

Please cite this paper as:

Kaplan, D. and A. Turner (2012), "Statistical Matching of PISA 2009 and TALIS 2008 Data in Iceland", *OECD Education Working Papers*, No. 78, OECD Publishing. <http://dx.doi.org/10.1787/5k97g3zzvg30-en>



OECD Education Working Papers
No. 78

Statistical Matching of PISA 2009 and TALIS 2008 Data in Iceland

David Kaplan, Alyn Turner

DIRECTORATE FOR EDUCATION

EDU/WKP(2012)15
Unclassified

Statistical Matching of PISA 2009 and TALIS 2008 data in Iceland

Directorate for Education Working Paper N°78
By David Kaplan and Alyn Turner

June 2012

This Working Paper was prepared as part of the Teaching and Learning International Survey (TALIS) programme. It presents a systematic evaluation of a set of statistical matching methods focused on the goal of creating a synthetic file of PISA 2009 and TALIS 2008 data for Iceland. The paper evaluates the extent to which each method provides a matched data set that maintains the essential properties of PISA and TALIS, concentrating on a set of validity criteria established by Rässler (2002).

Julie BÉLANGER, +33 (0) 1 45 24 91 93, julie.belanger@oecd.org

JT03323540

Complete document available on OLIS in its original format

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

OECD DIRECTORATE FOR EDUCATION

OECD EDUCATION WORKING PAPERS SERIES

This series is designed to make available to a wider readership selected studies drawing on the work of the OECD Directorate for Education. Authorship is usually collective, but principal writers are named. The papers are generally available only in their original language (English or French) with a short summary available in the other.

This document and any map included herein are without prejudice to the status of or sovereignty over any territory, to the delimitation of international frontiers and boundaries and to the name of any territory, city or area.

The opinions expressed in these papers are the sole responsibility of the author(s) and do not necessarily reflect those of the OECD or of the governments of its member countries.

Applications for permission to reproduce or translate all, or part of, this material should be sent to rights@oecd.org.

Comment on the series is welcome, and should be sent to edu.contact@oecd.org.

www.oecd.org/edu/workingpapers

Copyright OECD 2012.

ABSTRACT

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 policy-makers opportunities to identify particular educational institutional arrangements – that is, how aspects of educational systems are organised 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 relations between the inputs, processes, and outcomes of education. 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. This limitation can be partly addressed by statistically linking both surveys. This involves the creation of a synthetic cohort of data – that is, a new data file that combines information from both surveys. This paper presents a systematic evaluation of a set of statistical matching methods focused on the goal of creating a synthetic file of PISA 2009 and TALIS 2008 data for Iceland. We evaluate the extent to which each method provides a matched data set that maintains the essential properties of PISA and TALIS, concentrating on a set of validity criteria established by Rässler (2002). The experimental study provides a proof of concept that statistically matching PISA and TALIS is feasible for countries that wish to draw on the added value of both surveys for research and policy analysis.

Le Programme international pour le suivi des acquis des élèves (PISA) et l'Enquête internationale sur l'enseignement et l'apprentissage (TALIS) de l'OCDE constituent deux des plus grande enquêtes internationales auprès des étudiants et des enseignants. Les données générées par ces enquêtes permettent aux chercheurs et aux décideurs en matière de politiques d'identifier les systèmes éducatifs sont organisés afin de promouvoir l'égalité des chances dans l'enseignement tant à l'intérieur des pays qu'entre les pays. Évidemment, les décideurs sont intéressés à tous les trois niveaux du système scolaire, c'est-à-dire aux élèves, aux enseignants, et aux écoles, afin de comprendre pleinement les différences dans les relations entre les intrants, les processus, et les résultats de l'éducation qui sont observées dans les pays et entre les pays. Une limitation importante associée à ces enquêtes relève du fait qu'un élément important du système éducatif manque à chacune d'elles - à savoir, pour PISA l'aspect enseignant, et pour TALIS la perspective des élèves. Cette limitation peut être partiellement résolue en reliant statistiquement les deux enquêtes. Cela implique la création d'une cohorte synthétique des données - un nouveau fichier de données qui combine les informations provenant des deux enquêtes. Cet article présente une évaluation systématique d'un ensemble de méthodes statistiques qui ont pour but de créer un fichier de synthèse des enquêtes PISA 2009 et TALIS 2008 pour l'Islande. L'étude évalue la mesure dans laquelle chaque méthode fournit un ensemble de données qui maintient les propriétés psychométriques essentielles des enquêtes PISA et TALIS, en se concentrant sur un ensemble de critères de validité établis par Rässler (2002). Cette étude fournit la preuve de concept qu'il est possible de relier statistiquement PISA et TALIS pour les pays qui souhaitent s'appuyer sur la valeur ajoutée de chacune des enquêtes pour la recherche et l'analyse des politiques éducatives.

STATISTICAL MATCHING OF PISA 2009 AND TALIS 2008 DATA IN ICELAND

David Kaplan, Department of Educational Psychology

Alyn Turner, Department of Sociology

University of Wisconsin – Madison

Statement of Problem

1. Equality of educational opportunity varies within and between schools (Jencks & Tach, 2006). In other words, schools do not unequivocally provide every student, regardless of family background, an equal chance to achieve at the level of his or her potential. Research from the United States suggests that differences in teaching, learning processes, and the allocation and use of resources have important effects on the level of equality of opportunity for individual students (Barr & Dreeben, 1983; Hanushek & Lindseth, 2009; Gamoran & Dreeben, 1986; Gamoran, Secada, & Marrett, 2000).
2. Wide variation in students' educational outcomes also exists across countries. The gap between the highest and lowest performing OECD countries, for example, is the equivalent of about two years of schooling. Moreover, the gap between the lowest and highest performing OECD and non-OECD countries and economies is the equivalent of six years of schooling. These gaps remain after taking into account differences in national income; GDP explains about 6% of the differences in average student performance (OECD, 2009).
3. 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 policy-makers opportunities to identify particular educational institutional arrangements – that is, how aspects of educational systems are organised 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 relations between the inputs, processes, and outcomes of education. 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. One desirable approach to linking the PISA survey to the TALIS survey is to sample 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.
4. A more feasible 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 in terms of 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 report, we will use the generic term statistical matching with the understanding that some procedures involve imputation of missing data.

