JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

DOI: 10.1111/jedm.12261

Linking via Pseudo-Equivalent Group Design: Methodological Considerations and an Application to the PISA and PIAAC Assessments

Artur Pokropek
Educational Research Institute (IBE)
Francesca Borgonovi
University College London and Organisation for Economic Co-Operation and Development (OECD)

This article presents the pseudo-equivalent group approach and discusses how it can enhance the quality of linking in the presence of nonequivalent groups. The pseudo-equivalent group approach allows to achieve pseudo-equivalence using propensity score reweighting techniques. We use it to perform linking to establish scale concordance between two assessments. The article presents Monte-Carlo simulations and a real data application based on data from the Survey of Adult Skills (PIAAC) and the Programme for International Student Assessment (PISA). Monte-Carlo simulations suggest that the pseudo-equivalent group design is particularly useful whenever there is a large overlap across the two groups with respect to balancing variables and when the correlation between such variables and ability is medium or high. The example based on PISA and PIAAC data indicates that the approach can provide reasonable accurate linking that can be used for group-level comparisons.

Large-scale assessments are key components of accountability systems and have been increasingly used to monitor the performance of teachers, schools and education systems. The resulting growth in the availability of, and interest in large-scale assessments has in turn created awareness of the research and policy opportunities that could be gained from integrating results from different assessments.

At the national level integrating information from different assessments through linking procedures allows to monitor achievement growth when different assessments target different age groups or school grades or different geographical or temporal coverage. At the international level, linking assessments could support efforts to monitor progress towards the achievement of the Sustainable Development Goals in education (United Nations, 2017). Existing international assessments in fact cover different countries and world regions. Therefore, each study can paint only a partial picture of the progress made towards the provision of quality education to all. Linking could also allow to benchmark national results against international goals (see for example, Hanushek & Woessmann, 2013) and create opportunities to study learning growth at the cohort level across countries by combining linked cross-sectional studies conducted at different time points. Finally, linking could shed light on inconsistencies between national and international assessments (Cartwright, Lalancette, Mussio, & Xing, 2003, p. 6; Linn, McLaughlin, & Thissen, 2009; Szaleniec, Grudniewska, Kondratek, Kulon, & Pokropek, 2013) and between different international assessments (Wu, 2009).

However, proper linking can only be applied to very specific situations. The aim of this article is to illustrate how the pseudo-equivalent group approach could be used to link large-scale assessments in the absence of explicit linking designs. We present an analysis framework to estimate the expected performance of individuals in an achievement test given their performance in a different test in the absence of a design that allows direct comparisons across the two. We check the proposed method by running Monte-Carlo simulations under different scenarios. Finally, we apply this framework to compare the performance in two large-scale international assessments: the Programme for International Student Assessment (PISA) and the OECD Survey of Adult Skills (PIAAC).

Related Research

Concordance studies have been carried out in the past using large-scale international assessments. However, most existing concordance studies were conducted following the request of either policy makers responsible for test design, funding and implementation or were conducted directly by testing agencies to estimate the properties of different assessments. Therefore, such studies could implement adjustments during design and administration. The most common example of concordance analyses carried out using single, group, equivalent group or anchor test designs (Feuer, Holland, Green, Bertenthal, & Hemphill, 1999; Linn, 1975) is when large-scale international assessments have been linked to national assessments. An overview of previous studies that aimed to link such assessments is presented in Table 1.

Such linking studies used either single group design or equivalent group design assuming equivalence of groups through different linking procedures. The first type of linking (single group) is relatively rare since in most situations it involves the administration of an additional linking study on selected group of individuals and therefore implies additional costs. Moreover, conducting such a study requires designing proper test linking forms covering the core content from two different assessments, assuring that items will not be compromised, and assuring that test-takers answering linking forms and the original assessments do so in similar conditions. Equivalent group design is more common because it is less costly and easier to conduct, but its validity relies on the assumption of equivalence across groups. However, this assumption might not hold in practice, a problem that is either implicitly or explicitly recognized in most of the studies. In practice, in most instances one cannot rely on the availability of neither single group, equivalent group or anchor test designs and alternative approaches must be used.

In many large-scale international assessments, background questionnaires identify characteristics of respondents that can be used to develop pseudo-equivalent groups. The idea of using additional information to link different test forms has been investigated since the late 1980s. However, most applications have so far been limited to linkages based on common items or common persons and additional background information was used in addition to facilitate linking (Cook, Eignor, & Schmitt, 1988;

Table 1
Studies Establishing Linkages between International and National Assessments

Study	Tests	Method	Conclusions
Beaton and Gonzalez (1993)	U.S. National Assessment of Educational Progress (NAEP) and	Distribution-matching in a single group design and linear linking	The two approaches yielded similar results for average performing countries but were less consistent for low-achieving countries.
Pashley and Phillips (1993)	International Assessment of Educational Progress (IAEP)	Projection technique based on linear regression in a single group design	
Johnson (1998)	NAEP and The Third International Mathematics and Science Study (TIMSS)	Equivalent group design assuming equivalency of groups with various linking procedures	Findings revealed that the three approaches yielded very similar predicted TIMSS results for U.S. states based on their NAEP scores
Johnson, Cohen, Chen, Jiang, and Zhang (2003)			
Phillips (2007)			
Jia et al. (2014)		Equivalent group and common items design; calibration, statistical projection, and statistical moderation	
Lim and Sireci (2017)	TIMSS and NAEP For three time points	Equivalent group design assuming equivalency of groups; equipercentile linking	Problem of not strictly "randomly equivalent" group noted

(*Continued*)

Table 1
Continued

Study	Tests	Method	Conclusions
Cartwright et al. (2003)	British Columbia's Foundation Skills Assessment (FSA) and PISA 2000 reading assessment	Single group design (group of students sat both tests); several methods of linking	Different linking methods gives similar results
Radwan and Xu (2012)	Ontario Secondary School Literacy Test PISA 2009 reading		
ACT (2011)	PISA and PLAN		
Yamamoto (2002)	PISA was linked to the International Adult Literacy Survey (IALS	Common item and single group design using IRT methodology	Results from the linking exercise indicated that PISA students could be reliably placed on the IALS scale
Hambleton, Sireci, and Smith (2009)	NAEP, TIMSS, and PISA	Equivalent group design assuming equivalency of groups; equipercentile linking	Linking allowed to provide rough estimates of how well students from different countries perform Problem of not strictly "randomly equivalent" group noted

Lawrence & Dorans, 1990; Paek, Liu, & Oh, 2006; Yu, Livingston, Larkin, & Bonett, 2004).

