did not find significant differences in the mean reading performance between 2000 and 2009 for samples from the same 59 schools used in our study.

Comparisons of model fit showed that for both item sets and for both research questions, Model 1 had lower BIC and CAIC values compared to Model 2 (see Table 4a and Table 4b). However, both the significant χ^2-tests of the difference in deviance between the models and the AIC indicated that the more complex Model 2 fit better than Model 1. Thus, meaningful differences in item difficulty appear to exist. A closer investigation of the item difficulties revealed substantial differences for some items. Figure 1a shows the differences in item difficulty for study 2009 between the PISA 2000 instrument and the PISA 2009 instrument. Positive values indicate that the item was more difficult in the PISA 2009 instrument. Most reading items fall in the negligible or slight to moderate category of the ETS classification system (Zieky, 1993), but some clearly exceed the limit for moderate DIF, most notably R055Q03 which was extremely easy for students filling out reading 2009 compared to students filling out reading 2000 (-1.71 logits, SE = 0.09, 95% CI [-1.88, -1.53]). In Figure 1b, the differences in item difficulty for the reading items in the PISA 2000 instrument between study 2000 and study 2009 are depicted. Here, some of the same items as in Figure 1a showed large differences (e.g., R220Q05 with 1.14 logits, SE = 0.21, 95% CI [0.73, 1.55]), though others showing large differences in Figure 1a only showed smaller differences in Figure 1b (e.g., R055Q03 with -0.66 logits, SE = 0.08, 95% CI [-0.82, -0.51]). The pattern of the item difficulty differences is very similar between the common and link items. However, for some items (e.g., R220Q05) the difference is marginally larger when all 37 common items (1.14 logits, SE

= 0.21, 95% CI [0.73, 1.55]) are included compared to only the 26 link items (1.06 logits, SE = 0.21, 95% CI [0.65, 1.46]) in the comparison of study 2000 and study 2009 for the PISA 2000 instrument. For other items (e.g., R219Q02), the difference is slightly smaller with 37 items (-0.45 logits, SE = 0.14, 95% CI [-0.73, -0.17]) compared to 26 items (-0.52 logits, SE = 0.14, 95% CI [-0.80, -0.24]).

Model 1 was recomputed after removal of the items with the largest differences in item difficulty, namely R055Q03 and R220Q05 for the reading link items and additionally R101Q02 for the common reading items. First, group differences were re-assessed for the comparison of the PISA 2000 and PISA 2009 instruments in study 2009. The remaining 24 reading link items did not differ in difficulty between participants tested with reading 2000 in study 2009 compared to participants tested with reading 2009 in the same study (0.00 logits, SE = 0.03). The reduced number of 34 common items yielded a group parameter of 0.04 logits (SE = 0.03) for study 2009 with participants filling out reading 2000 having slightly better results. The group difference for the common items changed its direction and was slightly larger compared to the full item set.

Second, group differences were re-assessed for the comparison of reading 2000 in study 2000 and study 2009. Regarding the remaining 24 link items, the group parameter amounted to -0.37 logits (SE = 0.03), indicating that the PISA 2000 instrument was easier for students in study 2000 compared to students in study 2009. For the 34 common items the group parameter was -0.35 logits (SE = 0.03), indicating that it too was easier for students assessed in study 2000 compared to students assessed in study 2009. Thus, concerning reading 2000 in study 2000 and study 2009, the differences are in the same direction and slightly larger compared to the full item set for both the link items and the common items.

Link errors were computed between the instruments from PISA 2000 and PISA 2009 for the reading link and common items. These are reported in Table 5. For example, for the 37 common reading items, the link error was 6.43 points on the PISA reading scale for the link between the instruments from PISA 2000 and PISA 2009 both applied in study 2009. When only the 34 common reading items without large differences in item difficulty were used, the link error decreased to 5.48 points. The OECD reports a link error of 4.94 on the PISA reading scale between 2000 and 2009 (Table 12.36; OECD, 2011) which is lower than the link errors computed with our data. For comparisons of the magnitude of the link error in relation to the cluster position of the items in the booklets, the link error was computed separately by cluster position for the 37 common reading items using data from an international sample with $N =$ about 500 per OECD country. When the common reading items were at cluster position 1 in the test booklet, a link error of 6.69 points on the PISA reading scale resulted between PISA 2000 and PISA 2009 (see Table 5). Across the three cluster positions, the international sample with about 500 students per OECD country yielded a link error of 5.86 points while the international sample with about 2000 students per OECD country yielded a link error of 5.92 points on the PISA reading scale.