5. The current study is a systematic evaluation of a set of statistical matching methods focused on the goal of creating a synthetic file of PISA 2009 and TALIS 2008 data. We evaluate the extent to which each method provides a matched data set that maintains the essential properties of PISA and TALIS, concentrating on a set of validity criteria established by Rässler (2002) and described below. Our evaluation relies on an experimental comparison of the validity of each method relative to a standard. For this purpose, we use data from Iceland. We chose Iceland because it is the only OECD country that implemented PISA 2009 and TALIS 2008 on the population of PISA students, all TALIS teachers, and all PISA and TALIS schools. The experimental study will provide a proof of concept that statistically matching PISA and TALIS is feasible for countries that wish to draw on the added value of both surveys for research and policy analysis.

6. The organisation of this report is as follows. In the next section, we outline the problem of statistical matching with particular focus on validity criteria that can be used to evaluate the quality of statistical matching. Next, we outline the methods to be examined in this paper. It should be noted that a large number of methods exist for statistical matching. We will examine six methods that are representative of the broad array of statistical matching methods available, including non-parametric parametric statistical matching 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 statistical matching of PISA and TALIS. Next we will present the design of our study. The results will follow. The report closes with recommendations and limitations resulting from statistically matching PISA and TALIS. Annotated software code is made available in Annex A and Annex B.

The Policy Context

7. 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. The application of statistical matching is certainly not limited to this particular question.

8. PISA obtains samples of students across more than 60 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 utilising 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 standardised educational system have lower levels of inequality of opportunity compared to those with unstandardised systems (Werfhorst & Mijs, 2010).

9. 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 individualised approaches to learning. These education systems also provide clear and ambitious standards focused on complex, higher order thinking, and prioritise teacher and administration quality (OECD, 2010b).

10. 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.

11. 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 difficult, for practical and/or political reasons, to design and implement a large international survey with data gathered from students, teachers and school leaders. Ideally, then, the goal 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, school staff, and school administrators. Statistically combining two relatively distinct data sources is the goal of statistical matching.

Background on Statistical Matching

12. Statistical matching 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 sample representative of some population of relevance to the original surveys. It is convenient to categorise statistical matching methods as non-parametric (*i.e.* those not based on an underlying model for the observed and missing data), semi-parametric (combining non-parametric and parametric methods) or fully 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.

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

14. Following the seminal work Rubin (1976; see also Little & Rubin, 2002; Schafer, 1997; Enders, 2010) 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, say age and income, and assume that there is missing data on income. If the missing data on income is unrelated to the observed values of both age and income, then the

missing data are considered to be missing completely at random or MCAR. 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 income is unrelated to observed income, but may be related to observed age. For example, perhaps older individuals do not report their incomes. This type of missing data is referred to as missing at random or MAR. Under MAR, inferences will be valid, and there now exist many methods for handling missing data under the assumption of MAR.

15. 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 age and income, we may find that missing data on income is related to income. That is, perhaps individuals with higher incomes do not report their incomes, irrespective of their age. This type of missing data problem is referred to as not missing at random or NMAR. 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.

16. 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 is assessment plans that involved balanced incomplete block spiraling frameworks (see *e.g.* Kaplan, 1995) such as the design for the cognitive outcome assessments in PISA. Another example of concern to this report is the case of statistically matching 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.¹

Levels of Validity in Statistical Matching

17. 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 statistical match. This is of prime importance to our goal of matching 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 statistical matching can be found in (Rässler, 2002). Following Rässler, (2002) four levels of validity can be distinguished when considering the problem of statistical matching.

First Level Validity: Preserving Individual Values

18. The most difficult level of validity that can be achieved in statistical matching concerns the ability of the matching procedure 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, generally, it is usually impossible or at least unnecessary to achieve this level of validity. First, the algorithms that we will be examining are designed to reproduce expected values under a given model, and not individual values. Second, imputation algorithms are designed to produce a dataset that can be used for secondary analyses

¹ 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 and TALIS are at least MAR.

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

Second Level Validity: Preserving Joint Distributions

19. The idea behind second level validity is that the joint distribution of all of the variables in the synthetic data set be preserved after statistical matching. 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 only hold if the variables unique to PISA and unique to TALIS are conditionally independent given the variables common to both surveys.

Third Level Validity: Preserving Covariance/Correlation Structures

20. 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 method 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 of 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 $\widehat{cov}(x, y)$ 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 given the common variables z used in the match.²

Fourth Level Validity: Preserving Marginal Distributions

21. The lowest level of validity and a minimum requirement for statistical matching is that the marginal distributions of the individual variables in the original surveys be preserved after the statistical match. Formally, if \hat{f}_y is the marginal distribution of the PISA variables and $\hat{f}_{y,z}$ is the 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. This report provides tables and figures to assess fourth level validity.

Methodology

22. In this section, we describe statistical matching methods we evaluate in the context of the PISA-TALIS match. As noted earlier, there are scores of different methods that can be used for statistical matching, and it is beyond the scope of this report to evaluate every approach that is currently available. Our approach for this report, 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 statistical match. For our experimental study with Iceland, we 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.