The literature on linking without common items and nonequivalent or not fully equivalent groups using statistical matching is small but growing. Haberman (2015) for example proposed a linking method for nonequivalent groups of examinees where the test forms lack common linking items or have unsatisfactory linking items. The procedure used background information concerning examinees to construct sample weightings via minimum discriminant information (Haberman, 1984). Haberman (2015) showed that the pseudo-equivalent groups approach, which uses background questions to link different test forms, produced results similar to equivalent group approaches. Wiberg and Bränberg (2015) showed that covariates could be used as substitutes for common items in a nonequivalent groups with covariates (NEC) design. Sansivieri and Wiberg (2017) proposed new methods for the NEC design with covariates using information from covariates in IRT observed-score linking, while Wallin and Wiberg (2017) used propensity scores based on covariates for kernel equating. Statistical matching was used to link PISA with the Teaching and Learning International Survey (Kaplan & McCarty, 2013).

The overwhelming conclusion of the studies reviewed is that linking does not yield results that can be used to compare individual scores unless assessments are similar in inferences, constructs, populations and measurement characteristics (Feuer et al., 1999). However, even in circumstances in which these conditions are not met, valid comparisons for groups of the population can be made and scales can be compared. The validity and precision of such comparisons depend on the level of similarity (of constructs and testing conditions) between different assessments. In most cases the construct similarity condition was met (Wu, 2009).

General Method

A strategy to achieve pseudo-equivalence when attempting to link two large-scale assessments in the absence of random group design involves the following four steps. In the first step propensity scores (Rosenbaum & Rubin, 1983) are generated. In the second step propensity scores are used to balance the groups and achieve pseudo-equivalence. In the third step linking procedures are used on pseudo-equivalent groups to transform scales. Finally, in the fourth step a linking error is computed using bootstrap procedures (Efron 1982).

Propensity Score Model

Rosenbaum and Rubin (1983) show that in a group of subjects with the same propensity score, the distribution of observed covariates is the same and therefore conditional equivalence is achieved. Propensity scores can be estimated using various techniques, but the most popular approach is through logistic models (which we use), where treatment status (the version of the test) is the dependent variable and observed characteristics constitute the independent variables. Covariates that may need to be controlled for in order to effectively match the distribution of the two populations (sitting two different tests) include age, gender, socio-economic status or study programme (for example, because different assessments may have a different focus in terms of schools/study programme or demographic groups).

Balancing the Groups for Pseudo-Equivalence

After generating propensity scores different approaches can be used to obtain the pseudo-equivalence of groups for linking: propensity score matching, stratification on the propensity score, inverse probability of treatment weighting (IPTW) and covariate adjustment (Austin, 2011). There is no strong argument in favor of the application of a specific approach in the context of pseudo-equivalent group design. However, results from simulation studies suggest that propensity score matching and weighting techniques outperform others (Austin, 2009). While some authors advocate matching (Frölich, 2004) others prefer weighting (McCaffrey, Lockwood, & Setodji, 2013). We opt for weighting since this is computationally more parsimonious. Weighting involves generating predicted probabilities by using inverse probability weighting, with weights being constructed to reflect the average effect of treatment on the treated (ATT) (Austin, 2011). The choice of ATT rather than average treatment effect (ATE) should be the first choice whenever there are considerable differences in the distribution of key covariates in the populations captured in the two assessments and upper bound restrictions may be present. In order to reweight the B sample to match the A distribution for all individuals sitting test A, weights need to be set to 1 and for B such weights should reflect the propensity score.

Transforming Scales Using Reweighted Data

Applying propensity score weighting allows to obtain distribution parameters for the two tests in two pseudo-equivalent groups and to estimate score distributions for tests A and B. Concordance analysis can subsequently be performed by estimating a scale transformation function. In order to transform A scores in the B metric, the following linear transformation might be used (for a detailed discussion see Kolen & Brennan, 2004, p. 31–32):

$$l_B(A) = \frac{\sigma_B}{\sigma_A} A + \left[\mu_B - \frac{\sigma_B}{\sigma_A}\mu_A\right], \tag{1}$$

where A denotes scores from test A and B denotes scores from test B, μ_A and σ_A are the mean and standard deviation of A scores, and μ_B and σ_B denote mean and standard deviation of B scores after reweighting. Both equations express a simple linear function where $\sigma_A/\sigma_B(A)$ is the slope and $\mu_A - \sigma_A/\sigma_B\mu_B$ the intercept. Although other linking methods could be used (e.g., kernel linking), we opted for linear linking because it has the important advantage of simplicity: linear linking in fact results in two numbers (slope and intercept) which can be used easily by researchers interested in A-B concordance. In most situations simplicity does not come at the expense of precision. For example, the two methods yield virtually identical results in the empirical example presented (results can be requested from the authors).

