Large-scale Assessments in Education
a SpringerOpen Journal

RESEARCH Open Access

Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS surveys

David Kaplan[1*] and Alyn Turner McCarty[2]

*Correspondence:
dkaplan@education.wisc.edu
[1]Department of Educational Psychology, University of Wisconsin, Madison, USA
Full list of author information is available at the end of the article

Abstract

Background: In the context of international large scale assessments, it is often not feasible to implement a complete survey of all relevant populations. For example, the OECD Program for International Student Assessment surveys both students and schools, but does not obtain information from teachers. In contrast the OECD Teaching and Learning International Survey assesses teachers and schools but does not assess students. Clearly, important information is missing from both assessments. One approach to obtaining information from both surveys is through data fusion – a variety of methods that can be used to create a synthetic data set containing information from both surveys.

Methods: This paper presents an experimental evaluation of a representative group of data fusion methods using data from Iceland – the only OECD country that implemented both PISA and TALIS to all members of the relevant populations.

Results: On the basis of a set of validity criterion we find that Bayesian bootstrap predictive mean matching and the EM-bootstrap methods perform best with respect to creating a usable synthetic data file for research purposes.

The OECD Program for International Student Assessment (PISA) and the OECD Teaching and Learning International Survey (TALIS) constitute two of the largest ongoing international student and teacher surveys presently underway. Data generated from these surveys offer researchers and policymakers opportunities to identify particular educational institutional arrangements – that is, how aspects of educational systems are organized – to promote equality of educational opportunity both within and between countries.

Naturally, policy makers are interested in all three levels of the school system – students, teachers, and schools, in order to fully understand within and between country differences in the inputs, processes, and outcomes of education. However, a serious limitation of these data collection efforts is that each survey is missing an important component of the educational system in their design – namely, PISA is missing teacher level data and TALIS is missing student level data. The PISA and TALIS surveys are not, at present, linked. An ideal approach to linking the PISA survey to the TALIS survey is to sample

Springer

© 2013 Kaplan and McCarty; licensee Springer. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

schools and administer both PISA and TALIS. However, because a simultaneous administration of both surveys may not be feasible for many countries, this limits the extent to which information unique to each survey can be understood jointly.

In contrast to a design-based approach, a statistical approach to linking the PISA survey to the TALIS survey involves the creation of a synthetic cohort of data – that is, a new data file that combines information from both surveys. Two approaches are common and will be explored in this study. The first is *statistical matching*, which involves finding units in the two separate files that are "close" in some statistical sense, and then filling in missing data with the data from the unit and its match. The second approach involves *imputation*, which treats the goal of creating a synthetic file as a large missing data problem. The approach is to use information common to both surveys to impute plausible values of the missing data occurring in both surveys. Throughout this paper, we will use the generic term *data fusion* to mean the creation of the synthetic data file.

The current study is a systematic evaluation of a set of data fusion methods focused on the goal of creating a synthetic file of PISA and TALIS data. We evaluate the extent to which each method provides a synthetic data set that maintains the essential properties of PISA and TALIS, concentrating on a set of validity criteria established by Rässler (2002) as described below.

Our evaluation relies on an experimental comparison of the validity of each method relative to a clearly defined standard. For this purpose, we use data from Iceland. We chose Iceland because it is the only OECD country that implemented PISA and TALIS on the population of PISA students, all TALIS teachers, and all PISA and TALIS schools[a]. The experimental study will provide a *proof of concept* that fusing PISA and TALIS is feasible for countries that wish to draw on the added value of both surveys for research and policy analysis.

The organization of this paper is as follows. In the next section, we outline the problem of data fusion with particular focus on validity criteria that can be used to evaluate the quality of data fusion. Next, we outline the methods to be examined in this paper. It should be noted that a large number of methods exist for data fusion. We will examine six methods that are representative of the broad array of data fusion methods available, including non-parameteric and parametric algorithms. Our focus will also be on methodologies that are available within the R statistical programming environment (R Development Core Team, 2010). Our focus on the R statistical programming environment reflects our view that the open source and free nature of R can allow maximum accessibility across all countries to support the fusion of data sources generally and PISA and TALIS specifically. Next, we will present the design of our study. The results will follow. The paper closes with recommendations and limitations resulting from fusing PISA and TALIS. Annotated software code is made available in the appendices.

The policy context

Effective educational policy rests on the availability of reliable information about both the structure and process of educational systems. In this section we describe one potential policy question that can be more fully understood by fusing PISA and TALIS data. However, the application of data fusion is not limited to this particular question.

PISA obtains samples of students across more than 40 countries and economies, allowing researchers to relate variation in characteristics of national educational institutions to

levels of performance and inequality in student learning. In other words, researchers can use PISA to identify particular educational institutional arrangements that promote educational excellence and equality among students. For example, recent research utilizing data from PISA suggests that countries with a more strongly differentiated educational system tend to have higher levels of inequality of educational opportunity by social class and race/ethnicity; and countries with a more standardized educational system have lower levels of inequality of opportunity compared to those with unstandardized systems (Van de Werfhorst and Mijs 2010).

Although much has been documented relating institutional arrangements to student performance, more recently the focus has turned to detailed descriptions of how variation in the way educational systems are structured shapes what takes place in the classroom. In other words, more attention is being paid to how the *process* of education varies within and between countries. PISA administers surveys of students and school administrators, and the patterns revealed from their responses suggest that the best performing education systems embrace the diversity in students capacities, interests, and social background with individualized approaches to learning. These education systems also provide clear and ambitious standards focused on complex, higher order thinking, and prioritize teacher and administration quality (OECD 2010b).

An additional source of data on school processes comes from TALIS. While PISA links institutional characteristics to student performance, TALIS links institutional characteristics to aspects of school and classroom climate from the perspective of teachers and school administrators. For example, TALIS asks teachers and principals about the disciplinary climate of the school. Extant research suggests that classroom disciplinary climates affect student outcomes and attainment, and that many countries consider discipline a high priority policy issue (OECD 2009). However, only by linking the TALIS and PISA surveys can researchers fully model the relations between institutional differentiation, disciplinary climate, and student learning.

Because learning occurs in the context of classrooms, aspects of teacher practices and classroom climate are key to understanding the mechanisms through which policy decisions might impact educational performance and inequality in learning. However, at present it appears infeasible, for practical and/or political reasons, to design and implement a large three level-international survey with students nested in classrooms, and in turn nested in schools. An alternative approach, then, would be to statistically combine PISA and TALIS in order to more carefully and universally describe school systems, with the intent of reporting associations between performance, equality, and educational policy, and how these factors combine to produce a social system which can be described from the perspective of families, students, teachers, and school administrators. Statistically combining two, relatively distinct, data sources is the goal of data fusion.

Background on data fusion

Data fusion involves filling in missing data from two surveys in order to obtain a "synthetic" set of data that can be considered as generated from a population of relevance to the original surveys. It is convenient to categorize data fusion methods as either nonparametric (i.e. those not based on an underlying model for the observed and missing data) or parametric (i.e. methods based on assuming an underlying model for observed

and missing data described by a set of parameters). In both cases, however, the problem is one of addressing the issue of missing data – that is, TALIS is missing student level data available from PISA, and PISA is missing teacher level data available in TALIS.

In the context of PISA and TALIS, we can consider two types of missing data; unit and item non-response. However, when considering the fusion of the two data sets, a very large amount of unit missing data obtains because the surveys contain different items and units of analysis. What is required to move forward with data fusion is a general theoretical framework for the problem of missing data.

Following the seminal work of Rubin (1976, see also; Little and Rubin 2002; Schafer 1997; Enders 2010) let R be a missing data indicator, taking the value of 1 if the data are observed, and 0 if the data are missing. Let $f(R)$ be the associated probability distribution of R. Further, let y be the complete data, y_{obs} represent observed data and y_{miss} represent missing data. Finally, let ϕ be the scalar or vector-valued parameter describing the process that generates the missing data. The underlying mechanism that generates missing data can be considered either *ignorable* or *non-ignorable*. An ignorable missing data mechanism is one in which inferences are not affected by the process that generated the missing data.

There are two types of missing data mechanisms that can be considered ignorable. Take, for example, two variables from the TALIS teacher questionnaire, say teacher job satisfaction (jobsat) and teacher self-efficacy (selfef), and assume that there is missing data on selfef. If the missing data on selfef is unrelated to the observed values of both jobsat and selfef, then the missing data are considered to be *missing completely at random* or *MCAR*. More formally, MCAR implies that

$$f(R|y,\phi) = f(R|\phi), \qquad \text{for all } y,\ \phi. \tag{1}$$

Under the assumption of MCAR, such methods as listwise deletion or regression imputation can be used to treat missing data (although they might not be desirable approaches for other reasons). Next, imagine a situation in which the missing data on selfef is unrelated to observed selfef, but may be related to observed jobsat. For example, perhaps teachers with lower job satisfaction tend not to report their levels of self-efficacy. This type of missing data is referred to as *missing at random* or *MAR*. Again, in terms of our notation, MAR implies that

$$f(R|y,\phi) = f(R|y_{obs},\phi), \qquad \text{for all } y_{miss},\ \phi \tag{2}$$

Under MAR, inferences will be valid, and there now exist many methods for handling missing data under the assumption of MAR.