23. A common feature of all statistical matching 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

² To see this, let $\widehat{cov}(x, y) = cov[E(x|z), E(y|z)]$ be the synthetic covariances. Then, because $cov(x, y) = E[cov(x, y|z)] + cov[E(x|z), E(y|z)]$, the only way for $\widehat{cov}(x, y) = cov(x, y)$ is if the $E(cov(x, y|z)) = 0$. This report provides an assessment of third level validity.

must be aggregated to their respective school level before statistical matching can proceed. In doing so, the multilevel structure of each survey is lost. Statistical matching, 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 the match. 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.

24. We organise this section as follows. First, we describe a non-parametric approach based on so-called “hot deck imputation” – namely distance hot deck matching. The remaining approaches are parametric and based on file concatenation and multiple imputation (Rubin, 1986; 1987). The file concatenation perspective sees statistical matching 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.

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

Nonparametric Hot Deck Matching

26. 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, Di Zio, and Scanu (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, Di Zio, and Scanu (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 2009 and TALIS 2008, 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 2009 and TALIS 2008 did not result in important exogenous changes across schools.

27. In addition to these substantive concerns the sample sizes of the data sets is 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 recipient data set to the smaller of the two – in this case TALIS. We can see why this is reasonable. If TALIS were 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.

28. 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 as, 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, we use the R program StatMatch (D’Orazio, 2011) for non-parametric hot deck matching.

Distance Hot Deck Matching

29. 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 report, we chose the Euclidean distance metric. 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 Statistical Matching

30. As noted earlier, in addition to nonparametric methods based on variants of hot deck imputation, parametric statistical matching 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 matching methods followed by three statistical matching methods derived from the Bayesian perspective. The two frequentist approaches discussed next are implemented in the R software program mice (van Buuren & Groothuis-Oudshoorn, 2010).

Stochastic Regression Imputation

31. A common approach to statistical matching 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 statistically matching 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.

32. 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 plausible 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 multiply imputed data sets. Subsequent analyses are then based on analysing all of the data sets simultaneously and then pooling the results according to rules set down by Rubin (1987).

Semi-Parametric Predictive Mean Matching

33. Regression imputation and hot deck matching sets the groundwork 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. Thus, the combination of the parametric prediction equation with non-parametric hot deck matching yields this semi-parametric procedure.

Bayesian Approaches to Statistical Matching

34. In the previous section, we concentrated on two approaches to statistical matching 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 allows us to estimate the sample variability around the estimates of the model parameters.

35. In contrast to the frequentist school of statistics, the Bayesian school adopts an entirely different view of statistical inference. Specifically, the Bayesian school views any unknown quantity, and particularly parameters, as random, possessing 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 & Depaoli, in press, for an overview of Bayesian inference).

36. The central reason for adopting a Bayesian perspective to the problem of statistical matching (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 (Shafer, 1997) 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, which can further reduce uncertainty in model parameters.

Bayesian Regression Imputation via Chained Equations

37. In this section we concentrate our discussion on a Bayesianly proper form of multiple imputation using the method of chained equations.³

³ Another popular form of Bayesianly proper imputation involves the data augmentation algorithm of Tanner and Wong (1987). The method of chained equations recognises 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 previous work on chained equations can be found in Tanner and Wong (1987).

38. 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 were a count, then a Poisson model would be appropriate. This is a major advantage over other Bayesianly proper methods such as data augmentation 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.

39. Once the sequence is completed for all variables with missing data, the posterior distribution of the regression parameters is 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 algorithm that generates the sequence of iterations is based on the so-called Gibbs sampler (Geman & Geman, 1984) a very popular method for obtaining draws from posterior distributions. Finally, the algorithm can run these sequences simultaneously m number of times obtaining m imputed data sets. For the purposes of this report, we utilise the chained equation algorithm implemented in the R software program mice (van Buuren & Groothuis-Oudshoorn, 2010).

Bayesian Bootstrap Predictive Mean Matching

40. 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.

41. 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). 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. For specific details, see Rubin (1981). Next, 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 $m < 1$ times to create m multiply imputed data sets. For this report, we use the R software program BaBooN (Meinfelder, 2011).

A Hybrid Method: The EM Bootstrap

42. In this section we examine an approach that combines Bayesian imputation concepts with the frequentist idea of bootstrap sampling. Essentially, bootstrapping is a data-based simulation method which relies on drawing repeated samples from the data to estimate the sample distribution of almost any statistic, and was developed as a simplified alternative to inferences derived from statistical theory (Efron and

Tibshirani, 1993). Specifically, this section considers the implementation of the EM algorithm with bootstrapping.

43. Briefly, EM stands for expectation-maximisation and is an algorithm that is widely used to obtain maximum likelihood estimates of model parameters in the context of missing data problems. The essence of the EM algorithm proceeds as follows. Using a set of starting values for the means and the covariance matrix of the data (perhaps obtained from listwise deletion), the E-step of the EM algorithm creates the sufficient statistics necessary to obtain regression equations that yield the predictions of the missing data given the observed data and the initial set of model parameters. The next step is to use the “filled-I” data to obtain new estimates of model parameters via the M-step, which is simply the use of straightforward equations to obtain new estimates of the vector of means and the covariance matrix of the data. The algorithm then iterates back to the E-step to obtain new regression equations. The algorithm cycles between the E-step and the M-step until a convergence criterion has been met, at which point the maximum likelihood estimates have been obtained. The E-step and M-step are the likelihood counterparts of the Bayesian I-step and P-step in data augmentation.

44. The EM algorithm has been extended to handle the problem of multiple imputation without the need for computationally intensive draws from the posterior distribution, as with the data augmentation approach. The idea is to extend the EM algorithm using a bootstrap approach. This approach is labeled EMB (Honaker & King, 2010) and implemented in the R program Amelia (Honaker, King & Blackwell, 2010), which we use in our analyses below.