To investigate one possible reason for the differences in item difficulty we found, position effects were estimated. First, we compared the model fit between Model 1 and Mod-

el 3 (including the three-way interaction between item, cluster, and instrument) for the 17 reading items that appeared at all three cluster positions. Model 1 yielded an AIC of 12060.84 (BIC = 12194.63, CAIC = 12218.63) while Model 3 yielded an AIC of 12088.89 (BIC = 12412.22, CAIC = 12470.22). Thus, the simpler model showed a better fit indicating that overall differences in item position did not play an important role for differences in item difficulty between the two instruments.

Second, we investigated correlations between differences in cluster position and differences in item difficulty. The correlation between the difference in cluster position (reading 2009 – reading 2000) and the difference in item difficulty between the two instruments was $r = .29$ ($p = .08$; $N = 37$) for the 37 common reading items. When the three items with large item difficulty differences were not included, the correlation rose to $r = .41$($p = .02$). The mean difference in cluster position was -0.19 ($SD = 0.55$) which corresponds to about one fifth of a cluster's length. Thus, common reading items on average were at a slightly earlier position (about three to four items earlier) in the PISA 2009 instrument. The resulting potential bias (quantified as the predicted value for the average difference in item position in the regression of difference in item difficulty on difference in item position) for the 34 remaining common items amounted to -0.05 logits (CI [-0.11, 0.01]) in favor of students tested with the PISA 2009 instrument. For the 26 link items the correlation between the difference in cluster position and the difference in item difficulty was $r = .26$ ($p = .20$). For the reduced item set of 24 reading link items, this correlation increased to $r = .55$ ($p = .01$). The 24 link items yielded a mean difference in cluster position of -0.29 ($SD = 0.44$) and a potential bias of -0.13 logits (CI [-0.21, -0.05]). Thus, we would expect participants taking reading 2009 to be slightly better compared to participants taking reading 2000 solely based on the earlier position of the reading items for both reading item sets, though the bias is larger when only taking the link items into account. However, as noted above, in our data students taking reading 2009 were only better than students taking reading 2000 for all 37 common item and the 26 link items, but not for the reduced set of 34 common items. Note that confidence intervals were computed taking into account only the regression's standard error. Considering the measurement error, the sampling error, and the link error additionally would result in wider confidence intervals for the potential bias.

Table 4a:
Comparison of Model Fit for the PISA 2000 (2006) Instrument and the PISA 2009 Instrument for Reading and Science

Domain	Model	N	#par	-2 lnL	AIC	BIC	CAIC	χ^2	df	p
Reading (37 common items)	1 PCM + instrument	1948	45	24355.02	24445.02	24695.88	24740.88			
	2 PCM + instrument + instrument*item	1948	81	24171.27	24333.27	24784.80	24865.8	183.76	36	<.001
Reading (26 link items)	1 PCM + instrument	1948	34	18645.13	18713.13	18902.67	18936.67			
	2 PCM + instrument + instrument*item	1948	59	18483.56	18601.56	18930.46	18989.46	161.57	25	<.001
Science	1 PCM + instrument	1948	47	11337.18	11431.18	11693.19	11740.19			
	2 PCM + instrument + instrument*item	1948	90	11176.83	11356.83	11858.54	11948.54	160.36	43	<.001

Note. PCM = partial credit model, #par = number of parameters, L = Likelihood, BIC = Bayesian Information Criterion, CAIC = consistent Akaike's Information Criterion, χ^2 = -2lnL model 1 – (-2lnL model 2), df = #par model 1 - #par model 2.

Table 4b:
Comparison of Model Fit for Study 2000 and Study 2009

Domain	Model	N	#par	-2 lnL	AIC	BIC	CAIC	χ^2	df	p
Reading (37 common items)	1 PCM + study	2041	45	24262.32	24352.32	24605.28	24650.28			
	2 PCM + study + study*item	2041	81	24167.96	24329.96	24785.28	24866.28	94.36	36	<.001
Reading (26 link items)	1 PCM + study	2041	34	18928.30	18996.30	19187.42	19221.42			
	2 PCM + study + study*item	2041	59	18852.76	18970.76	19302.41	19361.41	75.54	25	<.001

Note. PCM = partial credit model, #par = number of parameters, L = Likelihood, BIC = Bayesian Information Criterion, CAIC = consistent Akaike's Information Criterion, χ^2 = -2lnL model 1 – (-2lnL model 2), df = #par model 1 - #par model 2.