In real data contexts, MCAR and MAR are fairly unrealistic assumptions. A more realistic situation is one in which the missing data mechanism is non-ignorable. Taking our example of job satisfaction and self-efficacy, we may find that missing data on self-efficacy is related to self-efficacy. That is, perhaps teachers with low self-efficacy do not report their levels of self-efficacy regardless of their level of job satisfaction. This type of missing data problem is referred to as *not missing at random* or *NMAR*. More formally,

$$f(R|y,\phi) = f(R|y_{obs},y_{miss},\phi), \qquad \text{for all } y,\ \phi \tag{3}$$

Under NMAR, inferences derived from conventional approaches are not valid, and what is required is a substantive model of interest that incorporates a model of the missing data process.

Despite the fact that NMAR is perhaps the more realistic scenario for missing data problems, advances in handling missing data have generally been made under the assumption of MAR, where the assumption of MCAR is considered mostly unrealistic. There is, however, one unique situation in which MCAR might be reasonably expected to hold – and that is where the missing data are missing by design. One example of missing by design are assessment plans that involved balanced incomplete block spiralling frameworks (see e.g. Kaplan 1995) – such as the design for the cognitive outcome assessments in PISA. Another example of concern to this paper is the case of statistically fusing different surveys. In the case of PISA and TALIS, the two surveys have no units in common but do have variables in common – in particular, variables from the survey of principals in both the PISA and TALIS samples. Because there are no units in common across the two surveys, the missing data are reasonably considered to be MCAR[b].

Levels of validity in data fusion

An immediate question that is raised when considering the problem of filling in missing data, particularly in the context of large sample surveys such as PISA and TALIS, is the validity of the method used to fuse the data sets. This is of prime importance to our goal of fusing PISA and TALIS insofar as the results of these surveys carry major policy consequences. An important discussion of the problem of validity in the context of data fusion can be found in Rässler (2002).

We begin by introducing notation that closely follows Rässler (2002). Let Y denote the $n_p \times p$ matrix of data from PISA, let X denote the $n_t \times t$ matrix of data from TALIS, and let Z be the $n_c \times c$ matrix of data common to both PISA and TALIS, $n_p = 1, 2, \ldots N_p; n_t = 1, 2, \ldots N_t; n_c = 1, 2, \ldots, N_c; p = 1, 2, \ldots P; t = 1, 2, \ldots T; c = 1, 2, \ldots C$. Further, denote by $f_{X,Y,Z}$ the joint density function of the X, Y, and Z, and $\tilde{f}_{X,Y,Z}$ be the corresponding joint density function after data fusion. Specific observations drawn from X, Y, and Z will be denoted as x_i, y_i, and z_i, respectively. Following Rässler (2002), four levels of validity can be distinguished when considering the problem of data fusion.

First level validity: preserving individual values

The most difficult level of validity that can be achieved in data fusion concerns the ability of the algorithm to reproduce the true but unknown individual values of the sample data. That is, does the algorithm provide the values for the missing data in PISA and TALIS that would have been observed had those variables been presented and answered? Because the true individual values are unknown, the only way that first level validity can be established is via a simulation study Rässler (2002). Although the data from Iceland provide an opportunity to assess first level validity, it is usually impossible or at least unnecessary to achieve this level of validity. First, in order for the algorithm to reproduce individual values perfectly, the missing data would have to be perfectly predicted by the data common to both surveys (as would be the case if a unique identifier is included in the two data sets). Second, the algorithms that we will be examining are designed to reproduce expected values under a given model, and not individual values. Third, imputation algorithms are designed to produced a dataset that can be used for secondary analyses based

on summary statistics and not individual values. Thus, for this paper, we will not assess first level validity.

Second level validity: preserving joint distributions

The idea behind second level validity is that the joint distribution of all of the variables in the synthetic data set be preserved after data fusion. For this to be true, we must first assume that the PISA and TALIS schools (within a country) were sampled independently within and across the surveys. Then, we can assume that the synthetic file is a random sample from a synthetic distribution. Rässler (2002) shows that this will hold only if the variables unique to PISA and unique to TALIS are conditionally independent given the variables common to both surveys. That is, $f_{X,Y|Z} = f_{X|Z} f_{Y|Z} = \tilde{f}_{X,Y|Z}$.

Third level validity: preserving covariance/correlation structures

Both PISA and TALIS not only inform education policy for countries, but they both serve as important sources for research and analysis. In that case, statistical modeling methods that rely on the covariances and higher order moments of the data – such as regression analysis and factor analysis – are often employed as analytic methodologies. If the goal is statistical modeling of the synthetic data, then the covariance structure of the data before and after matching should be the same. As with second level validity, the synthetic data set should represent a sample from a synthetic population that has the same covariance structure as the actual population of interest. Following Rässler (2002), if we let $\widetilde{cov}(X, Y)$ be the covariances of the variables in the synthetic population, and $cov(X, Y)$ be the covariances of the true population, then the only way in which these two covariances are equal to each other is if X and Y are on average conditionally uncorrelated give the common variables z used in the match. To see this, let

$$\widetilde{cov}(X, Y) = cov[\, E(X|Z), E(Y|Z))] \,, \tag{4}$$

be the synthetic covariances. Then, because

$$cov(X, Y) = E[cov(X, Y|Z)] + cov[E(X|Z), E(Y|Z)] \,, \tag{5}$$

it follows that the only way for $\widetilde{cov}(X, Y) = cov(X, Y)$ is if the $E(cov(X, Y|Z)) = 0$. This paper will provide an assessment of third level validity.

Fourth level validity: preserving marginal distributions

The lowest level of validity and a minimum requirement for data fusion is that the marginal distributions of the individual variables in the original surveys be preserved after the statistical match. Formally, if $\hat{\tilde{f}}_y$ is the empirical marginal distribution of the PISA variables and $\hat{\tilde{f}}_{y,z}$ is the empirical joint distribution of the PISA variables and variables common to PISA and TALIS in the synthetic sample, then after the match they should not differ meaningfully from $\hat{f}_y$ and $\hat{f}_{y,z}$, the marginal and joint distributions from PISA, respectively. We will provide a descriptive assessment of fourth level validity.

Methods

In this section, we describe the data fusion methods we will evaluate in the context of the PISA-TALIS fusion. As noted earlier, there are scores of different methods that can be used for data fusion, and it is beyond the scope of this paper to evaluate every approach that is currently available. Our approach for this paper, therefore, is to examine a handful

of the most representative approaches and to provide a detailed evaluation of their usefulness and validity in providing a synthetic file. For our experimental study with Iceland, we will concentrate on the third and fourth levels of validity described earlier because these are the most important levels for research and policy analysis using PISA and TALIS.

A common feature of all data fusion methods, and, admittedly, a limitation in the context of PISA and TALIS, is that the data must be aggregated to a common unit of analysis. For PISA and TALIS, the only level of analysis common across the surveys is the school. Thus, student and teacher data must be aggregated to their respective school level before data fusion can proceed. In doing so, the multilevel structure of each survey is lost. Data fusion, therefore, takes place by identifying school level variables that are common across PISA and TALIS. Any number of variables will do, but the more variables in common, the more information can then be brought to bear to create a synthetic file. In cases in which a variable has been measured on a different scale across the two surveys, the extant literature suggests that they should be converted to *z*-scores, even if the variables are categorical (e.g. Rässler 2002). Differences in the scales of categorical variables can also be handled by collapsing one, or both, to a common set of categories.

We organize this section as follows. First, we will describe a non-parametric approach based on so-called "hot deck imputation" – namely *distance hot deck matching*. The remaining approaches will be parametric and based on file concatenation and multiple imputation (Rubin 1986,1987). The file concatenation perspective sees data fusion as a missing data problem with the goal of imputing values for the missing data. However, rather than imputing a single value for a missing data point and treating it as fixed, the multiple imputation framework accounts for uncertainty about the missing data by creating multiple plausible missing values resulting in multiple data sets. The data sets are then combined in specific ways for analysis purposes.

Within the multiple imputation perspective, we will describe approaches derived from the frequentist and Bayesian frameworks of statistics. Within the frequentist framework, we will examine two methods – *stochastic regression imputation* and *predictive mean matching*. Within the Bayesian framework, we will describe *Bayesian linear regression imputation via chained equations*, *Bayesian bootstrap predictive mean matching*, and the *EM boostrap* – the latter being a hybrid of Bayesian and frequentist methods.

Nonparametric hot deck matching

Hot deck imputation procedures require that a distinction be made between a "donor" data set and a "recipient" data set. As noted by D'Orazio et al. (2006), there are several factors that need to be considered when designating a donor and recipient data set. The two most important, according to D'Orazio et al. (2006) concerns the phenomenon under study and the accuracy of information contained in the two surveys. In the former case, matching PISA and TALIS should yield a synthetic data set that retains the ability to draw valid and reliable inferences of policy relevance. In the latter case, it does not make much sense to match two data sets in which the information from either or both surveys is inaccurate. An example concerns matching data sets when the matching units were obtained at very different time points. In such cases, it may not be reasonable to assume that the synthetic file represents independent and identically distributed observations from the same population. In the case of PISA and TALIS, it is true that these surveys were not implemented at the same time. At the school level within a country, the

argument would have to be made that TALIS schools are different from their corresponding PISA schools, perhaps due to the implementation of some country level policy during the interim in which PISA and TALIS were implemented. We are assuming that within a country, the time difference between the implementation of PISA and TALIS did not result in important exogenous changes across schools.