45. Following (Honaker & King, 2010) and Honaker (personal communication, 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 bootstrapped data set, the EM algorithm is run. It is here that Honaker & King (2010) allow for the inclusion of prior distributions on the model parameters estimated via the EM algorithm. Notice that because m bootstrapped samples are obtained, and that each EM run on these samples 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 priori* (MAP) estimates; the Bayesian counterpart of the maximum likelihood estimates. Finally, 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.

Data and Methods

46. The PISA 2009 survey design samples schools proportional to size followed by a sample of the target student population within those schools.⁴

47. 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 2009 collected 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 during the survey window of March to September 2009.

⁴ There are additional complexities to the sampling designs of PISA and TALIS that can be found in their respective technical reports.

48. For TALIS 2008, 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 2008 provides teacher-level and school-level data from reports by teachers and school administrators across 16 OECD-member countries and 7 partner countries and economies.⁵

An Experimental Study Using Data from Iceland

49. In total, 142 schools participated in the TALIS survey and/or the PISA survey. Of these, 122 PISA and TALIS schools could 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 statistical matching procedures utilise data from 78 schools in Iceland with full information from the PISA and TALIS data sets.

50. 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.⁶

Variables

51. 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.

Matching Variables

52. 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.⁷

⁵ To be included in the TALIS study, a minimum of 200 schools must participate in the surveys.

⁶ 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 statistical matching 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.

⁷ Information about the average disciplinary climate for each school was drawn from student surveys in PISA and the teacher surveys in TALIS, 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. 204) these variables are suitable to use as matching variables.

53. 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 standardised 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 statistical matching 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.

54. 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.

Unique Variables

55. 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. summarising skills and memorization strategies) promote 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 summarising 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% for Iceland). Also, 21% of the variation in reading performance across OECD countries can be explained by variation in summarising skills (OECD, 2010b, pg. 47) (20% for Iceland). Both measures are averaged to the school level for analysis.

56. 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' behaviour 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 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 for Iceland

57. Software code for the statistical matching methods is presented in Annex A and software code for implementing the validity checks is given in Annex B for the hot deck matching method only. Validity checking for the other methods would be implemented in the same way.

58. An inspection of Table 1 shows the descriptive statistics for the Iceland data for the original data and Tables 2 to 7 show the results for each statistical matching 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.

59. An inspection of Tables 1 to 7 shows 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 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 these validity criteria.

60. Figures 1 to 6 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 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. A Q-Q plot is a graphical approach for 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 Q-Q plots overall, and particularly better than the EM bootstrap method.

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	78.00	0.30	0.36	0.31	0.31	0.39	-0.58	1.02	1.60	-0.27	-0.41	0.04
jobsat	78.00	3.13	0.20	3.12	3.13	0.18	2.75	3.57	0.82	0.16	-0.49	0.02
joyread	78.00	-0.05	0.31	-0.09	-0.06	0.30	-0.93	0.91	1.84	0.55	1.38	0.04
metasum	78.00	-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

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	780.00	0.30	0.37	0.31	0.31	0.39	-0.93	1.42	2.35	-0.13	0.03	0.01
jobsat	780.00	3.12	0.20	3.11	3.12	0.18	2.41	3.82	1.40	-0.02	-0.01	0.01
joyread	780.00	-0.05	0.34	-0.08	-0.06	0.31	-0.94	1.26	2.20	0.25	0.53	0.01
metasum	780.00	-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	780.00	0.29	0.35	0.30	0.30	0.39	-0.58	1.02	1.60	-0.27	-0.48	0.01
jobsat	780.00	3.12	0.21	3.11	3.11	0.19	2.75	3.57	0.82	0.13	-0.61	0.01
joyread	780.00	-0.06	0.32	-0.11	-0.06	0.27	-0.93	0.91	1.84	0.24	0.73	0.01
metasum	780.00	-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

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	780.00	0.30	0.39	0.31	0.30	0.39	-1.07	1.92	2.99	-0.12	-0.02	0.01
jobsat	780.00	3.13	0.22	3.11	3.12	0.20	2.47	3.88	1.41	0.12	0.07	0.01
joyread	780.00	-0.04	0.35	-0.06	-0.04	0.33	-1.39	1.42	2.82	0.19	0.78	0.01
metasum	780.00	-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	780.00	0.27	0.35	0.30	0.28	0.41	-0.58	1.02	1.60	-0.18	-0.52	0.01
jobsat	780.00	3.12	0.19	3.11	3.12	0.16	2.75	3.57	0.82	0.22	-0.51	0.01
joyread	780.00	-0.08	0.31	-0.12	-0.08	0.26	-0.93	0.91	1.84	0.23	0.63	0.01
metasum	780.00	-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

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	780.00	0.32	0.36	0.34	0.33	0.36	-1.11	1.38	2.49	-0.26	0.06	0.01
jobsat	780.00	3.12	0.20	3.11	3.11	0.17	2.38	3.70	1.31	0.00	-0.04	0.01
joyread	780.00	-0.03	0.33	-0.05	-0.03	0.31	-1.02	0.99	2.01	0.07	0.32	0.01
metasum	780.00	-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

Discussion

61. The purpose of this report 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 report that statistically matching PISA and TALIS might be a reasonable option for countries that are unable to administer both surveys to the same sample of schools. Our analyses suggest that statistically matching PISA and TALIS is feasible and can be considered by countries interested in gleaning added value from both surveys.