Table 5:
Link errors for Reading and Science

Domain and sample	**Number of items**			
Reading	**37 common items**	**34 reduced common items**	**26 link items**	**24 reduced link items**
Reading 2000 and 2009 in study 2009	6.43	5.48	6.33	6.30
Reading 2000 in study 2000 and study 2009	6.30	4.34	8.02	5.66
International sample				
Cluster position 1 (N = 500 per country)	6.69			
Cluster position 2 (N = 500 per country)	8.13			
Cluster position 3 (N = 500 per country)	8.13			
Cluster positions 1-3 (N = 2000 per country)	5.92			
Cluster positions 1-3 (N = 500 per country)	5.86			

Science	**44 link items**	**42 reduced link items**
Science 2006 and Science 2009 in study 2009	10.50	7.55

Note. Link errors for reading are reported on the PISA 2000 scale. Link errors for science are reported on the PISA 2006 scale.

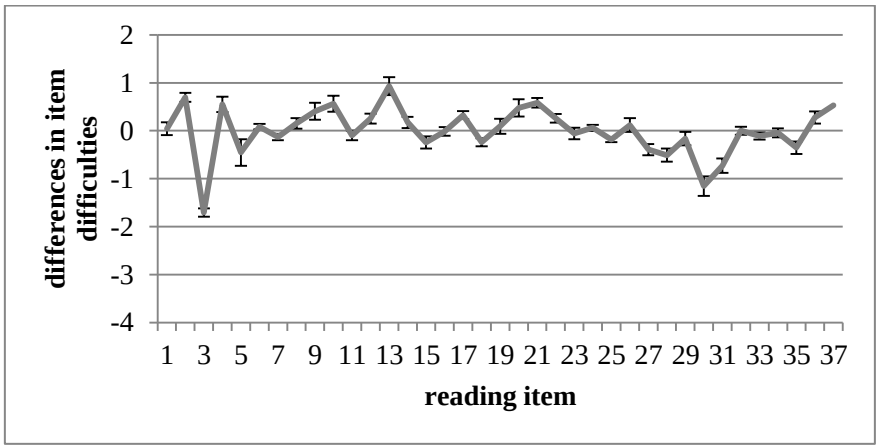

Figure 1a:
Differences in item difficulties for reading between the PISA 2009 instrument and the PISA 2000 instrument in study 2009. Positive values indicate that the item was more difficult in the PISA 2009 instrument. Error bars represent standard errors.

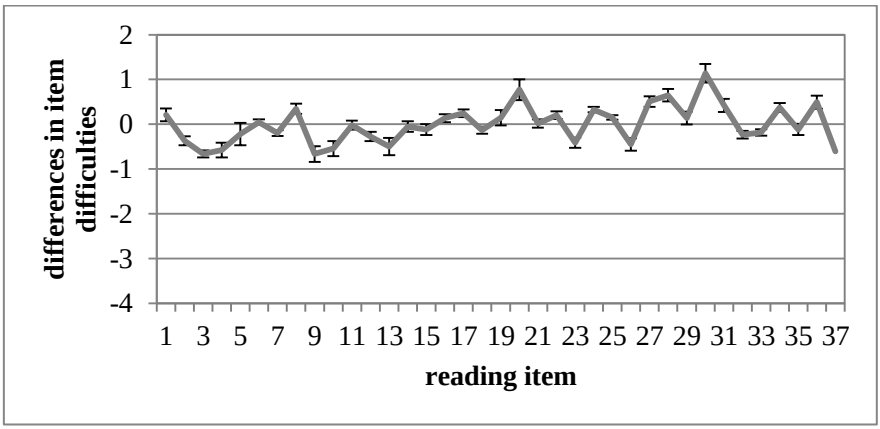

Figure 1b:
Differences in item difficulties between study 2009 and study 2000 for the PISA 2000 instrument. Positive values indicate that the item was more difficult in study 2009. Error bars represent standard errors.

Due to a model identification constraint on the item parameters, no SE are estimated for the last item. Item labels are listed in the Appendix.