In addition to these substantive concerns, the sample sizes of the data sets are also a consideration. In the case of PISA and TALIS, the school sample sizes are markedly different; PISA, on average, samples twice as many schools as TALIS. Thus, it is common practice to assign the role of the recipient data set to the smaller of the two - in this case TALIS. We can see why this is reasonable. If TALIS was the donor survey, then records in TALIS would be imputed more than once into PISA, which could then artificially reduce the variability of the distributions of the variables in the synthetic data set.

The essence of hot deck imputation is that missing data in a recipient file (TALIS) are filled in with actual values from a donor data file (PISA) based on a pre-specified algorithm. This approach requires that the donor data set be at least as large or larger than the recipient data set. Once a PISA donor is found for a TALIS recipient, the missing data for the TALIS recipient is given the value of the PISA donor. The resulting synthetic data set has a sample size equivalent to that of the original TALIS sample. A number of algorithms exist for hot deck matching, however for this paper we will focus our attention only on nearest neighbor hot deck matching. For our analyses, will use the R program *StatMatch* (D'Orazio 2011) for non-parametric hot deck matching.

Distance hot deck matching

Distance hot deck matching is perhaps the oldest form of hot deck matching and has been used in a variety of applications. The idea is simple. The algorithm finds a school in PISA that is closest to a school in TALIS based on a chosen metric of "distance". For the purposes of this paper, we chose the Euclidean distance metric. For simplicity, following D'Orazio et al. (2006) let z be a single matching variable. Then, a donor from PISA for the t^{th} record in TALIS is chosen such that

$$d_{pt*} = \min_{1<t<n_t} |z_t^T - z_p^P|, \tag{6}$$

Once that school is found, the missing data for the TALIS school is given the value obtained from the PISA school. If two or more donor schools are found to match a TALIS school, then one school is chosen at random.

Frequentist approaches to data fusion

As noted earlier, in addition to nonparametric methods based on variants of hot deck imputation, parametric data fusion in the form of file concatenation and multiple imputation can also be considered. In this case, the resulting synthetic data set has a sample size which is the sum of the sample sizes of the separate surveys. In this section, we consider two frequentist-based statistical data fusion methods, both of which are implemented in the R software program *mice* (van Buuren and Groothuis-Oudshoorn 2010).

Stochastic regression imputation

A common approach to imputation is based on linear regression analysis. Under the assumption that the missing data are at least MAR, the regression imputation approach

uses linear regression to obtain predicted values for the missing observations. Thus, in the case of PISA and TALIS, variables that are unique to TALIS would be regressed on the variables common to PISA and TALIS. From here, missing data is filled in using the predicted values of the TALIS missing data. The method proceeds similarly for filling in missing PISA data.

The difficulty with linear regression imputation is that because the imputed values are predictions from a regression equation, they will lie precisely on the regression line and hence lead to underestimation of residual variability. This lack of variability in the imputed values is clearly not realistic, and, moreover, will result in an overestimation of the correlations (and hence R^2) in subsequent analyses. To remedy this problem, a residual value is drawn from a normal distribution with a mean of zero and a variance equal to the residual variance of the regression equation. This residual value is added to the predicted value, yielding *stochastic regression imputation*.

With only one residual drawn from a normal distribution, the imputed missing data value is still treated as unique and fixed. Given that missing data are, by definition, unknown, it may be more reasonable to obtain multiple predicted values of the missing data by drawing multiple residual values from the normal distribution. These multiple draws, when added to the regression equation, will yield multiple data sets each with a different predicted value for the missing data. In this way, uncertainty in the predicted values of the missing data are addressed. Subsequent analyses are then based on analyzing all of the data sets simultaneously and then pooling the results according to rules set down by Rubin (1987)[c]. For this study, we use the *norm.nob* option in the R software program *mice* (van Buuren and Groothuis-Oudshoorn 2010) to conduct stochastic regression imputation.

Predictive mean matching

Regression imputation and hot deck matching set the ground work for so-called *predictive mean matching* introduced by Rubin (1986). In our context, the essential idea is that a missing value in PISA is imputed by matching its predicted value based on regression imputation to the predicted values of the observed data on the basis of some distance metric. Then, the procedure uses the actual observed value for the imputation. That is, for each regression, there is a predicted value for the missing data and also a predicted value for the observed data. The predicted value for the observed data is then matched to a predicted value of the missing data using, say, a nearest neighbor distance metric. Once the match is found, the actual observed value (rather than the predicted value) replaces the missing value. In this sense, predictive mean matching operates much like hot deck matching. For this study, we use the *pmm* option within *mice* to conduct predictive mean matching.

Bayesian approaches to data fusion

In the previous section, we concentrated on two approaches to data fusion that lie within the so-called "frequentist" paradigm of statistics. This paradigm is most closely associated with Sir R. A. Fisher and rests on a view that equates probability with long run frequency and the idea of identically repeatable experiments. Along with likelihood theory (also associated with Fisher), the general frequentist paradigm views parameters (such as population means, variances, and regression coefficients) as unknown and fixed. A sample,

taken from the population is then used to provide an estimate of the unknown parameters, and the notion of identically repeatable samples from the population allow us to estimate the sampling variability around the estimates of the model parameters.

In contrast to the frequentist school of statistics, the Bayesian school adopts an entirely different view of statistical inference. Specifically, the Bayesian school views all unknown quantities, and in particularly parameters, as random variables that can be described by a probability distribution that characterizes our uncertainty about the average value and variation of the parameter. This probability distribution is referred to in the Bayesian literature as the *prior distribution*. Bayes' theorem is used to link the prior distribution to the actual data distribution (analogously, the likelihood) yielding a *posterior distribution* of the model parameters (see Kaplan and Depaoli 2013, for an overview of Bayesian inference).

The central reason for adopting a Bayesian perspective to the problem of data fusion (and other missing data problems more generally) is that by viewing parameters probabilistically and specifying a prior distribution on the parameters of interest, the imputation method (described next) is *Bayesianly proper* (Rubin 1987) insofar as the imputations reflect uncertainty about the missing data as well as uncertainty about the unknown model parameters. Moreover, this view of statistical inference allows for the incorporation of prior knowledge of model parameters, which can further reduce uncertainty in model parameters. In the context of international large scale assessments, such prior knowledge can come from previous fusion studies, where the prior information can be encoded as values of the hyperparameters of the fusion model. It is important to point out that although the method of stochastic regression imputation described above has a Bayesian flavor, it is not Bayesianly proper insofar as it does not account for parameter uncertainty, but rather only uncertainty in the predicted missing data values.

Bayesian regression imputation via chained equations

In this section we concentrate our discussion on a *Bayesianly proper* form of multiple imputation using the method of chained equations[d]. The method of chained equations recognizes that in many instances, it might be better to engage in a series of single univariate imputations along with diagnostic checking rather than a omnibus multivariate model for imputation that might be sensitive to specification issues. An overview of imputations via chained equations can be found in van Buuren (2012).

The essence of the chained equations approach is that a univariate regression model consistent with the scale of the variable with missing data is used to provide predicted values of the missing data given the observed data. Thus, if a variable with missing data is continuous, then a normal model is used. If a variable is a count, then a Poisson model would be appropriate. This is a major advantage over other Bayesianly proper methods such as data augmentation (Tanner and Wong 1987) that assume a common distribution for all of the variables. Once a variable of interest is "filled-in", that variable, along with the variables for which there is complete data, is used in the sequence to fill in another variable. In general, the order of the sequence is determined by the amount of missing data, where the variable with least amount of missing data is imputed first, and so on.

Once the sequence is completed for all variables with missing data, the posterior distribution of the regression parameters are obtained, and the process is started again. Specifically, the filled-in data from the previous cycle, along with complete data are used

for the second and subsequent cycles (Enders 2010). The Gibbs sampler (Geman and Geman 1984) is used to generate the sequence of iterations. Finally, the algorithm can run these sequences in parallel m number of times obtaining m imputed data sets. For the purposes of this paper, we will utilized the *norm* option as implemented in *mice*.

Bayesian bootstrap predictive mean matching

Multiple imputation via chained equations is inherently a parametric method. That is, in estimating a Bayesian linear regression, the posterior distributions are obtained via Bayes' theorem which requires parametric assumptions. It may be desirable, however, to relax assumptions regarding the posterior distributions of the model parameters, and to do this requires a replacement of the step that draws the conditional predictive distribution of the missing data given the observed data. A hybrid of predictive mean matching, referred to as *posterior predictive mean matching*, proceeds first by obtaining parameter draws using classical multiple imputation approaches. However, the final step then uses those values to obtain predicted values of the data followed by conventional predictive mean matching.