Computation of Linking Error

If the means and standard deviations used to construct the linking function were error-free, the transformed values could be treated as if they were observed values. However, transformation constants should be corrected for uncertainty arising from the use of estimation techniques. Such uncertainty can be introduced by adding a linking error to the estimated standard errors associated with the parameters of interest. Linking errors are used, for example, in PISA whenever trends in achievement for individual countries are conducted. The estimation of linking errors in the context of pseudo-equivalent groups lacks anchor items, test takers are different and there is only a partial overlap in the distribution of test takers along specified observed characterized across the two tests. Linking errors can be estimated by considering the discrepancies between linking parameters obtained using different samples of respondents.

Linking procedures based on pseudo-equivalent group designs are prone to two systematic sources of error, which should be reflected in the uncertainty associated with estimates: error induced by weighting and measurement error. In order to correctly account for uncertainty, it is possible to use bootstrapping to account for weighting-derived uncertainty and plausible values to account for measurement error (Efron, 1982, p. 29–35).

Additional Considerations

Working with large scale assessments like PISA and PIAAC involves using the plausible values (PVs) methodology (Wu, 2005) because test takers are not assigned a single score, but, rather, a set of plausible values representing realizations from random draws of the estimated posterior distribution of individual abilities. For each ability measure and for each participant five PVs were generated in PISA and ten PVs were generated in PIAAC. The analysis presented in this article is based on the five PVs from PISA and the first five PVs from PIAAC (to enable matching with PISA). All analyses involving PVs were performed five times and results were combined using Rubin's rule (Rubin, 1987).

In the next two sections, we test the approach described above using Monte-Carlo simulations. We compare pseudo-equivalent group design to a design that assumes full equivalence. Next we illustrate how the approach works using an empirical example built on PISA and PIAAC data.

Study 1: Monte-Carlo Simulation

A key threat to the validity of the pseudo-equivalent approach is that the overlapping parts of the A group and the B group along covariate *k* may have different ability distributions. We run a Monte-Carlo simulation to compare the pseudo-equivalent approach and the equivalent group approach and consider the impact of different levels of population overlap over a key covariate (in our case age) and different levels of correlations between age and ability. We also present results from a scenario that allows us to bridge the analysis conducted in Study 1 to Study 2, that is, a real data example using PIAAC and PISA data.

Data and Method

For the first scenario data were generated by sampling random variables that mimic age (the main covariate in the two assessments) from two distributions and generating ability scores correlated with age. For the first group, the age distribution had a mean of 15 and a standard deviation of 1. For the second group, we generated 5 scenarios with the following means: 15, 16, 17, 18, or 20. The standard deviation was set to two in each scenario. Additionally, the first age distribution was truncated to have a maximum age of 17 years and 0 months. The age distributions of the 5 scenarios were truncated to have a minimum of 15 and 0 months.

This meant that only the first distribution included simulated data points for individuals younger than 15 and only the second distribution included simulated data points for individuals older than 17. Data points between the age of 15 and 17 constitute the overlapping region. Depending on the scenario considered, the overlap accounted for between 3% and 33% of the total simulated sample, representing between 300 and 3,300 data points. The total size for the Monte-Carlo experiment was set to 10,000 (5,000 for each group and scenario).

An ability variable was generated to have a mean of 500 and a standard deviation of 100 in the pooled sample of individuals from two groups. In order to explore the sensitivity of our findings to an association between ability and age, ability was generated to correlate with age. We evaluated three levels of correlation: strong (.5), medium (.3), and weak (.1). An additional auxiliary variable was generated from a standard normal distribution with a correlation of .35 with the ability variable. The auxiliary variable mimicked in the analysis phase the inclusion of additional covariates typically considered in analyses of assessment data. A correlation of .35 corresponds to an explained variance in achievement of around 12%, which is the lower bound of explanatory power of background variables such as socioeconomic variables.

We put the results of the two assessments on different scales by subtracting 100 from the ability variable in the second group and divided the scale by 2 (no change for group 1).

The two datasets allowed us to test two approaches to establish a link. The first was the pseudo-equivalent approach described in the General Method section. The second approach was linear linking: we used only the overlapping part of the age distributions (between 15 and 17) and treated groups as equivalent in this age range. We repeated the procedure just described 10,000 times for each scenario. The performance of the two approaches was evaluated using standard indicators: bias (in original metric averaged across all replications), and root mean squared error (RMSE).

The second scenario for the simulations was designed to mimic real data used in Study 2 where PISA and PIAAC data are involved. We generated the data in the same way data were generated in the first scenario but we changed conditions, setting a medium level correlation between ability and age (–.25), and a small overlap in age distributions. The size of the shared age group was set to 215 with a sample size of 4,837 for PISA and 9,366 for PIAAC). Similar to the first condition we compare the pseudo-equivalent and the equivalent group approach. In the latter we impose age restrictions that mimic real data (observations with an age smaller than 15.25 and greater than 17 were excluded from the sample mimicking the PISA data, while observations with an age smaller than 16 were excluded from the sample mimicking the PIAAC data).

Performance was evaluated using bias and RMSE supplemented by mean squared error (MSE) empirical SE, % gain in precision (indicating the inverse squared ratio of the empirical standard error of pseudo-equivalent group to that of the equivalent group).

Study 1 Results

Figure 1 illustrates how the two approaches perform in terms of recovering the true mean for group 2 on the initial scale given different levels of correlations between age and ability. On the horizontal axis, we present variations in the age distributions of group 2 (mean 15, 16, 17, 18, and 20), which determines the share of common persons between the two groups.