Science

As only five of the seven PISA 2006 science clusters were used in study 2009, only 44 science link items (out of the full set of 53 items) could be analyzed. Since these five clusters from PISA 2006 were positioned at the fourth cluster position in study 2009, only this cluster position was used for the data from science 2009 as well. The group parameter included in Model 1 revealed that the science link items were easier for participants tested in 2009 with science 2009 than for participants tested in 2009 with science 2006 (-0.27 logits, SE = 0.04). A comparison of the model fit for Model 1 and Model 2 yielded a better fit for Model 2 according to the χ^2-test and the AIC (see Table 4a). Thus, as for reading, the science items also showed an interaction between instruments (2006 vs. 2009) and items, indicating that differences in item difficulty between the two instruments need to be taken into account. As can be seen in Figure 2, some items differed substantially between science 2006 and science 2009, especially S413Q05 (-3.36 logits, SE = 0.20, 95% CI[-3.75, -2.96]) and S256Q01 (-1.86 logits, SE = 0.39, 95% CI [-2.61, -1.10]) which are both easier in science 2009. When these two items were removed and Model 1 was recomputed, the difference between the two subsamples was reduced to -0.20 logits (SE = 0.04), again favoring students tested with science 2009 in study 2009, though slightly smaller in size compared to the full item set.

The link error for the 42 science link items (without S413Q05 and S256Q01) was 7.55 points on the PISA science scale (see Table 5). This link error is not comparable to the one reported in the PISA 2009 Technical Report (2.57 points; OECD, 2012) which was based on the full set of 53 link items at all four cluster positions. The correlation between the difference in the index for item cluster position and the difference in item difficulty was r = -.03 (p = .84) for the 42 science link items remaining after removal of the two items with the largest differences in item difficulty. The mean difference in item cluster

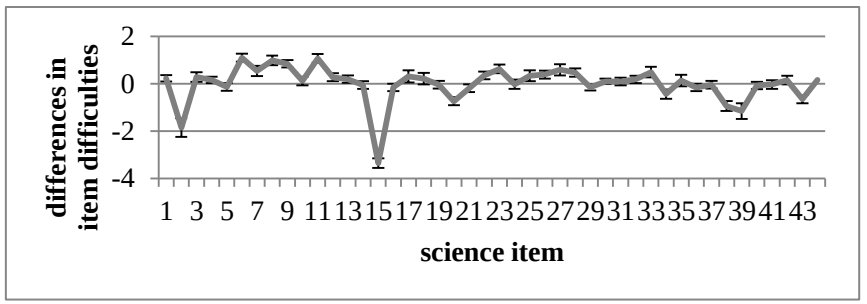

Figure 2:
Differences in item difficulties for science between the PISA 2009 instrument and the PISA 2006 instrument. Positive values indicate that the item was more difficult in the PISA 2009 instrument. Error bars represent standard errors.

Due to a model identification constraint on the item parameters, no SE are estimated for the last item. Item labels are listed in the Appendix.

position between science 2009 and science 2006 (for the fourth cluster) was negligible at -0.01 (*SD* = 0.18). Thus, for the 42 science link items and the fourth cluster position, position effects did not appear to play a role for the differences in item difficulty.

Discussion

In this paper, the measurement invariance – as an important pre-requisite of trend analyses – of PISA reading and science link items was analyzed for items from PISA 2000 and PISA 2009 for reading and from PISA 2006 and PISA 2009 for science. Furthermore, we analyzed whether position effects accounted for differences in item difficulties across instruments and assessments.

Our analyses showed that some of the reading and science link items changed in their difficulty between 2000 and 2009. One possible reason are variations in item wording between the assessments. Regarding the German test booklets applied here, five reading link items (R055Q03, R067Q04, R104Q02, R220Q04, and R227Q02) were phrased slightly differently in PISA 2009 compared to PISA 2000. For R055Q03 this explanation appears especially plausible, since the wording in the German booklets was simplified which may have led to the item being easier in the PISA 2009 instrument compared to the PISA 2000 instrument. As an aside, R055Q03 was deleted at the national level in the German-speaking countries for PISA 2000 and PISA 2003 but has been retained since PISA 2006 (presumably after the wording was changed).

The comparison of model fit for the RCML models making different equality assumptions confirms that differences in item difficulties exist. Thus, the PISA 2000 and PISA 2009 instruments are not measurement invariant regarding the 37 common reading items as well as the 26 reading link items. Furthermore, the instrument from PISA 2000 also was not measurement invariant between two studies (study 2000 and study 2009) regarding the reading items. For 44 of the science link items, measurement invariance was also shown to be violated between the instruments from PISA 2006 and PISA 2009.