Posterior predictive mean matching sets the groundwork for *Bayesian bootstrap predictive mean matching* (BBPMM). The goal of BBPMM is to further relax the distribution assumptions associated with draws from the posterior distributions of the model parameters. The algorithm begins by forming a Bayesian bootstrap of the observations Rubin (1981,1987). The Bayesian bootstrap (BB) is quite similar to conventional frequentist bootstrap (Efron 1979), except that it provides a method for simulating the posterior distribution of the parameters of interest rather than the sampling distribution of parameters of interest, and as such, is more robust to violations of distributional assumptions associated with the posterior distribution.

Following Rubin (1981), one replication of the Bayesian bootstrap is formed by generating $n-1$ random variates from a uniform (0,1) distribution $u_1, u_2, \ldots, u_{n-1}$. Next, these variates are sorted in ascending order and gaps $d_i = u_i - u_{i-1}$ are calculated, with $i = 1, 2, \ldots, n-1$, and where $d_0 = 0$ and $d_n = 1$. Then, the d_i are the probabilities attached to the data $y_i, (i = 1, 2, \ldots, n)$. For each BB replication, the parameters of interest are calculated. The resulting distributions of the parameters over all BB replications are the posterior distributions of the parameters of interest. Rubin (1981) shows that the Bayesian bootstrap gives results that are very similar to the conventional bootstrap, except that the interpretation is different: the Bayesian bootstrap simulates the posterior distributions of the parameters of interest and provides likelihood statements regarding the parameters, whereas the conventional bootstrap simulates the sampling distribution of the parameters and provides frequency statements regarding the parameters (Rubin 1981, pg. 131).

In the context of missing data, the Bayesian bootstrap can be used for multiple imputation. The algorithm follows closely the description in the previous paragraph. Again, let $y_{obs} = (y_1, y_2, \ldots, y_{n_{obs}})$. We draw n_{obs} random numbers from a uniform (0,1) distribution, sort them in ascending order, and let $u_0 = 0$ and $u_1 = 1$. The gaps d_i are calculated as above. Then a uniform(0,1) random number is drawn independently n_{miss} times. We impute y_i, $(i = 1, 2, \ldots, n_{obs})$ if $d_{i-1} < u < d_i$. For multiple imputation, this procedure is repeated m times.

With the Bayesian bootstrap as the foundation, BBPMM obtains estimates of the regression parameters from the BB sample. This is followed by the calculation of predicted values of the observed and missing data based on the regression parameters from the BB

sample. Then, predictive mean matching is performed as described earlier. As with conventional MI, these steps can be carried out $m \geq 1$ times to create m multiply imputed data sets. For this paper, we use the R software program *BaBooN* (Koller-Meinfelder 2011) to implement BBPMM.

A hybrid method: the EM bootstrap

In this section we examine an approach that combines Bayesian imputation concepts with the frequentist idea of bootstrap sampling Efron (1979) along with the use of the EM algorithm. The idea is to extend the EM algorithm using a bootstrap approach but also allowing for the incorporation of priors on model parameters.

Following Honaker and King (2010) and Honaker (personal communcation, June 2011) the first step is to bootstrap the PISA and TALIS concatenated data set to create m versions of the incomplete data, where m ranges typically from 3 to 5 as in other multiple imputation approaches. Bootstrap resampling involves taking a sample of size n with replacement from the original dataset. Here, the m bootstrap samples of size n are obtained from the PISA and TALIS concatenated file, where n is the total sample size of the file. Second, for each bootstapped data set, the EM algorithm is run. It is here that Honaker and King (2010) allow for the inclusion of prior distributions on the model parameters estimated via the EM algorithm.

Notice that because m boostrapped samples are obtained, and that each EM run may contain priors, then once the EM algorithm has run, the model parameters will be different. Indeed, with priors, the final results are the *maximum a posteriori* (MAP) estimates – the Bayesian counterpart of the maximum likelihood estimates. Because we have complete data for Iceland, in this paper we elicit informative priors based on the known distributions of the missing variables. Specifically, we apply a matrix of means and standard deviations from the complete Iceland data. In practice, however, this information would be unknown. Nevertheless, for data fusion problems of this sort, we reasonably assume MCAR and so it may be justified to use the means and standard deviations of the observed variables to obtain priors. Also, findings from matching current cycles of PISA and TALIS could be used to inform the specification of priors for future data fusion exercises.

With priors in hand, missing values are imputed based on the final converged estimates for each of the m datasets. These m versions can then be used in subsequent analyses. This approach is referred to as the EM-Bootstrap *EMB* (Honaker and King 2010) and implemented in the R program *Amelia* (Honaker et al. 2010) which we will use in our analyses below.

Data source

The PISA 2009 survey design samples schools proportional to size followed by a sample of the target student population within those schools[e]. The target student population was based on target age rather than school grade levels to allow for international comparability. The eligible age range at the time of the assessment was between 15 years and 3 months and 16 years and 2 months to ensure that students were assessed before they completed compulsory education. Also, only those who had completed at least 6 years of formal schooling were eligible for the study and those with intellectual disabilities or limited language proficiency in the language of the test were excluded.

PISA collects student-level and school-level data from reports by students, school administrators, and parents across 34 OECD member countries and 41 partner countries and economies[f].

For TALIS, a two-stage stratified probability sample was employed with lower secondary education teachers (level 2 of the 1997 revision of the International Standard Classification of Education, ISCED 97) as second stage units randomly selected from randomly selected schools. The surveys were in the field from October 2007 to May 2008. TALIS provides teacher-level and school level data from reports by teachers and school administrators across 16 member countries and 7 partner countries and economies[g].

Data from Iceland

Recall that we are utilizing data from Iceland because Iceland implemented both the PISA and TALIS surveys and thus provides a unique opportunity to study the validity of the data fusion algorithms. In total, 142 schools participated in either the TALIS survey or the PISA survey. Of these, 122 PISA and TALIS schools were able to be matched. The 20 schools that were unmatched were eligible for TALIS or PISA, but not both. An additional 39 schools were excluded due to large amounts of missing data on variables needed for the matching procedures. Finally, 5 schools were excluded because they were identified to be influential outliers. Thus, the data fusion procedures utilize data from 78 schools in Iceland with full information from the PISA and TALIS data sets.

For our experiment with data from Iceland, preliminary analyses indicated that randomly deleting data would yield a sample size that was likely too small to effectively judge the quality of the matching procedures. To address this problem, we duplicated the Iceland data and then removed PISA data for half the sample and TALIS data for the other half of the sample. This led to a sample of 78 schools with PISA data and 78 schools with TALIS data. Because the duplication and subsequent deletion of the data were not dependent on any of the observed PISA, TALIS or common variables, the missing data are missing completely at random[h].

Variables

PISA administers surveys to school principals and to students. TALIS administers surveys to school principals and to teachers. Common variables are drawn from the school principal surveys from PISA and TALIS. These are the variables that are used in the matching methods to generate the matched data sets.

Table 1 Summary statistics and conditional covariance matrix for original Iceland data

Variable	n	Mean	sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Selfef	156.00	0.30	0.35	0.31	0.31	0.39	-0.58	1.02	1.60	-0.27	-0.44	0.03
Jobsat	156.00	3.13	0.20	3.12	3.13	0.18	2.75	3.57	0.82	0.17	-0.51	0.02
Joyread	156.00	-0.09	0.33	-0.13	-0.09	0.27	-0.93	0.91	1.84	0.28	0.89	0.03
Metasum	156.00	-0.17	0.37	-0.16	-0.18	0.32	-0.96	1.20	2.17	0.44	1.69	0.03
					Joyread	Metasum						
				Selfef	-0.02	-0.04						
				Jobsat	0.01	-0.01						

Table 2 Summary statistics and conditional covariance matrix for Iceland data: hot deck distance matching

Variable	n	Mean	sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Selfef	78	0.30	0.36	0.31	0.31	0.39	-0.58	1.02	1.60	-0.27	-0.41	0.04
Jobsat	78	3.13	0.20	3.12	3.13	0.18	2.75	3.57	0.82	0.16	-0.49	0.02
Joyread	78	-0.05	0.31	-0.09	-0.06	0.30	-0.93	0.91	1.84	0.55	1.38	0.04
Metasum	78	-0.18	0.39	-0.16	-0.18	0.38	-0.96	1.20	2.17	0.33	1.02	0.04
					Joyread	Metasum						
				Selfef	-0.00	0.02						
				Jobsat	-0.01	-0.01						

Matching variables

We were able to match on several indicators and indices that are similar in both the PISA and TALIS school administrator surveys. Both sets of data include information on school sector, the size of the school community, the total enrollment in the school, a measure of the availability of school material resources, the extent to which teacher absenteeism interferes with student learning, a measure of the extent to which student-related factors affect the school climate, and a measure of the disciplinary climate of the school[i].

There may be other variables in the student or parent surveys from PISA, or the teacher surveys from TALIS that can be used for the matching procedure. These would need to be standardized and averaged to the school level prior to applying the matching procedures. Including more variables for the match is generally better, although increasing the variables included in the matching procedure necessitates a larger school sample in both PISA and TALIS. Also, in certain contexts, a reduced set of variables may be used depending on their usefulness for the data fusion procedure. For example, in Iceland there are very few private schools. Because of the lack of variation in the school sector variable, it is not useful for the match.