62. Among the methodologies that were considered in this report, two stand out as deserving serious consideration for matching PISA and TALIS – Bayesian bootstrap predictive mean matching, and the EM-

bootstrap. Both methodologies worked quite well with respect to Rässler's (2002) third and fourth level validity criteria. It should be noted that both algorithms were implemented without the specification of priors on the model parameters. We anticipate that the implementation of priors would influence the comparability of the method to other methods depending on the precision of the priors. That is, highly precise priors around incorrect model parameters would likely result in poor performance compared to precise priors around correct values. An issue regarding the use of priors for the BBPMM or the EM-bootstrap concerns how priors might be elicited. Findings from matching current cycles of PISA and TALIS could be used to inform the specification of priors for future statistical matching exercises.

63. The use of statistical matching and imputation methods has consequences for subsequent modeling activities. Specifically, the type of matching method being employed can influence the variation in model parameters. For example, in the case of regression imputation (without a stochastic component), the standard error of regression coefficients would be underestimated, leading to an inflated R^2 and thus an increase in Type I errors. Adding a stochastic component improves on this problem, but the imputed data is still being treated as though it is known. Methods that involve multiple imputation restore variability back into the data through the creation and analysis of multiple data sets. From Rubin (1987) we know that the efficiency of an estimate is a function of the number of imputations and the rate of the missing data. For hot deck matching the issue of sampling variability in the estimates is complicated. First, because the sample size is the size of the recipient sample, there may likely be a loss of power. However, variability in the estimates may be reduced because the recipient data set treats the imputed data as though it were known. Moreover, we suspect that the sampling variability of the estimates will differ as a function of the type of distance metric being used (where here we used a nearest neighbour distance metric). At present, we know of no systematic comparative examination of the methods used in this report with respect to sampling variability of parameter estimates, and thus, we cannot provide systematic guidance on this issue. A conservative approach, however, would be to create different synthetic data sets using the methods proposed here with differing assumptions as a rough “sensitivity analysis” to gauge changes in estimates and standard errors.

64. As noted earlier, statistical matching 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 statistical matching algorithm in the software program mice (van Buuren & Groothuis-Oudshoorn, 2010) based on the Gibbs sampling algorithm. For future cycles, should countries participating in PISA use a teacher questionnaire, and assuming that there are teacher-level variables common to TALIS and the PISA teacher questionnaires, two-level statistical matching may be feasible and certainly worth exploring.

65. In the context of cross-national education research, statistical matching within countries may allow 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 teachers' 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 2008 survey, and each of these 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 statistical matching to provide complete information on multiple countries is promising. For example, PISA data suggest that the best performing education systems prioritise 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 individualised approaches to learning (OECD, 2010a). TALIS 2008 data suggest that professional

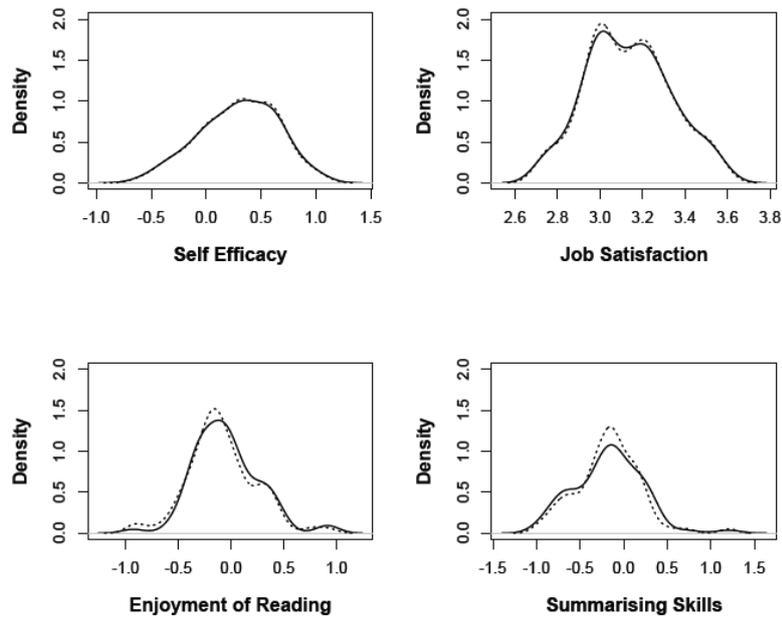
development, teaching practices, teachers' beliefs and attitudes, school and teacher evaluation methods are important for understanding and improving educational processes (OECD, 2009).

66. In the absence of a new design that formally links through the administration of PISA and TALIS jointly, statistical matching provides the next best approach for addressing these important policy questions.

67. To conclude, this report 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).

68. In the absence of a direct implementation of both surveys, countries may wish to pursue this line of investigation. It should be noted, however, that substantive applications for these matching procedures are still in their infancy. The case of Iceland's data for PISA and TALIS offers an exciting opportunity to investigate these methods, but the small sample size limited the additional manipulations that could be done to investigate these methods further (and determine, for example, the effect of the level of error in the imputed data on further analyses). More research is therefore needed to form firm conclusions regarding how and when matching would be desirable