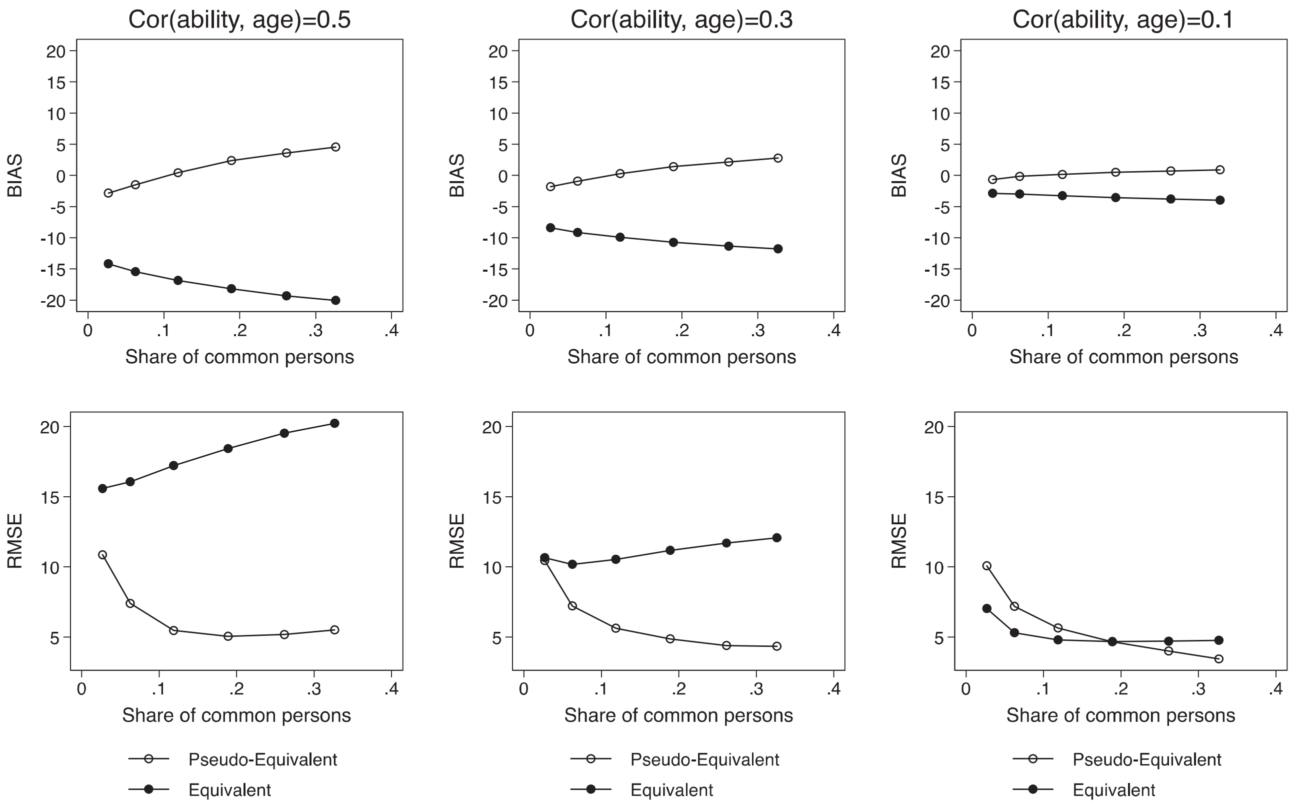

Figure 1. Monte-Carlo simulation: Pseudo-equivalent vs. equivalent group approaches for mean recovery.

Table 2
Results of Monte-Carlo Simulation Study Based on 10,000 Replications

Approach	Bias	Empirical SE	% Gain in Precision	Mean Squared Error	Root Mean Squared Error
Equivalent group	.720	6.785	—	48.390	5.537
Pseudo-equivalent group	.889	6.089	24.175	39.636	5.027

The pseudo-equivalent approach yields a smaller bias than the equivalent group approach. In general, the higher the share of the common persons in the groups, the higher was the bias. This is not intuitive but it is because of differences in the shapes of ability distributions in different scenarios. When the share of common people is larger, ability distributions for the two groups in the common part were more different than when this share is lower. The RMSE for correlations of .5 and .3 is smaller for the pseudo-equivalent than for the equivalent group approach for all scenarios and the advantage of the pseudo-equivalent approach is particularly large when the common part and the distribution is larger and when the correlation between age and achievement is stronger. By contrast, when the correlation between age and achievement is low (.1), the two approaches perform essentially on a par.

In Table 2 we present conditions mimicking the empirical analysis presented in Study 2. Both methods yield small positive biases, with the equivalent-group approach yielding a smaller bias than the pseudo-equivalent-group approach. However, according to the empirical SE, gain in precision, MSE, and RMSE indicators the pseudo-equivalent-group approach is superior and brings higher overall precision. These results confirm that for the presented example, the pseudo-equivalent group approach is preferable to the equivalent group approach, although the difference is not large. Moreover, the absolute values of RMSE suggest that we can achieve reasonable estimates of PIAAC population mean abilities on the PISA scale with the average error of 5 points on a scale that has a standard deviation of 100 points.

In summary, results of Study 1 indicate that the pseudo-equivalent group approach should be used when the correlation between measured abilities and main covariates is high and the overlap between the distributions to be matched provide reasonable sample sizes (preferably higher than 1,000). In other cases, the pseudo-equivalent group approach may yield only small improvements or even be worse than an approach that uses the overlapping part of the age distribution as is done in equivalent group design. Preferably, researchers considering pseudo-equivalent group designs should undertake a simulation study like the one presented in Table 2 to determine whether the conditions they are faced with favour the use of the pseudo-equivalent group approach or not. Table 2 suggests that in the case of PISA and PIAAC, a small improvement can be obtained over approaches that ignore imbalances across groups while achieving reasonable accuracy for scale transformation. These results motivate us to use the pseudo-equivalent group design in Study 2.

Study 2: Real Data Example Based on PISA and PIAAC

PISA is a triennial large-scale low-stakes standardized assessment targeting the schooled population of children between the ages of 15 years and three months and 16 years and two months. The PIAAC target population is defined as "all non-institutionalised adults between the ages of 16 and 65 (inclusive) whose usual place of residence is in the country at the time of data collection." PIAAC is a household-based study while PISA is a school-based study. Both assessments measure reading and mathematical skills (in PIAAC named literacy and numeracy).