Another consideration for matching is to match within meaningful subpopulations. Researchers may wish to match within private schools and within public schools, for example. This would be a useful strategy if schools within sub-populations differ greatly from each other. Sub-populations could be defined within school sector, regions, governance structures, etc. We did not do this for the current analysis because the private schools were dropped from the sample for reasons unrelated to their school sector designation.

Table 3 Summary statistics and conditional covariance matrix for Iceland data: stochastic regression imputation

Variable	n	Mean	sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Selfef	78	0.30	0.37	0.31	0.31	0.39	-0.93	1.42	2.35	-0.13	0.03	0.01
Jobsat	78	3.12	0.20	3.11	3.12	0.18	2.41	3.82	1.40	-0.02	-0.01	0.01
Joyread	78	-0.05	0.34	-0.08	-0.06	0.31	-0.94	1.26	2.20	0.25	0.53	0.01
Metasum	78	-0.12	0.38	-0.14	-0.13	0.36	-1.18	1.20	2.38	0.30	0.57	0.01
					Joyread	Metasum						
				Selfef	-0.08	0.02						
				Jobsat	-0.03	0.03						

Table 4 Summary statistics and conditional covariance matrix for Iceland data: predictive mean matching

Variable	n	Mean	sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
selfef	78	0.29	0.35	0.30	0.30	0.39	-0.58	1.02	1.60	-0.27	-0.48	0.01
jobsat	78	3.12	0.21	3.11	3.11	0.19	2.75	3.57	0.82	0.13	-0.61	0.01
joyread	78	-0.06	0.32	-0.11	-0.06	0.27	-0.93	0.91	1.84	0.24	0.73	0.01
metasum	78	-0.13	0.38	-0.11	-0.14	0.35	-0.96	1.20	2.17	0.54	1.74	0.01

	Joyread	Metasum
Selfef	-0.01	0.00
Jobsat	0.00	-0.03

Unique variables

The central focus of PISA 2009 was proficiency in reading. PISA identified a cumulative or cyclical model of how engagement in reading activities (e.g., enjoyment of reading and diversity of reading materials) and approaches to learning (e.g., summarizing skills and memorization strategies) promotes reading performance at the end of compulsory education (OECD, 2010b, pg. 25). These skills are of interest to researchers studying inequality because they have been shown to mediate the effects of socioeconomic advantage on reading achievement (OECD, 2010b, pg. 91). To measure students' engagement in reading and learning strategies, we chose one indicator of each: "enjoyment of reading" (joyread) and "summarizing skills "(metasum). According to analysis of PISA 2009, 18% of the student variation in reading performance across OECD countries can be explained by variation in students' enjoyment of reading (OECD, 2010b pg. 28) (22% ISL). Also, 21% of the variation in reading performance across OECD countries can be explained by variation in summarizing skills (OECD 2010b, pg. 47) (20% ISL). Both measures are averaged to the school level for analysis.

We chose two predictor variables of interest that measure teachers' job-related attitudes: "teacher job satisfaction" (jobsat) and "teacher self-efficacy" (selfef). Job satisfaction influences aspects of teachers' behavior such as performance, absenteeism, and turnover (OECD, 2009, p.111). Similarly, teachers' self-efficacy influences their instructional standards and coping strategies (OECD, 2009, p. 111). Both job satisfaction and teacher self-efficacy are linked to instructional practices and student achievement (Ashton and Webb, 1986; Ross, 1998).The job satisfaction measure is taken from one item in the TALIS teacher survey which asks the teachers to indicate how strongly they agree with the statement "All in all, I am satisfied with my job." The self-efficacy measure is a composite of four items in the teacher survey. Teachers are asked to indicate how strongly

Table 5 Summary statistics and conditional covariance matrix for Iceland data: Bayesian regression imputation

Variable	n	Mean	sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Selfef	78	0.30	0.39	0.31	0.30	0.39	-1.07	1.92	2.99	-0.12	-0.02	0.01
Jobsat	78	3.13	0.22	3.11	3.12	0.20	2.47	3.88	1.41	0.12	0.07	0.01
Joyread	78	-0.04	0.35	-0.06	-0.04	0.33	-1.39	1.42	2.82	0.19	0.78	0.01
Metasum	78	-0.10	0.41	-0.11	-0.11	0.38	-1.14	1.29	2.43	0.33	0.47	0.01

	Joyread	Metasum
Selfef	-0.03	-0.10
Jobsat	0.01	0.01

Table 6 Summary statistics and conditional covariance matrix for Iceland data: Bayesian bootstrap predictive mean matching

Variable	n	Mean	sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Selfef	78	0.27	0.35	0.30	0.28	0.41	-0.58	1.02	1.60	-0.18	-0.52	0.01
Jobsat	78	3.12	0.19	3.11	3.12	0.16	2.75	3.57	0.82	0.22	-0.51	0.01
Joyread	78	-0.08	0.31	-0.12	-0.08	0.26	-0.93	0.91	1.84	0.23	0.63	0.01
Metasum	78	-0.15	0.36	-0.14	-0.15	0.33	-0.96	1.20	2.17	0.31	1.22	0.01

	Joyread	Metasum
Selfef	0.01	0.01
Jobsat	0.01	0.02

they agree with the statements: "I feel that I am making a significant educational difference in the lives of my students", "If I try really hard, I can make progress with even the most difficult and unmotivated students", "I am successful with the students in my class", and "I usually know how to get through to students." Both the job satisfaction measure and the teacher self-efficacy measure are averaged to the school level for analysis.

Results

Software code for the data fusion methods is presented in Additional file 1: Annex A and software code for implementing the validity checks is given in Additional file 1: Annex B for the hot deck matching method only. Validity checking for the other methods would be implemented in the same way.

An inspection of Table 1 shows the descriptive statistics for the Iceland data for the original data and Tables 2, 3, 4, 5, 6, 7 and 8 show the results for each data fusion algorithm. A complete set of descriptive statistics are provided including the mean, standard deviation, median, trimmed mean, mean absolute deviation, minimum, maximum, range, skewness, kurtosis, and standard error of the mean. A visual comparison of the results suggests that most of the methods do a reasonably good job of reproducing marginal descriptive values. Exceptions include stochastic regression imputation and Bayesian regression imputation using chained equations. Hot deck matching does a reasonable job except with respect to skewness and kurtosis estimates.

An inspection of Tables 1, 2, 3, 4, 5, 6 7 and 8 also present an assessment of third level validity – namely the preservation of the correlation/covariance structure of the data. Recall, that preservation of the correlation/covariance structure requires that the

Table 7 Summary statistics and conditional covariance matrix for Iceland data: EM bootstrap

Variable	n	Mean	sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Selfef	78	0.32	0.36	0.34	0.33	0.36	-1.11	1.38	2.49	-0.26	0.06	0.01
Jobsat	78	3.12	0.20	3.11	3.11	0.17	2.38	3.70	1.31	0.00	-0.04	0.01
Joyread	78	-0.03	0.33	-0.05	-0.03	0.31	-1.02	0.99	2.01	0.07	0.32	0.01
Metasum	78	-0.11	0.37	-0.11	-0.12	0.37	-1.06	1.20	2.26	0.20	0.45	0.01

	Joyread	Metasum
Selfef	0.01	0.00
Jobsat	-0.00	-0.02

Table 8 Summary statistics and conditional covariance matrix for Iceland data: EM bootstrap with priors

Variable	n	Mean	sd	Median	Trimmed	Mad	Min	Max	Range	Skew	Kurtosis	se
Selfef	78	0.30	0.33	0.30	0.31	0.37	-0.65	1.39	2.03	-0.19	-0.27	0.01
Jobsat	78	3.13	0.19	3.12	3.12	0.18	2.63	3.72	1.09	0.14	-0.24	0.01
Joyread	78	-0.06	0.30	-0.08	-0.07	0.27	-0.93	0.91	1.84	0.18	0.65	0.01
Metasum	78	-0.13	0.35	-0.13	-0.13	0.34	-0.97	1.20	2.18	0.19	0.91	0.01

	Joyread	Metasum
Selfef	-0.00	0.00
Jobsat	0.02	0.01

conditional correlations among the unique variables given the matching variable should be close to zero. As an example, inspection of Table 2 for hot deck matching reveals that the conditional correlations are very small and not greater than 0.02. When compared to the values in Table 1, we see that hot deck matching does an excellent job of preserving correlation/covariance structure of the data. Overall, the results indicate that while most methods do a reasonably good job of meeting third level validity, BBPMM and the EM bootstrap stand out as being the best methods in terms of this validity criteria.

Figures 1, 2, 3, 4, 5, 6 and 7 provide a visual inspection of the descriptive statistics results presented above. Specifically, the *kernel density* plots represent smoothed histograms. We compare the distribution of the synthetic data (solid line) against the original data (dotted lines). We find that most all procedures yield a kernel density plot that matches the distribution of the original variables quite well. In addition, we also present *quantile-quantile* (Q-Q) plots in Figures 8, 9, 10, 11, 12, 13 and 14. A Q-Q plot is a graphical approach for

Figure 1 Kernel density plots for hot deck matching.