**Figure 1. Kernel Density Plots for Matched Iceland Data
Hot Deck Distance Matching**



**Figure 2. Quantile-Quantile Plots for Matched Iceland Data
Hot Deck Distance Matching**

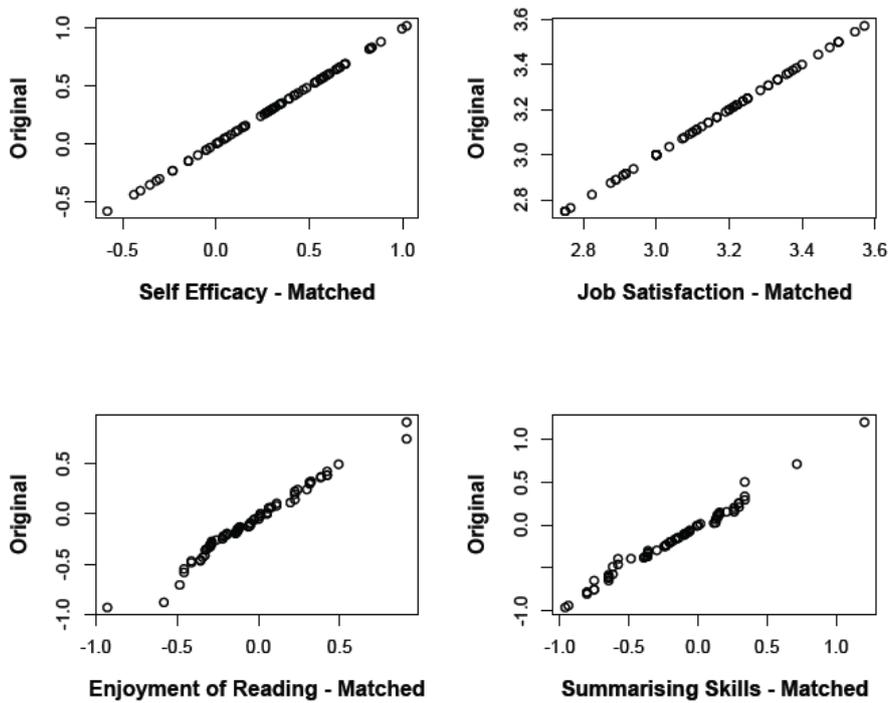


Figure 3. Kernel Density Plots for Matched Iceland Data

Linear Regression Ignoring Model Error

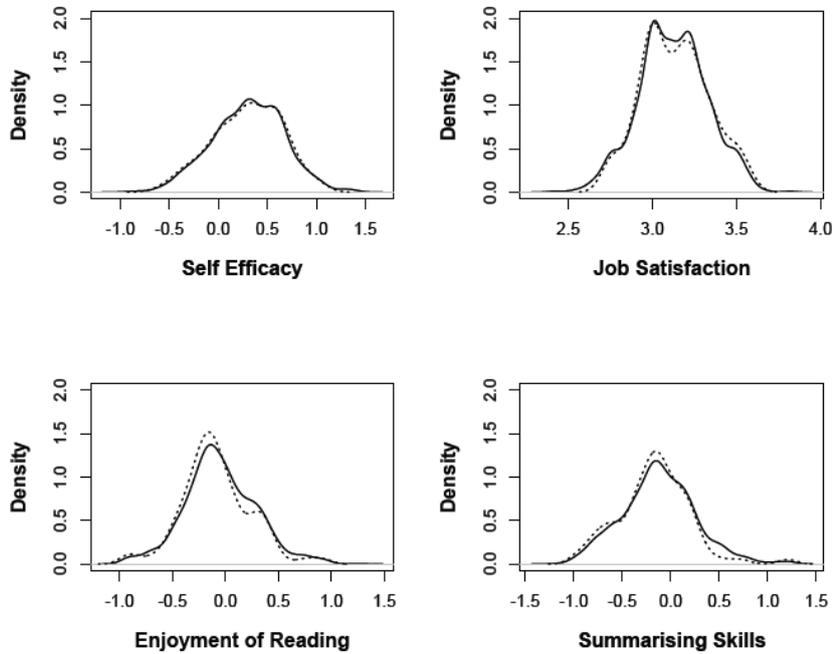