The link errors computed with our data were larger compared to the ones reported in the Technical Report for PISA 2009 (OECD, 2012) for the reading link 00/09 and the science link 06/09. However, when items with large differences in item difficulty between the instruments were removed, the link error was reduced by approximately 0 to 3 points on the PISA scale (mean reduction 20.83%). Thus, the few items that changed their difficulties between assessments appear to have had a strong influence on the size of the link error. Furthermore, using the reduced item sets, link errors were larger for the 24 reading link items compared to the 34 common reading items for the link 00/09 in study 2009. For science, the link error computed from items on cluster position four was much larger than the one computed by the OECD for all cluster positions. For the international sample the link error was smallest at cluster position 1 and largest at cluster position 3. It is conceivable that the link error was increased by fatigue effects for cluster positions 2 and 3. Differences between the link errors from the international samples drawn in this study and the one reported by the OECD are probably due to the different data used: the OECD link error is based on data from all four cluster positions while our link errors are

based on data from only the first three cluster positions for reading. Since the link error is computed using the differences in item difficulty between assessments, it can be assumed that differing results on the differences in item difficulty between the OECD sample and our sample contributed to differing link errors. The size of the link error also appears to be influenced by sample size since the link error was slightly larger for an international sample of about 500 students per OECD country compared to an international sample of about 2000 students per OECD country. Large link errors can impair the measurement invariance of PISA instruments and in consequence limit the conclusions that can be drawn from trend analyses. It follows that eliminating factors that lead to large link errors is important. Our results confirm the previous finding by Wu (2010) and Mazzeo and von Davier (2009) that rather more than fewer items should be used to establish the link.

Differences in item position between instruments are a possible explanation for differences in item difficulty. Position effects are generally of concern in the ability domain, where proficiency scores may be biased if the position on which items are presented has an influence on item difficulties over and above the items' content. This is illustrated by Hohensinn et al. (2008) in an application of the LLTM to test item position effects in mathematics data in a large scale assessment. Hohensinn et al. showed a small fatigue effect taking place. Other frameworks than the LLTM as in this study and Hohensinn et al. can also be applied to the investigation of item position effects. For example, Schweizer and Ren (2013) demonstrate how confirmatory factor analysis can be used to represent the position effect in speed tests where individual differences in working speed also play a role. Our study showed that in the PISA 2009 instrument, the reading link items were on average positioned earlier compared to the instrument used in PISA 2000. Thus, these items may have been easier for participants in PISA 2009 due to position effects. The model including an interaction between item, cluster, and instrument did not show a better fit than the simpler model not including this interaction for the subset of items that allowed estimating this model. This indicates that overall differences in item position had a negligible effect on differences in item difficulty. Nevertheless, position effects may have played a role for some items. For example, the difference in cluster position and the difference in item difficulties between the instruments from PISA 2000 and PISA 2009 for the reading items showed a small to medium correlation, indicating that on average, reading items were positioned earlier and were easier in reading 2009 compared to reading 2000. It follows that the recommendation expressed by Mazzeo and von Davier (2009) as well as Wu (2010) of changing as little as possible and assuming that all changes have an effect can only be emphasized as even minor differences between assessments can limit possibilities for trend analyses.

Strongly related to the issue of item position effects is the issue of booklet effects which can also influence changes in item difficulty and in turn enlarge the link error. Booklet effects refer to the position of items in test booklets. According to Wu (2010), link items should be placed at the same position since difficulty changes resulting from position effects may increase the link error. Booklet effects affected item parameter estimates in PISA 2000 (Adams & Carstensen, 2002). Since the test design has been balanced from PISA 2003 on, item parameter estimates in PISA 2006 and PISA 2009 should not be

affected by booklet effects. However, the different location of domains within each booklet had an effect on proficiency distributions (OECD, 2012). This is taken into account in PISA 2009 by estimating booklet parameter estimates and adding or subtracting the resulting booklet effects from the proficiencies of students. Lastly, carry-over effects may also contribute to difference in item difficulties. This is especially relevant for the comparison between science 2006 and science 2009 as well as reading 2000 and reading 2009 both assessed in study 2009 since here differing items preceded the link and common items we analyzed, possibly contributing to differences in item difficulties, while for the comparison between study 2000 and study 2009 regarding reading the composition of the clusters was identical.

Limitations

The results reported here are based on samples from a single country, namely Germany. Results at the international level or in other countries participating in PISA may differ. The samples used in this study both consisted of high school students. Thus, the results are not generally valid for other school types. Furthermore, while the sample assessed in 2000 was part of the official German PISA 2000 sample, the sample assessed in 2009 formed part of a study conducted by the German PISA consortium in addition to PISA 2009. However, since this study was conducted in adherence to the PISA procedure (e.g., concerning standardization), it can be assumed that the data collection and analyses for this sample did not differ systematically from those of the PISA sample.