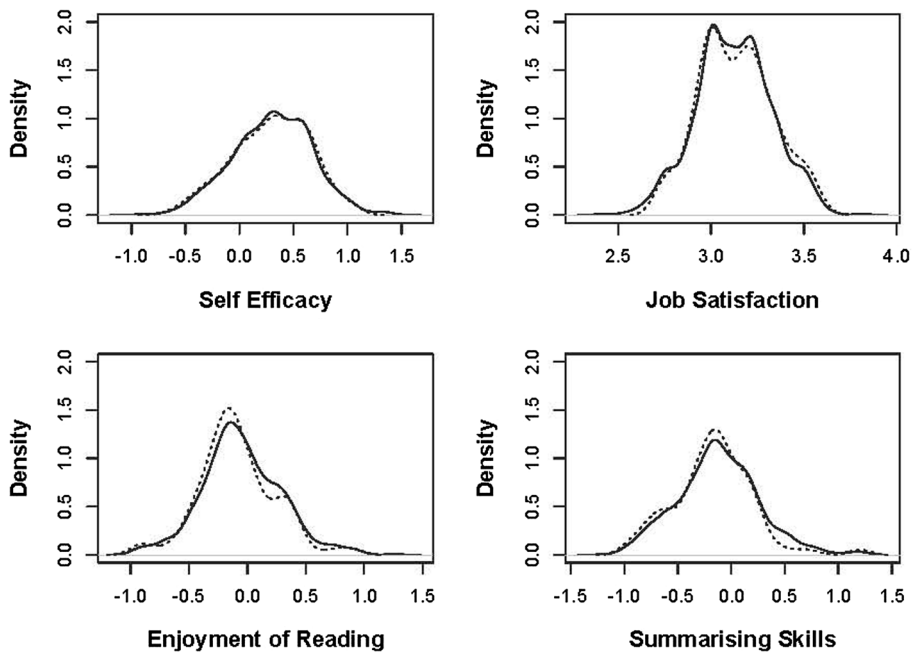

Figure 2 Kernel density plots for stochastic regression.

comparing two probability distributions by plotting their quantiles against each other. If the two probability distributions being compared are similar, the points in the Q-Q plot will lie approximately on a straight line. A close inspection reveals that BBPMM provides the best quantile-quantile plots overall, and particularly better than the EM bootstrap method.

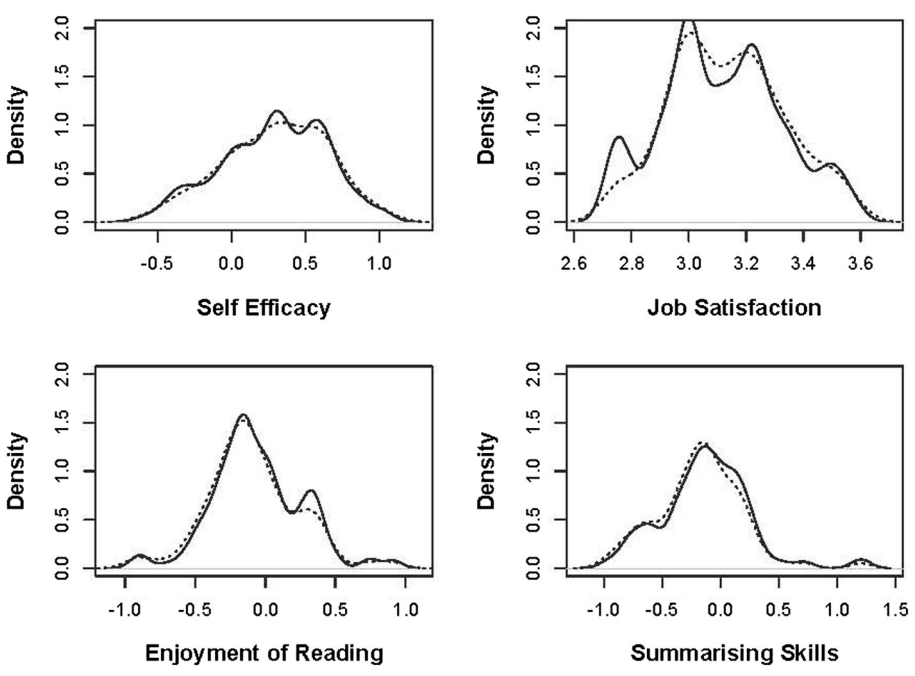

Figure 3 Kernel density plots for predictive mean matching.

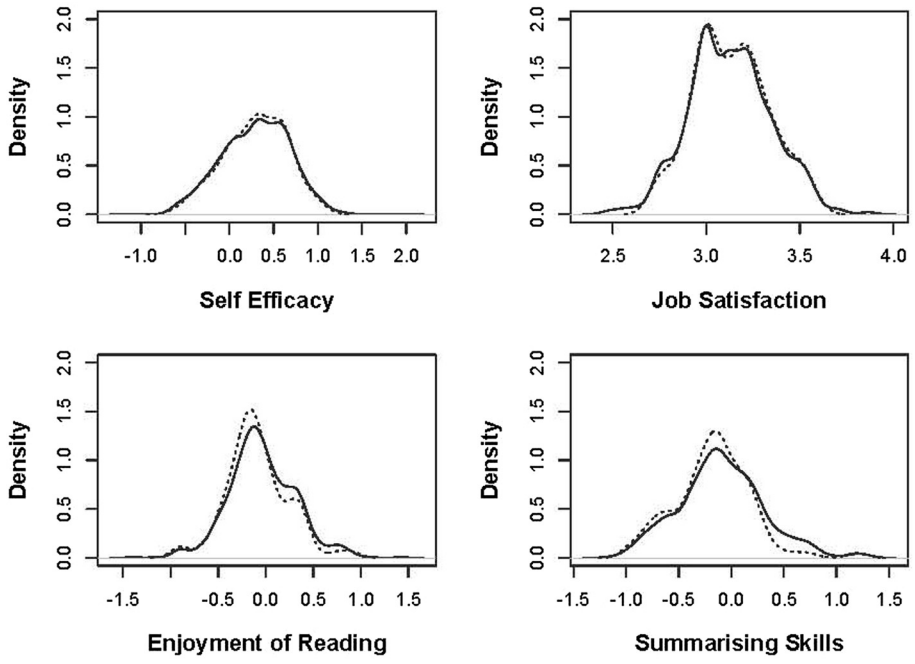

Figure 4 Kernel density plots for Bayesian regression.

Discussion

The purpose of this paper was to provide a *proof of concept* on how one might implement a statistical match of PISA and TALIS. We argued at the beginning of the paper that statistically matching PISA and TALIS might be a reasonable option for countries that are unable to administer both surveys to the same sample schools. Our

Figure 5 Kernel density plots for Bayesian bootstrap predictive mean matching.

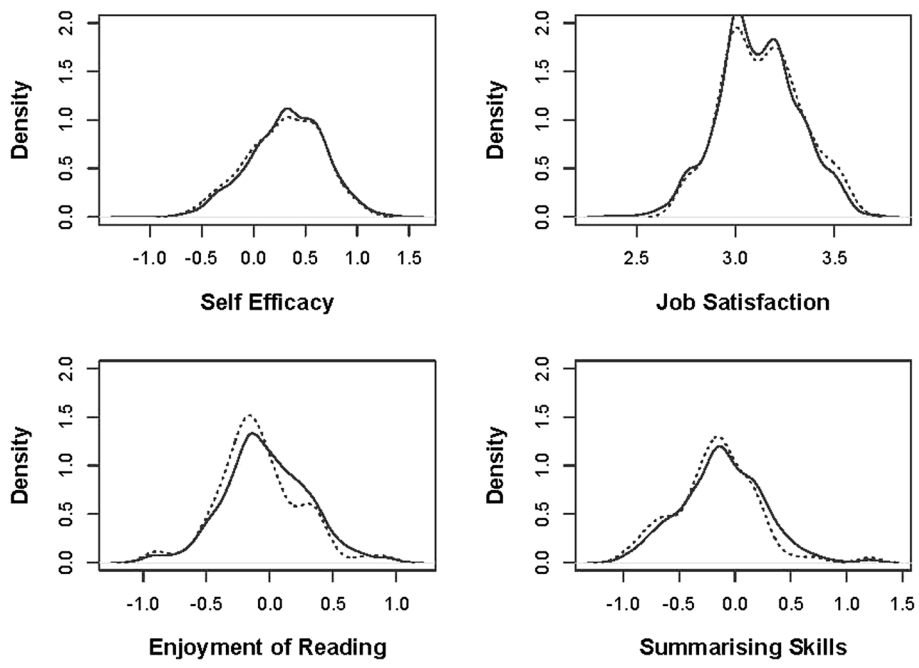

Figure 6 Kernel density plots for the EM bootstrap.

analyses suggest that statistically matching PISA and TALIS is feasible and should be seriously considered by countries interested in gleaning added value from both surveys.