Figure 8. Quantile-Quantile Plots for Matched Data

Linear Regression Ignoring Model Error

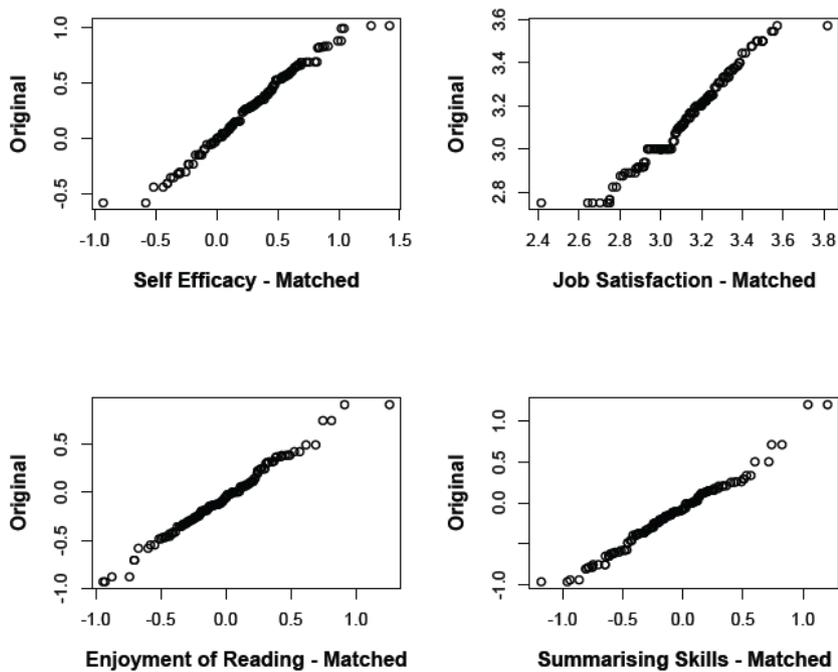


Figure 5. Kernel Density Plots for Matched Iceland Data

Predictive Mean Matching

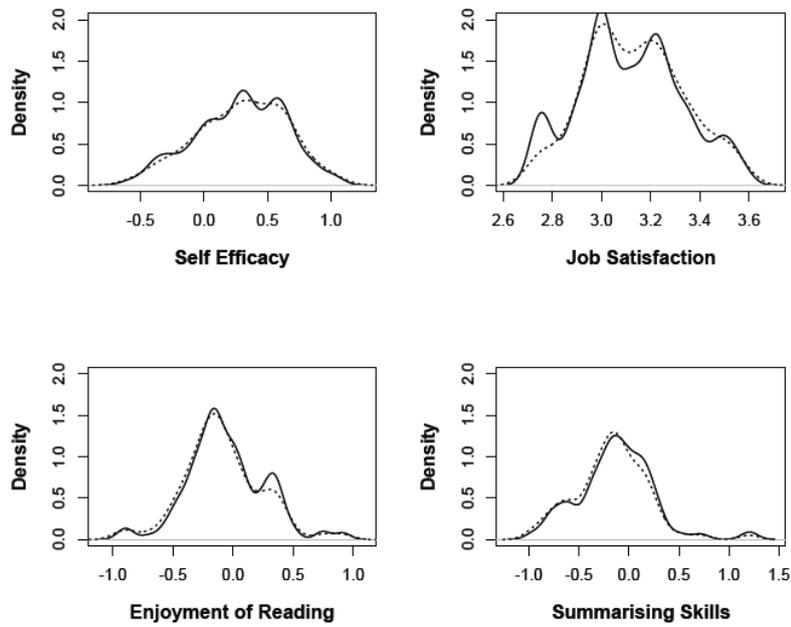
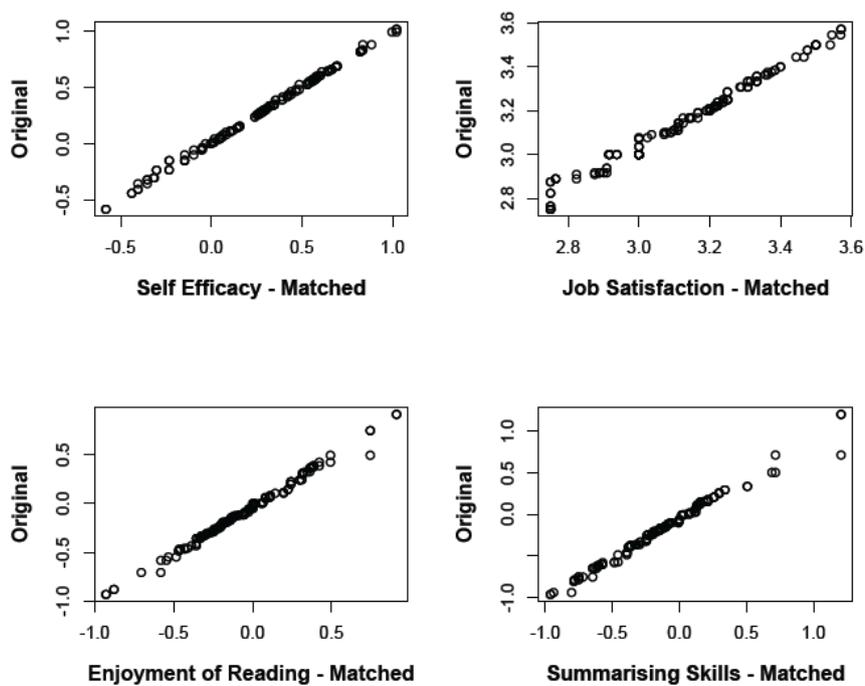
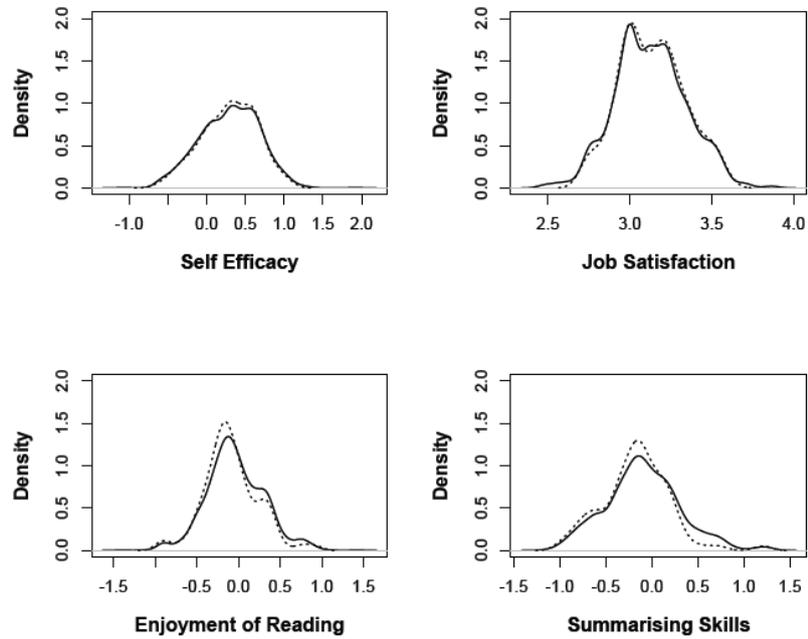