The PIAAC and the PISA frameworks are very similar. Both share the same (action-oriented or functional) definition of skills. They share a common approach to the specification of constructs, a comparable definition of measured abilities, and similar content definitions and contexts in which tasks are embedded (for more details see OECD, 2013b, p. 86–91).

While PIAAC and PISA share many features, they differ along important dimensions. They have different target populations and operational procedures. Furthermore, while the main PISA instruments were paper-based in 2012, PIAAC was the first computer-based assessment, although individuals who lacked familiarity with a computer (or a willingness to sit a test with a computer) were offered a paper-based version of the test. PISA and PIAAC also vary in the response formats to test questions. PISA uses a greater variety of response formats than PIAAC.

Data

No attempts were made to link PIAAC and PISA at the international level during the design of the two studies. However, in PISA 2012, countries had the opportunity to extend the PISA target population through national options. In Poland students from grade 10 (16+) were sampled. These students participated in the study following the administration protocols and procedures that were implemented for the population of 15-year-olds. Results for the additional national sample were scaled together with the international sample. An important difference is that sampling of the international sample was based on students' age while the national sample was grade based, capturing not only the age group that typically attends grade 10 (16- and 17-year-old students in Poland) but also older ones. Older students could be attending grade 10 for different reasons, for example, 36% of students aged 18 and 19 (39 students in the sample) had repeated a year or more. Students might also have started school at a late age. It is reasonable to assume that 18- and 19-year-olds in the grade 10 sample are a highly selected group and that background variables are unlikely to adequately reflect the selection process leading such students to be in grade 10. Therefore, we excluded 18- and 19-year-old students participating in PISA from the analysis. Table 3 presents the age distribution of the Polish samples for PISA 2012 and PIAAC.

The age distributions of the two assessments overlap but are highly unbalanced. Age differences between the two groups mean that it is not possible to use equivalent group design to derive concordance between the two assessments. Since populations in both surveys can be defined by age, which is observable, the concordance design

Table 3
Distribution of Age in the Polish Samples of PISA and PIAAC

	Age						
Dataset	<16	16	17	18	19	>19	Total
PISA	3,526*	3,642	1,088	79*	28*	0	8,152
PIAAC	0	93	123	157	702	8,291	9,989
Total	3,526	3,735	1,211	236	730	8,291	18,141

*Excluded from linking.

could be treated as a missing data design, under the assumption of random missing cases, as detailed in Rubin (1974).

Method

Two sets of covariates were used to generate propensity scores. The first was age (recorded to monthly precision) and its square term to account for nonlinear effects. Age was the most important predictor in the propensity score model since the random process assigning participants to each study operated mainly through this variable. Because compulsory schooling for the young cohorts captured in PISA and PIAAC in 2012 lasted until age 18 and selection into upper secondary had already occurred for students in our samples, education selection should not bias our results. Therefore, the probability of being sampled in either study can be considered to be a function of age and, conditional on age, samples should theoretically be equivalent.

Because some discrepancies might occur, other selection-related variables were added to the second model to account for such unintended selection process. Since the response rate in PIAAC is lower than PISA (in Poland in 2012 the response rate was 56% for PIAAC and 88% for PISA) (OECD, 2013a,b), the primary selection mechanism unaccounted for in our study is the one induced by differential response rates. However, nonresponse bias analyses indicate that the Polish PIAAC sample is unbiased with respect to the underlying target population. We nonetheless add additional controls to the model to reduce any potential selection bias. PIAAC and PISA share few background variables. We control for the number of books at home at age 15, a strong predictor of cognitive ability and that was included in both studies (Sikora, Evans, & Kelley, 2019). The variable was coded as follows: 1 = 10 books or fewer; 2 = 11 to 25 books; 3 = 26 to 100 books; 4 = 101 to 200 books; 5 = 201 to 500 books; 6 = more than 500 books. Gender was also added as a control and full factorial interactions between all predictors were added. Gender differences in test-taking motivation have been well-documented (Borgonovi & Biecek, 2016).

We generated weights using propensity scores to balance the two populations and achieve pseudo-equivalence. Linear linking was then applied to link the scales. We performed linking 5 times because a set of 5 PVs was involved in linking. Finally linking errors were computed as described in the General Method section.

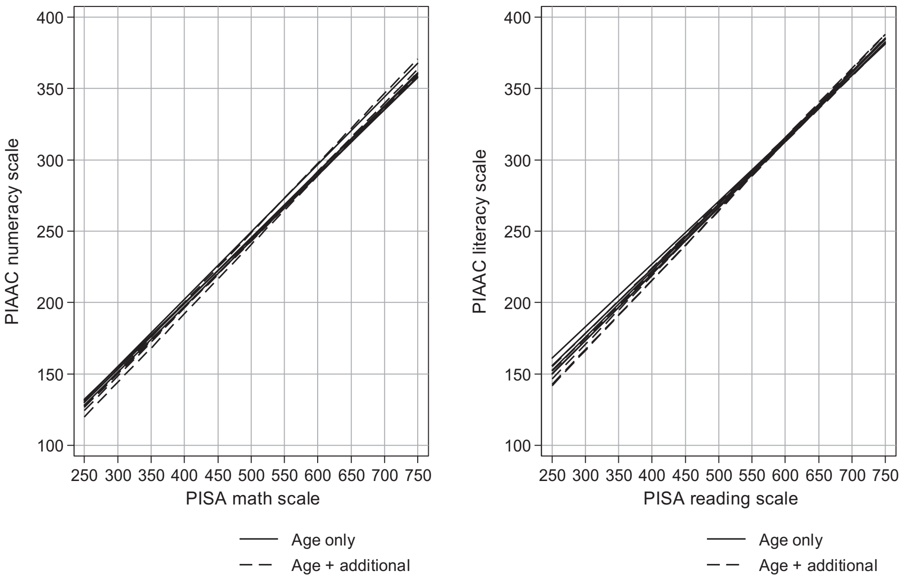

Figure 2. Concordance between PISA and PIAAC using different weights and five plausible values.