Among the methodologies that were considered in this paper, two stand out as deserving serious consideration for matching PISA and TALIS – Bayesian

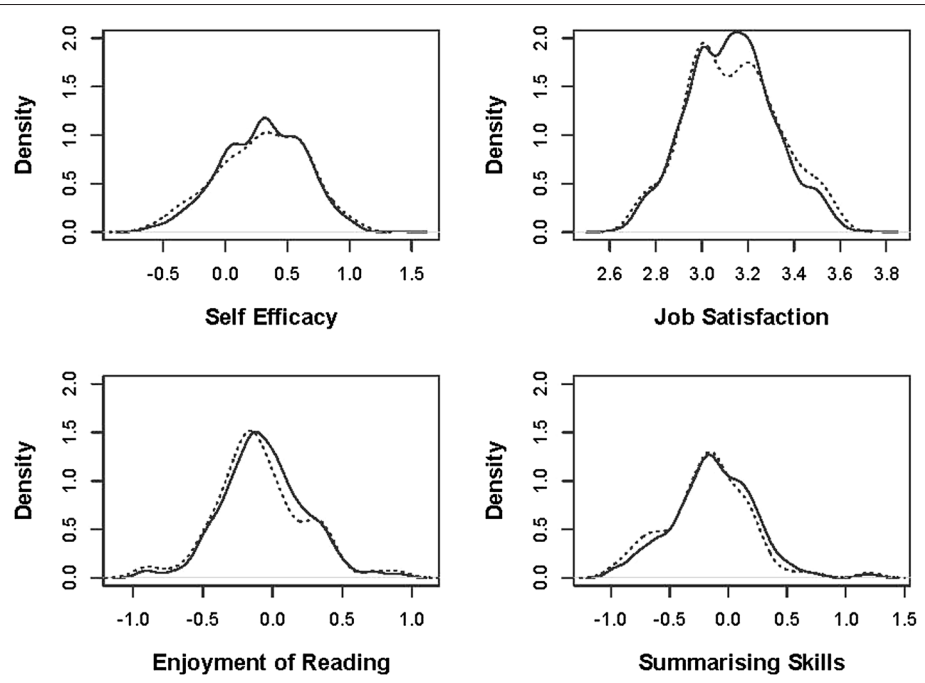

Figure 7 Kernel density plots for the EM bootstrap with priors.

Figure 8 QQ plots for hot deck matching.

bootstrap predictive mean matching, and the EM-bootstrap. Both methodologies worked quite well with respect to Rassler's (2002) third and fourth level validity criteria.

Our simulation study made use of data from Iceland. Because Iceland implemented PISA and TALIS on all relevant students and schools, we were able to evaluate the ability

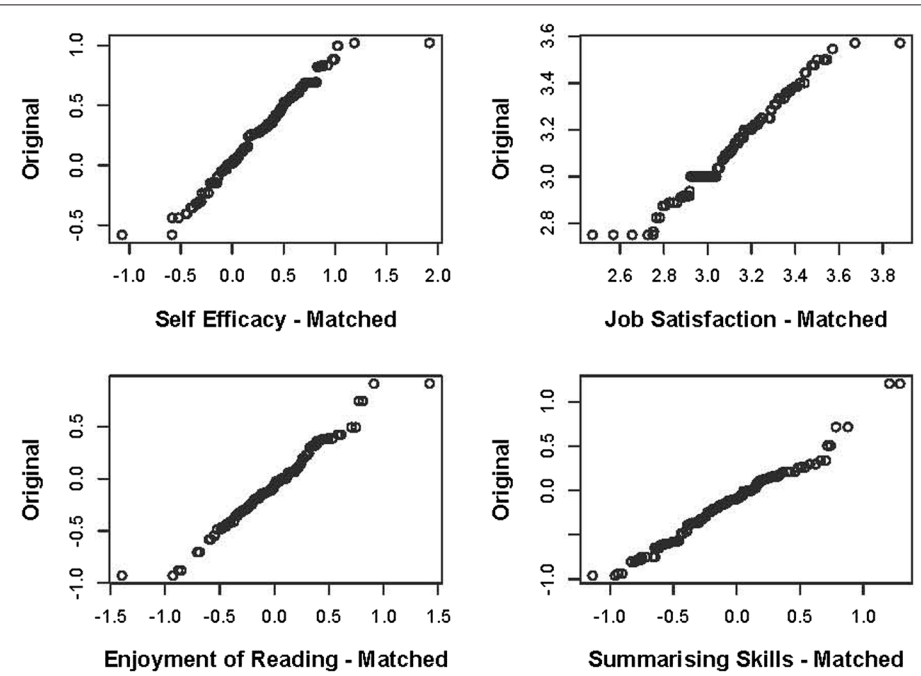

Figure 9 QQ plots for stochastic regression.

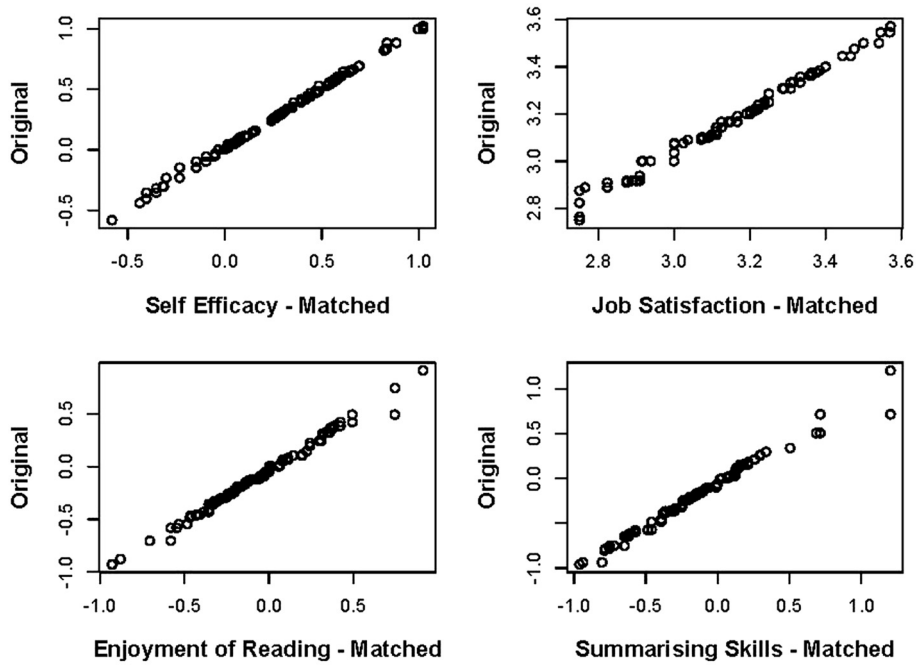

Figure 10 QQ plots for predictive mean matching.

of alternative methods in reproducing true correlations. In practice, however, the true estimates of the correlation structure would be unknown. In this case, researchers would need to make use of sensitivity analyses, the details of which are outlined in Rubin (1986). A study of sensitivity to violations of assumptions was beyond the scope and purpose of this paper.

Figure 11 QQ plots for Bayesian regression.

Figure 12 QQ plots for Bayesian bootstrap predictive mean matching.

As noted earlier, data fusion is typically limited to single level data structures. In the case of PISA and TALIS, this requires aggregation of student and teacher level data to the school level, respectively. Thus, the well known problems associated with data aggregation are present in the statistically matched file. However, there does exist a two-level data fusion algorithm in the software program *mice* (van Buuren

Figure 13 QQ plots for the EM bootstrap.

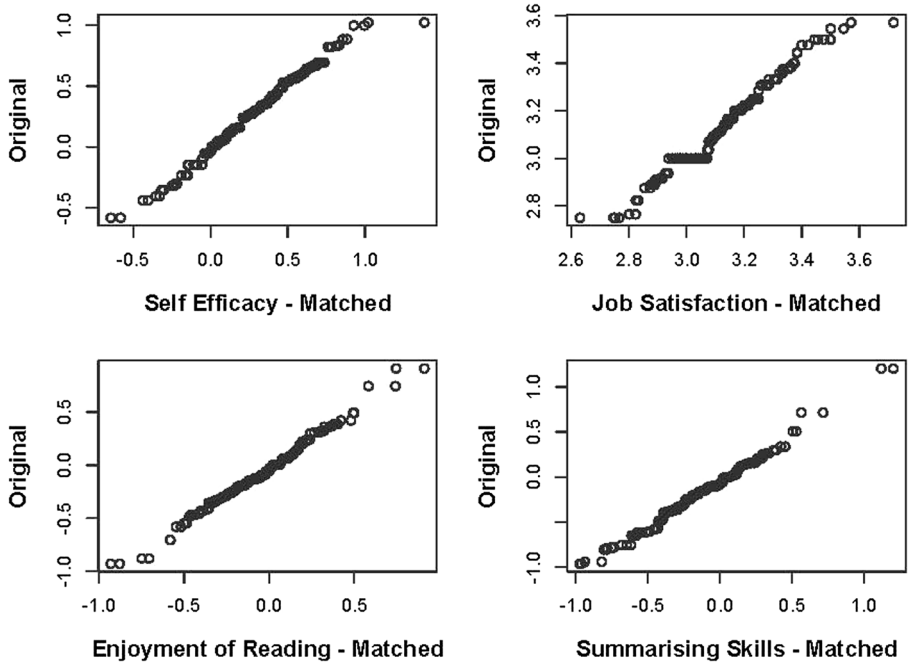

Figure 14 QQ plots for the EM bootstrap with priors.

and Groothuis-Oudshoorn 2010) based on the Gibbs sampling algorithm. For those PISA participating countries that opted for the international teacher questionnaire, and assuming that there are teacher level variables common to TALIS and the PISA teacher questionnaires, two level data fusion may be feasible and certainly worth exploring.