Conclusion

The interaction between items and study or instrument, respectively, indicates that measurement invariance between the PISA instruments for 2000 (2006) and 2009 for reading (science) is not given. For some items, differences in item difficulty are substantial. These may partly be attributed to position effects, though other factors play a role as well. For the reading items, the link 2000/2009 works quite well with all common items and shows a smaller link error compared to the link error computed with only the link items. Items with large differences in item difficulty between assessments appear to increase the link error and thus should be removed from linking.

References

Adams, R. J., & Carstensen, C. H. (2002). Scaling outcomes. In OECD, *PISA 2000 Technical Report* (pp. 149–162). Paris, France: OECD.

Adams, R. J., & Wilson, M. R. (1996). Formulating the Rasch model as a mixed coefficients multinomial logit. In G. Engelhard & M. R. Wilson (Eds.), *Objective measurement: Theory into practice. Vol III* (pp. 143–166). Norwood, NJ: Ablex.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest, Hungary: Academiai Kiado.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581–594. doi:10.1177/014662167800200413

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., … Weiß, M. (Eds.). (2002). *Pisa 2000 – die Länder der Bundesrepublik Deutschland im Vergleich: PISA-E* [PISA 2000 – comparison of the German states: PISA-E]. Opladen, Germany: Leske + Budrich.

Bischof, L., Hochweber, J., Hartig, J., & Klieme, E. (in press). Schulentwicklung im Verlauf eines Jahrzehnts – Erste Ergebnisse des PISA Schulpanels [School development in a decade – First results from the PISA school panel]. *Zeitschrift für Pädagogik*.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. doi:10.1007/BF02293801

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. doi:10.1007/BF02294361

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*, 359–374. doi:10.1016/0001-6918(73)90003-6

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

Hohensinn, C., Kubinger, K. D., Reif, M., Holocher-Ertl, Khorramdel, L., & Frebort, M. (2008). Examining item-position effects in large-scale assessment using the Linear Logistic Test Model. *Psychology Science Quarterly*, *50*(3), 391-402.

Kolen, M. J., & Brennan, R. L. (2004). *Test Equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Linacre, J. M. (1994). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. doi:10.1007/BF02296272

Mazzeo, J. & Davier, M. von. (2009). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*. Retrieved August, 2010, from http://edsurveys.rti.org/PISA.

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

OECD. (2002). *PISA 2000 Technical report*. Paris, France: OECD Publications.

OECD. (2009a). *PISA 2009 Assessment framework – Key competencies in reading, mathematics, and science*. Paris, France: OECD Publications.

OECD. (2009b). *PISA 2006 Technical report*. Paris, France: OECD Publications.

OECD. (2010). *PISA 2009. Learning trends: Changes in student performance since 2000*. Paris, France: OECD Publications.

OECD. (2012). *PISA 2009 Technical Report*. Retrieved from http://dx.doi.org/10.1787/9789264167872-en

OECD, & Statistics Canada. (2000). *Literacy in the Information Age: Final report of the International Adult Literacy Survey*. Paris, France: OECD Publications.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danmarks Paedagogiske Institut.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. doi:10.1214/aos/1176344136

Schweizer, K., & Ren, X. (2013). The position effect in tests with a time limit: the consideration of interruption and working speed. *Psychological Test and Assessment Modeling, 55*(1), 62-78.

Wu, M. L. (2010). Measurement, sampling, and equating errors in large-scale assessments. *Educational Measurement: Issues and Practice*, *4*(29), 15–27.

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. (2007). ConQuest (Version 2.0) [Computer software]. Camberwell, Australia: Australian Council for Educational Research.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 337–364). Hillsdale, NJ: Lawrence Erlbaum.

Appendix 1

Reading and Science Items with Item Parameters (SE) in each Subsample and Mean Position