Study 2 Results

We evaluated two models to generate propensity scores: a model that controls for age only and a model that includes the additional controls described above. The model with additional covariates fits the data slightly better (BIC –132840.554 vs. –131890.072 and Efron's R^2: .940 vs .942) than the one with only age as a covariate, but the difference between the two is small. The percentage of correctly classified individuals and the R^2s in the two models are virtually identical. This confirms that "age" trumps all other effects when predicting propensity scores.

Results show that the estimated probabilities and weights associated with PIAAC participants being in the PISA sample approach zero at the age of 20, the upper bound for the PIAAC sample used in the analysis. This means that effectively only 1,000 of the youngest respondents from PIAAC were used for linking: the very small weight associated with older PIAAC participants means that, in practical terms, they do not contribute any information to the estimation (see Figure A1 in the Annex.)

Table A1 in the Annex presents descriptive statistics for key variables used in the analyses: age, number of books, % females, % of respondents who report being still in education (this variable is presented for validation purposes only). For PIAAC, results are presented before and after weighting. Table A1 suggests that the reweighted data match the PISA sample well.

Figure 2 illustrates concordance scores estimated using age-only propensity score weighting and weighting based on age and additional controls. The linear linking of the PISA and PIAAC math-numeracy assessment is presented in the left panel while the linking between the PISA and PIAAC reading-literacy assessment is presented in the right panel.

Figure 2 suggests that the variability in the slope of linking functions is very small overall but is largest in the upper and lower tails of the proficiency distributions. Since linear functions are used, the larger differences observed in the tails reflect variability in the slopes of the five sets, but not variability within each of the five functions, which is constant. This suggests that linking is less accurate in the upper and lower tails of proficiency. The difference in linking estimates between the two weighting procedures (age-only and age + additional controls) is small. Most of the variance in the estimated slopes is due differences in estimates across plausible values (i.e., variance due to measurement error) rather than variance across weighting estimates.

Table 4 presents linking constants that can be used to transform the PISA scales into PIAAC scales and vice versa. Constants are presented for each of the five plausible values and can be applied to respondent level data. To transform scales from aggregates, constants should be derived from average slopes and intercepts from the five plausible values.

The parameters reported in Table 4 allow to convert individual PISA scores into PIAAC scores and *vice versa.* In practice, many researchers and policy makers are interested in identifying concordance values and the associated margin of error for specific scores. Important scores are absolute benchmarks of proficiency defining levels of competencies according to the PISA and PIAAC assessment frameworks (the PISA and PIAAC proficiency levels) (OECD, 2013a,b) and specific percentiles characterising low levels of achievement (10th percentile), high levels of achievement (90th percentile) and the median.

Table A2 in the Annex illustrates the linking conversion (and the precision of such conversion) for the six PISA mathematics proficiency levels and the seven PISA reading proficiency levels by reporting the transformed scores estimated with age-only weighting and with age and additional variables weighting, as well as standard errors associated with each transformed score. Table A3 illustrates the linking conversion for the scores characterising the five PIAAC proficiency levels in numeracy and literacy. Linking errors for scores in the middle part of the distributions are considerably smaller than at the extremes.

Table 5 presents two sets of standard errors for the mean, the 10th, 50th, and the 90th percentiles. The first set takes into account measurement, sampling, and linking uncertainty, while the second set takes into account measurement and sampling uncertainty. Results indicate that by neglecting linking error, standard errors are heavily underestimated at the 90th and 10th percentiles, while at the mean and at the median, discrepancies are smaller.

Discussion

We show that the pseudo-equivalent group design can enhance the quality of linking in the presence of nonequivalent groups and illustrate how it performs compared to equivalent group design in the presence of variables that can be used to balance the groups. Monte-Carlo simulations suggest that pseudo-equivalent group design is particularly useful whenever there is a large overlap across the two groups with

Table 4

Linking Constants for Linear Linking

		Mathematics/Numeracy						Reading/Literacy			
		From PISA to PIAAC		From PIAAC to PISA				From PISA to PIAAC		From PIAAC to PISA	
Weighting procedure	PV	Slope	Intercept	Slope	Intercept	Weighting procedure	PV	Slope	Intercept	Slope	Intercept
Age only	1	.459	15.867	2.178	–34.561	Age only	1	.452	43.196	2.214	–95.624
	2	.450	19.649	2.221	–43.637		2	.469	32.929	2.134	–70.267
	3	.472	13.557	2.117	–28.706		3	.465	36.207	2.150	–77.830
	4	.461	12.997	2.171	–28.222		4	.465	34.269	2.153	–73.771
	5	.462	14.957	2.167	–32.412		5	.440	51.457	2.276	–117.090
Average		.461	15.406	2.171	–33.507	Average		.458	39.612	2.185	–86.916
Age + additional	1	.464	11.734	2.155	–25.286	Age + additional	1	.459	37.329	2.178	–81.299
	2	.465	10.538	2.152	–22.675		2	.492	18.956	2.033	–38.541
	3	.488	5.147	2.051	–10.558		3	.483	25.738	2.069	–53.254
	4	.482	–.411	2.074	.851		4	.486	21.561	2.060	–44.411
	5	.478	5.222	2.093	–10.929		5	.458	40.773	2.182	–88.946
Average		.475	6.446	2.105	–13.719	Average		.476	28.871	2.104	–61.290