In the context of cross-national education research, data fusion within countries allows for a more nuanced analysis of cross-national differences. Recall that while both PISA and TALIS allow researchers to link institutional characteristics to aspects of school and classroom climate, only PISA offers measures of student learning, and only TALIS provides information about teachersï£¡ job-related attitudes. In order to fully understand cross-national differences in outcomes, it is necessary to provide a complete description of the inputs and processes that relate to differences in outcomes across countries. In all, 24 countries participated in the TALIS survey, and each of these has also participated in PISA 2009. Matching the TALIS and PISA surveys for each of these 24 countries is beyond the scope of the current study, however the potential for data fusion to provide complete information on multiple countries is promising. For example, PISA data suggest that the best performing education systems prioritize teacher and administration quality, provide clear and ambitious standards focused on complex, higher order thinking, and embrace the diversity in students capacities, interests, and social background through individualized approaches to learning (OECD 2010a). TALIS data suggest that professional development, teaching practices, teachers beliefs and attitudes, school and teacher evaluation methods are important for understanding and improving educational processes (OECD 2009). In the absence of a new design that formally links through the administration of PISA and TALIS jointly, data fusion provides the next best approach for addressing these important policy questions.

Conclusions

To conclude, this paper demonstrated the feasibility of statistically matching PISA and TALIS, as well as demonstrated the effectiveness of six algorithms that could be employed for this purpose. The feasibility of statistically matching PISA and TALIS is supplemented by the accessibility of free and open source software - specifically, software packages found within the R statistical computing environment (R Development Core Team 2010). In the absence of a direct implementation of both surveys, national project managers for PISA and TALIS may wish to invest in analytic and software training on methods of data fusion.

Endnotes

[a]Rather than drawing a sample of targeted students and teachers, Iceland surveyed the entire population, which for PISA included all 15 year old students and and for TALIS included all teachers who teach ISCED level 2 students (usually aged between 11 and 16). Our study includes schools where these two populations overlapped, that is schools serving PISA students taught by TALIS teachers.

[b]Of course, within a survey, missing data on some variables, including those that are common across PISA and TALIS might be MAR or NMAR. We will assume that missing data on variables in common to both PISA or TALIS are at least MAR.

[c]However, see Reiter (2012) for a recent discussion of bias in variance estimates based on Rubin's combining rules and a Bayesian alternative.

[d]Another popular form of Bayesianly proper imputation involves the data augmentation algorithm of Tanner and Wong (1987).

[e]There are additional complexities to the sampling designs of PISA and TALIS that can be found in their respective technical reports.

[f]To be included in the PISA study, a minimum of 150 schools must participate in the surveys.

[g]To be included in the TALIS study, a minimum of 200 schools must participate in the surveys.

[h] Doubling the sample is not recommended in general. The results of the match will be artificially improved as a result of the duplication. However, for the purposes of our experiment, which compares data fusion strategies using the same duplicated data, we do not expect this strategy to influence our recommendations for the preferred matching method. This is because each method is equally subject to the artificial improvements risked by the data duplication.

[i]Information about the average disciplinary climate for each school was drawn from student surveys in PISA and the teacher surveys in TALIS, and are averaged to the school level. It has been shown that there is a high level of agreement on indicators of disciplinary climates among teachers and students (OECD 2009 pg. 104), so we argue that these variables are suitable to use as matching variables.

Additional file

Additional file 1: Annex A. # R Scripts for PISA-TALIS Match. **Annex B.** ## Script for calculating marginal distributions and conditional covariance matrix. ## Needed to check third and fourth order validity.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This research was supported by contract # EDU/JA00066381 from the Directorate for Education, Organization for Economic Cooperation and Development. An earlier version of this research was published as an OECD working paper

and can be found at http://www.oecd-ilibrary.org/education/statistical-matching-of-pisa-2009-and-talis-2008-data-in-iceland_5k97g3zzvg30-en. The comments and suggestions in this paper are solely those of the authors and do not necessarily reflect the views of the OECD.
The authors are grateful to Suzanne Rässler for valuable comments on an earlier draft of this paper, and to Stef van Buuren for addressing questions regarding the *mice* program.

Author details
[1]Department of Educational Psychology, University of Wisconsin, Madison, USA. [2]Department of Sociology, University of Wisconsin, Madison, USA.

Received: 29 August 2013 Accepted: 3 September 2013
Published: 16 September 2013

References

Ashton, P, & Webb, N (1986). *Making a Difference: Teacher Efficacy and Student Achievement*. Monogram, Longman, White Plains, New York.
D'Orazio, M (2011). Statmatch: Statistical matching [Computer software manual]. Available from http://CRAN.R-project.org/package=StatMatch (R package version 1.0.1).
D'Orazio, M, Di Zio, M, Scanu, M (2006). *Statistical matching: theory and practice*. New York: Wiley.
Efron, B (1979). Bootstrap methods: another look at the jacknife. *The Annals of Statistics*, *7*, 1–26.
Enders, CK (2010). *Applied missing data analysis*. New York: Guilford.
Geman, S, & Geman, D (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*, 721–741.
Honaker, J, & King, G (2010). What to do about missing values in time-series cross-section data. *American Journal of Political Science*, *54*, 561–581.
Honaker, J, King, G, Blackwell, M (2010). Amelia II: a program for missing data [Computer software manual]. Available from http://CRAN.R-project.org/package=Amelia (R package version 1.2-18).
Kaplan, D (1995). The impact of BIB spiraling-induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Statistics*, *20*, 69–82.
Kaplan, D, & Depaoli, S (2013). Bayesian statistical methods. In TD Little (Ed.), *Oxford Handbook of Quantitative Methods* (pp. 407–437). Oxford: Oxford University Press.
Koller-Meinfelder, F (2011). BaBooN: Bayesian bootstrap predictive mean matching – multiple and single imputation for discrete data [Computer software manual]. Available from http://CRAN.R-project.org/package=BaBooN (R package version 2.14.0).
Little, RJA, & Rubin, DB (2002). *Statistical analysis with missing data (2nd. ed.)* New York: Wiley.
OECD (2009). *Creating effective teaching and learning results: First results from TALIS*. Paris: OECD.
OECD (2010a). *Pisa 2009 results: executive summary*. Paris: OECD.
OECD (2010b). *Pisa 2009 results: Learning to learn – student engagement, strategies and practices (Vol. 3)*. Paris: OECD.
R Development Core Team (2010). R: a language and environment for statistical computing [Computer software manual]. Vienna Austria. Available from http://www.R-project.org (ISBN 3-900051-07-0).
Rässler, S (2002). *Statistical matching: a frequentist theory, practical applications, and alternative Bayesian approaches*. New York: Springer.
Reiter, JP (2012). Bayesian finite population imputation for data fusion. *Statistica Sinica*, *22*, 795–811.
Ross, JA (1998). The Antecedents and Consequences of Teacher Efficacy. In Brophy J (Ed.), *Advances in Research on Teaching*, Vol. 7, 49–74. JAI Press. Greenwich, Connecticut.
Rubin, DB (1976). Inference and missing data. *Biometrika*, *63*, 581–592.
Rubin, D B (1981). The Bayesian bootstrap. *The Annals of Statistics*, *9*, 130–134.
Rubin, D B (1986). Statistical matching using file concatenation with adjusted weights and multiple imputation. *Journal of Business and Economic Statistics*, *4*, 87–95.
Rubin, D B (1987). *Multiple imputation in nonresponse surveys*. Hoboken: Wiley.
Schafer, JL (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.
Tanner, MH, & Wong, WA (1987). The calculation of posterior distributions by PISA/TALIS Link 34 data augmentation (with discussion). *Journal of the American Statistical Association*, *82*, 528–550.
van Buuren, S, & Groothuis-Oudshoorn, K (2010). Multivariate imputation by chained equations, version, *2*, 3. http://www.multiple-imputation.com/.
van Buuren, S (2012). *Flexible imputation of missing data*. New York: Chapman & Hall.
Van de Werfhorst, H, & Mijs, JM (2010). Achievement inequality and the institutional structure of educational systems: a comparative perspective. *Annual Review of Sociology*, *36*, 407–428.

doi:10.1186/2196-0739-1-6
Cite this article as: Kaplan and McCarty: **Data fusion with international large scale assessments: a case study using the OECD PISA and TALIS surveys.** *Large-scale Assessments in Education* 2013 **1**:6.