Domain	Item nr.	Item label	Item parameter (SE) S00 – I00 *N* = 1487	Item parameter (SE) S09 – I00 *N* = 554	mean position I00	Item parameter (SE) S09 – I09 *N* = 1394	mean position I09
Reading	1	R055Q01*	-3.07 (.21)	-2.56 (.20)	2.13	-2.52 (.17)	2.53
	2	R055Q02*	-1.59 (.12)	-1.66 (.15)	2.20	-0.96 (.11)	2.60
	3	R055Q03*	0.48 (.10)	0.16 (.12)	2.27	-1.54 (.13)	2.67
	4	R055Q05*	-2.96 (.20)	-3.25 (.27)	2.33	-2.70 (.19)	2.73
	5	R067Q01*	-4.12 (.34)	-4.06 (.38)	2.60	-4.51 (.42)	2.21
	6	R067Q04*	-0.86 (.08)	-0.60 (.08)	2.67	-0.46 (.07)	2.29
	7	R067Q05*	-1.37 (.10)	-1.19 (.09)	2.73	-1.29 (.09)	2.36
	8	R083Q01	-2.27 (.15)	-1.63 (.18)	2.74	-1.48 (.12)	2.00
	9	R083Q02	-2.73 (.18)	-3.12 (.30)	2.79	-2.71 (.18)	2.06
	10	R083Q03	-2.76 (.18)	-3.03 (.29)	2.85	-2.47 (.17)	2.13
	11	R083Q04	-1.45 (.12)	-1.17 (.16)	2.91	-1.27 (.12)	2.19
	12	R101Q01	-1.46 (.12)	-1.43 (.17)	2.41	-1.18 (.12)	2.69
	13	R101Q02	-3.10 (.21)	-3.32 (.33)	2.47	-2.39 (.16)	2.75
	14	R101Q03	-2.12 (.15)	-1.88 (.19)	2.53	-1.70 (.13)	2.81
	15	R101Q04	-2.20 (.15)	-2.03 (.20)	2.59	-2.27 (.16)	2.88
	16	R101Q05	-0.74 (.11)	-0.29 (.14)	2.65	-0.30 (.10)	2.94
	17	R102Q04A*	-1.40 (.12)	-0.85 (.12)	2.50	-0.52 (.11)	2.43
	18	R102Q05*	-0.42 (.10)	-0.23 (.12)	2.56	-0.46 (.10)	2.50
	19	R102Q07*	-3.46 (.25)	-3.03 (.24)	2.67	-2.93 (.21)	2.57
	20	R104Q01*	-4.26 (.36)	-3.21 (.31)	3.26	-2.73 (.19)	2.80
	21	R104Q02*	0.55 (.10)	0.91 (.16)	3.32	1.50 (.13)	2.87
	22	R104Q05*	0.63 (.11)	1.39 (.24)	3.44	1.69 (.20)	2.93
	23	R111Q01*	-2.08 (.14)	-2.20 (.17)	2.72	-2.26 (.16)	2.27
	24	R111Q02B*	-0.57 (.08)	0.02 (.09)	2.78	0.09 (.08)	2.33
	25	R111Q06B*	-0.43 (.06)	0.04 (.06)	2.94	-0.14 (.06)	2.47
	26	R219Q02*	-2.39 (.16)	-2.55 (.24)	2.11	-2.43 (.17)	2.14
	27	R220Q01*	-1.04 (.11)	-0.21 (.21)	3.69	-0.60 (.11)	2.64

Domain	Item nr.	Item label	Item parameter (SE) S00 – I00 *N* = 1487	Item parameter (SE) S09 – I00 *N* = 554	mean position I00	Item parameter (SE) S09 – I09 *N* = 1394	mean position I09
Reading	28	R220Q02B*	-2.14 (.15)	-1.19 (.24)	3.75	-1.69 (.13)	2.71
	29	R220Q04*	-1.87 (.13)	-1.42 (.25)	3.81	-1.58 (.13)	2.79
	30	R220Q05*	-3.73 (.27)	-2.29 (.33)	3.88	-3.45 (.26)	2.86
	31	R220Q06*	-2.21 (.15)	-1.49 (.26)	3.94	-2.21 (.16)	2.93
	32	R227Q01*	-1.00 (.11)	-0.92 (.12)	2.00	-0.92 (.11)	2.00
	33	R227Q02T*	-1.85 (.15)	-1.79 (.17)	2.06	-1.92 (.16)	2.07
	34	R227Q03*	-2.01 (.14)	-1.34 (.14)	2.11	-1.38 (.12)	2.13
	35	R227Q06*	-2.54 (.17)	-2.36 (.19)	2.22	-2.72 (.19)	2.20
	36	R245Q01	-3.01 (.20)	-2.23 (.21)	1.50	-1.95 (.14)	2.56
	37	R245Q02	-3.10 (.21)	-3.42 (.35)	1.58	-2.89 (.20)	2.63
Science	1	S131Q02T	NA	-1.04 (.30)	4.05	-1.10 (.11)	4.18
	2	S256Q01	NA	-2.18 (.36)	4.00	-4.30 (.72)	4.22
	3	S269Q01	NA	-1.94 (.34)	4.18	-1.90 (.29)	4.00
	4	S269Q03T	NA	-0.61 (.29)	4.23	-0.74 (.10)	4.06
	5	S269Q04T	NA	0.66 (.29)	4.27	0.18 (.22)	4.11
	6	S326Q01	NA	-1.45 (.31)	4.10	-0.65 (.22)	4.00
	7	S326Q02	NA	-2.41 (.37)	4.15	-2.14 (.30)	4.06
	8	S326Q03	NA	-2.32 (.36)	4.20	-1.61 (.26)	4.11
	9	S326Q04T	NA	0.09 (.28)	4.25	0.59 (.22)	4.17
	10	S408Q01	NA	-1.28 (.31)	4.30	-1.45 (.26)	4.17
	11	S408Q03	NA	1.03 (.29)	4.35	1.73 (.28)	4.22
	12	S408Q04T	NA	-1.07 (.30)	4.40	-1.09 (.24)	4.28
	13	S408Q05	NA	-0.21 (.28)	4.45	-0.33 (.22)	4.33
	14	S413Q04T	NA	-0.45 (.29)	4.84	-0.82 (.22)	4.50
	15	S413Q05	NA	2.10 (.35)	4.89	-1.59 (.26)	4.56
	16	S413Q06	NA	-0.09 (.29)	4.79	-0.56 (.22)	4.44
	17	S415Q02	NA	-2.74 (.40)	4.90	-2.69 (.38)	4.88
	18	S415Q07T	NA	-2.52 (.38)	4.85	-2.56 (.37)	4.82
	19	S415Q08T	NA	-0.34 (.28)	4.95	-0.69 (.23)	4.94
	20	S425Q02	NA	0.02 (.28)	4.48	-1.04 (.24)	4.89
	21	S425Q03	NA	-0.20 (.29)	4.38	-0.71 (.22)	4.78