Table 5

Standard Errors for Mean 10th, 50th, and 90th Percentiles

		Sampling + Measurement + Linking				Sampling + Measurement				Underestimation of Error by			
Statistic		Math num age	Math num age+	Read lit age	Read lit age+	Math num age	Math num age+	Read lit age	Read lit age+	Math num age	Math num age+	Read lit age	Read lit age+
PISA linked to PIAAC	Mean	4.80	5.24	4.10	4.85	2.96	3.44	2.36	2.54	62%	52%	73%	91%
	p10th	7.13	8.30	6.88	8.29	2.73	3.23	4.46	4.87	161%	157%	54%	70%
	p50th	4.79	5.27	4.05	4.80	3.08	3.52	2.50	2.67	56%	50%	62%	80%
	p90th	6.39	7.26	4.67	5.80	4.85	5.71	2.12	2.78	32%	27%	120%	109%
PIAAC linked to PISA	Mean	10.65	11.22	9.87	10.93	7.57	8.38	6.36	6.32	41%	34%	55%	73%
	p10th	14.99	17.11	18.36	20.45	7.36	8.12	11.71	11.83	104%	111%	57%	73%
	p50th	11.20	11.83	9.61	10.65	8.28	9.15	6.11	6.12	35%	29%	57%	74%
	p90th	15.81	17.27	9.88	11.70	11.05	12.56	5.72	6.65	43%	38%	73%	76%

respect to balancing variables and when the correlation between such variables and ability is medium or high.

We applied the pseudo-equivalent group approach to establish linkages between the PISA and PIAAC assessments and find that the results we obtain provide reasonable accurate linking that can be used for group-level comparisons. We applied the pseudo-equivalent group approach to link PISA and PIAAC because, although the two assessments differ along a number of dimensions (in particular age), there is a high degree of similarity in the two assessment frameworks and both tests are low-stakes for test-takers: similarities in frameworks and test motivation are key preconditions for using the pseudo-equivalent group approach. We estimate scale transformations across the two studies using a combination of statistical matching through propensity scores and linear transformation in matched samples.

A successful production of a concordance in this setting depends on the ability of the propensity score technique to capture the selection process. Theoretically, the selection process should be captured by age only since random age-restricted samples were taken from both assessments. Therefore, samples, conditional on age, should, in theory, be equivalent. However, different sampling frames between the studies and access to respondents might have a bearing on results. In PISA, schools are the primary sampling unit and students are sampled within selected schools. PIAAC samples were drawn directly from the national registry. Discrepancies in the overlap between the national registry and the schooled population should not be a major issue in Poland, a country where the PISA sample selectivity is low and the PISA target population (which comprises students enrolled in schools at grade 7 or above) reflects well the overall population of the same age group. Education is compulsory in Poland up to the age of 18 and graduation rates are high: 94% of 25- to 34-year-olds held the equivalent of a high-school degree in 2012 (OECD, 2013a). However, since sample selection processes might vary between samples, reducing their equivalence we employed two additional conditional variables: gender and the number of books individuals reported having in their homes at the age of 15 (as well as interactions between these variables), since these might play a role in the selection process.

The application of the pseudo-equivalent group approach to PISA and PIAAC is interesting for both methodological and substantive considerations. At the methodological level, linking the PISA and PIAAC samples requires to link two studies that overlap with respect to age, but the overlap is small because PIAAC captures a considerably wider population in terms of age spectrum than PISA (such that a large proportion of the PIAAC sample consists of individuals who are over 20, while the primary PISA target population consists of 15-year-old students). Furthermore, because ability is correlated with age in PIAAC, linking PISA and PIAAC allows us to apply the Monte-Carlo simulation framework that we developed to explore under which conditions pseudo-equivalent group approach outperforms the equivalent group approach. At the substantive level, establishing a link between PISA and PIAAC could enable policy makers and researchers to explore issues of achievement growth and how this may differ across population groups. However, because all our estimates were conducted using PISA and PIAAC Polish samples, they may not be easily generalizable to other countries.

Acknowledgments

The authors would like to thank Matthias von Davier, William Thorn, Kentaro Yamamoto as well as anonymous referees and the associate editor Stephen Sireci for very helpful comments on previous versions of this manuscript.

The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the Organisation for Economic Co-Operation and Development.

References

ACT. (2011). Technical manual for the PLAN-PISA international benchmarking linking study. Iowa City, IA: ACT.

Austin, P. C. (2009). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The International Journal of Biostatistics*, *5*(1), 171–184.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, *46*(3), 399–424.

Beaton, A. E., & Gonzales, E. J. (1993). Comparing the NAEP trial state assessment results with the IAEP international results, National Academy of Education Panel on the NAEP Trial State Assessment (Research Report).

Borgonovi, F., & Biecek, P. (2016). An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learning and Individual Differences*, *49*, 128–137.

Cartwright, F., Lalancette, D., Mussio, J., & Xing, D. (2003). *Linking provincial student assessments with national and international assessments*. Statistics Canada.

Cook, L., Eignor, D., & Schmitt, A. (1988). *The effects on IRT and conventional achievement test equating results of using equating samples matched on ability: ETS Research Report RR-88-52)*. Princeton, NJ: Educational Testing Service.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. *(Vol. 38)*: SIAM. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (1999). *Equivalence and linkage of educational tests*. Washington: National Academies Press.

Frölich, M. (2004). Finite-sample properties of propensity-score matching and weighting estimators. *Review of Economics and Statistics*, *86*(1), 77–90.

Haberman, S. J. (1984). Adjustment by minimum discriminant information. *The Annals of Statistics*, *12*(3), 971–988.

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics*, *40*(3), 254–273.

Hambleton, R. K., Sireci, S. G., & Smith, Z. (2009). Evaluating NAEP achievement levels in the context of international assessments. *Applied Measurement in Education*, *22*, 376–393.