Domain	Item nr.	Item label	Item parameter (SE) S00 – I00 *N* = 1487	Item parameter (SE) S09 – I00 *N* = 554	mean position I00	Item parameter (SE) S09 – I09 *N* = 1394	mean position I09
Science	22	S425Q04	NA	-0.61 (.29)	4.52	-0.58 (.23)	4.94
	23	S425Q05	NA	-1.57 (.32)	4.43	-1.23 (.24)	4.83
	24	S428Q01	NA	-1.66 (.33)	4.32	-1.95 (.29)	4.29
	25	S428Q03	NA	-2.40 (.38)	4.37	-2.31 (.32)	4.35
	26	S428Q05	NA	-1.16 (.31)	4.42	-1.07 (.24)	4.41
	27	S438Q01T	NA	-2.61 (.40)	4.53	-2.27 (.32)	4.65
	28	S438Q02	NA	-1.16 (.31)	4.58	-0.98 (.24)	4.71
	29	S438Q03T	NA	0.01 (.29)	4.63	-0.46 (.10)	4.76
	30	S465Q01	NA	-0.39 (.24)	4.16	-0.62 (.15)	4.00
	31	S465Q02	NA	-0.64 (.29)	4.21	-0.86 (.23)	4.06
	32	S465Q04	NA	0.14 (.29)	4.26	-0.01 (.21)	4.12
	33	S466Q01T	NA	-2.20 (.36)	4.79	-1.96 (.31)	4.83
	34	S466Q05	NA	-1.16 (.31)	4.89	-1.87 (.30)	4.94
	35	S466Q07T	NA	-2.40 (.38)	4.84	-2.50 (.37)	4.89
	36	S478Q01	NA	0.23 (.29)	4.37	-0.26 (.21)	4.28
	37	S478Q02T	NA	-0.64 (.29)	4.42	-0.99 (.23)	4.33
	38	S478Q03T	NA	-1.21 (.31)	4.47	-2.43 (.33)	4.39
	39	S514Q02	NA	-2.44 (.38)	4.62	-3.83 (.60)	4.47
	40	S514Q03	NA	-0.20 (.29)	4.67	-0.59 (.22)	4.53
	41	S514Q04	NA	-1.22 (.31)	4.71	-1.53 (.26)	4.59
	42	S527Q01T	NA	1.54 (.31)	4.55	1.31 (.26)	4.67
	43	S527Q03T	NA	-0.48 (.29)	4.59	-1.42 (.27)	4.72
	44	S527Q04T	NA	-0.57 (.29)	4.64	-0.70 (.24)	4.78

Note. S00 = study 2000. S09 = study 2009. I00 = PISA 2000 instrument. I06 = PISA 2006 instrument. I09 = PISA 2009 instrument.

* reading link items