Hanushek, E. A., & Woessmann, L. (2013). The role of international assessments of cognitive skills in the analysis of growth and development. In M. Von Davier, E. Gonzalez, I. Kirsch, & K. Yamamoto (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 47–65). Dordrecht, Netherlands: Springer.

Jia, Y., Phillips, G., Wise, L. L., Rahman, T., Xu, X., Wiley, C., & Diaz, T. E. (2014). *2011 NAEP-TIMSS linking study: Technical report on the linking methodologies and their evaluations* (NCES 2014–461). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Johnson, E. G. (1998). Linking the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS): A technical report. Research and Development Report: ETS.

Johnson, E. G., Cohen, J., Chen, W.-H., Jiang, T., & Zhang, Y. (2003). *2000 NAEP–1999 TIMSS linking report* (NCES 2005-01). Washington, DC: National Center for Education Statistics, U.S. Department of Education.

Kaplan, D., & McCarty, A. (2013). Data fusion with international large scale assessments: A case study using the OECD PISA and TALIS surveys. *Large-Scale Assessments in Education*, *1*(1), 1–26.

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

Lawrence, I. M., & Dorans, N. J. (1990). Effect on equating results of matching samples on an anchor test. *Applied Measurement in Education*, *3*(1), 19–36.

Linn, R. L. (1975). Anchor test study: The long and the short of it. *Journal of Educational Measurement*, *12*(3), 201–2014.

Linn, R. L., McLaughlin, D., & Thissen, D. (2009). Utility and validity of NAEP linking efforts. *American Institutes for Research*. Washington, DC: American Institutes for Research, NAEP Validity Studies Panel.

Lim, H., & Sireci, S. G. (2017). Linking TIMSS and NAEP assessments to evaluate international trends in achievement. *Education Policy Analysis Archives*, *25*, 11.

McCaffrey, D. F., Lockwood, J., & Setodji, C. M. (2013). Inverse probability weighting with error-prone covariates. *Biometrika*, *100*(3), 671–680.

OECD. (2013a). *PISA 2012 Results: What students know and can do*. Paris: OECD Publishing.

OECD. (2013b). *The survey of adults skills. Reader's companion*. Paris: OECD Publishing.

Paek, I., Liu, J., & Oh, H. (2006). *Investigation of propensity score matching on linear/nonlinear equating method for the PSAT/NMSQT*. Princeton, NJ: Educational Testing Service.

Pashley, P. J., & Phillips, G.W. (1993). *Toward world-class standards: A research study linking international and national assessments*. Princeton, NJ: Educational Testing Service.

Phillips, G. W. (2007). *Expressing international education achievement in terms of U.S. performance standards: Linking NAEP achievement levels to TIMSS*. Washington, DC: American Institutes for Research.

Radwan, N., & Xu, Y. (2012). Comparison of the Performance of Ontario Students on the OSSLT/TPCL and the PISA 2009 Reading Assessment. Ontario: Education Quality and Accountability Office.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley & Sons.

Sansivieri, V., & Wiberg, M. (2017). IRT observed-score with the non-equivalent groups with covariates design. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology—81st annual meeting of the psychometric society*, Asheville, North Carolina, 2016. New York: Springer.

Sikora, J., Evans, M. D. R., & Kelley, J. (2019). Scholarly culture: How books in adolescence enhance adult literacy, numeracy and technology skills in 31 societies. *Social Science Research*, *77*, 1–15.

Szaleniec, H., Grudniewska, M., Kondratek, B., Kulon, F., & Pokropek, A. (2013). Results of the 2002–2010 lower secondary school leaving exams on a common scale. *EDUKACJA*, *1*, 115–134.

United Nations. (2017). The Sustainable Development Goals Report 2017. New York: United Nations Publishing.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2), 114–128.

Wu, M. (2009). A comparison of PISA and TIMSS 2003 achievement results in mathematics. *Prospects*, *39*(1), 33–46.

Wallin, G., & Wiberg, M. (2017). Non-equivalent groups with covariates design using propensity scores for kernel equating. In L. A. van der Ark, M. Wiberg, S. A. Culpepper, J. A. Douglas, & W.-C. Wang (Eds.), *Quantitative psychology—81st annual meeting of the psychometric society*, Asheville, North Carolina, 2016. New York: Springer.

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement*, *39*(5), 349–361.

Yamamoto, K. (2002). Estimating PISA student scores on the IALS prose literacy scale. Technical report. Princeton, NJ: Educational Testing Service.

Yu, L., Livingston, S., Larkin, K., & Bonett, J. (2004). Investigating differences in examinee performance between computer-based and handwritten essays (ETS Research Report RR-04-18). Princeton, NJ: *Educational Testing Service*.

Authors

ARTUR POKROPEK is a Head of Data Science and Machine Learning Group (DSMLG), Educational Research Institute, 8 Górczewska Street, Warsaw, Poland; a.pokropek@ibe.edu.pl. His primary research interests include psychometric and statistical methods.

FRANCESCA BORGONOVI is a British Academy Global Professor, Institute of Education, University College London, 55-59 Gordon Square, London WC1H 0NU, United Kingdom; f.borgonovi@ucl.ac.uk. Her primary research interests include inequalities in academic achievement, attitudes, and self-rated beliefs.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure A1. Predicted probabilities and weights from Model 2 against age in PIAAC sample.
Table A1. Mean Values for Age, Number of Books, % Females, % Respondents in School in the PISA and PIAAC Samples before and after Weighting. Standardized Mean Differences in Brackets.
Table A2. Concordance of PISA to PIAAC Scores for Numeracy and Literacy Scales.
Table A3. Concordance of PIAAC to PISA Scores for Mathematics and Reading Scales.