Figure 6. Quantile-Quantile Plots for Matched Iceland Data

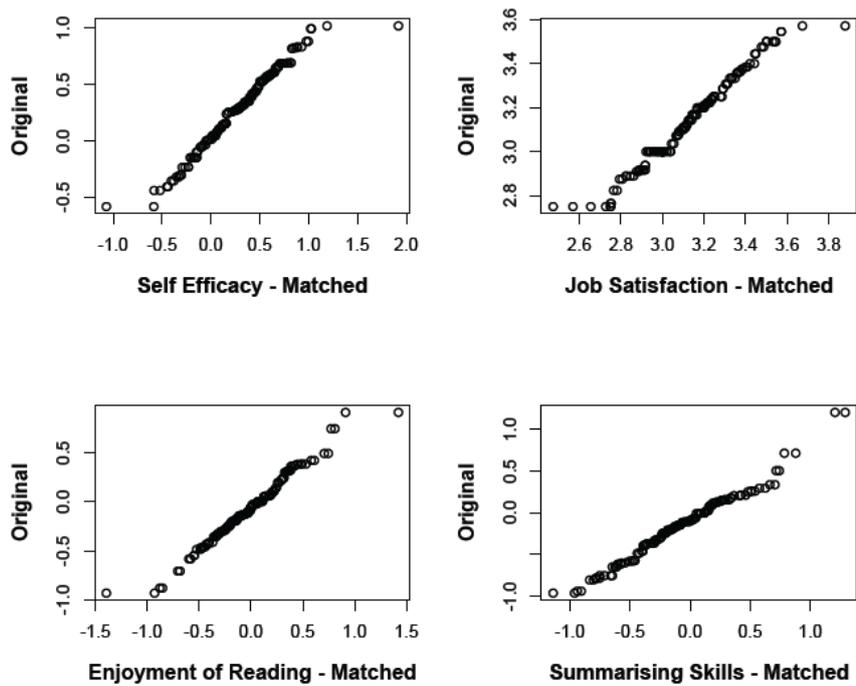
Predictive Mean Matching



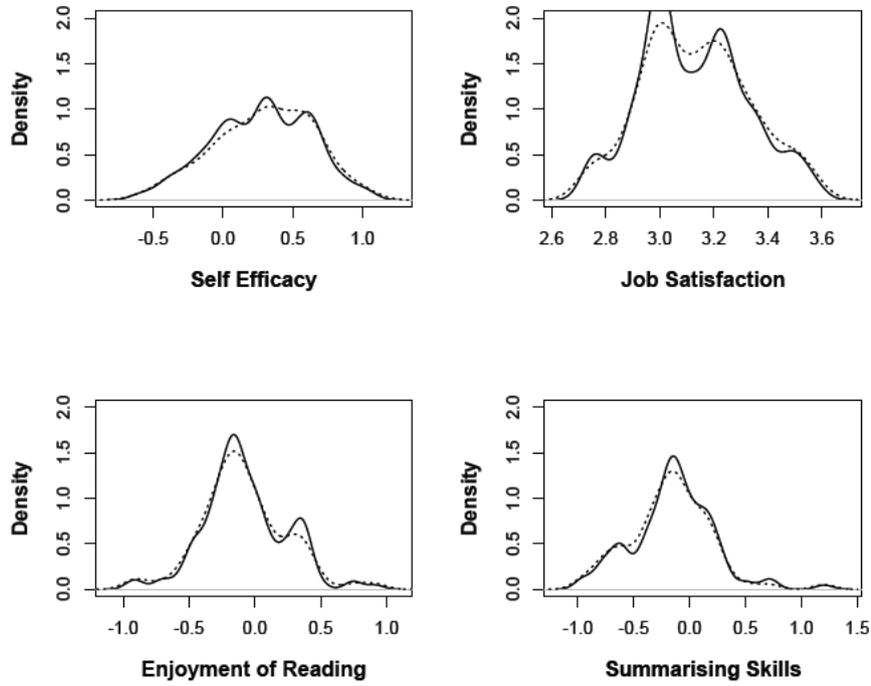
**Figure 7. Kernel Density Plots for Matched Iceland Data
Multiple Imputation with Chained Equations**



**Figure 4. Quantile-Quantile Plots for Matched Iceland Data
Multiple Imputation with Chained Equations**



**Figure 9. Kernel Density Plots for Matched Iceland Data
Bayesian Predictive Mean Matching**



**Figure 10. Quantile-Quantile Plots for Matched Iceland Data
Bayesian Predictive Mean Matching**

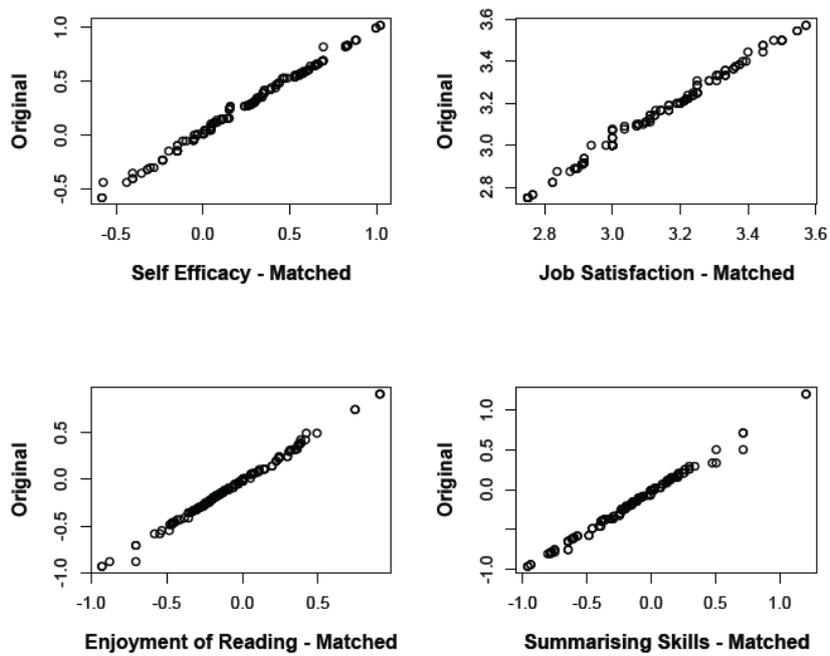


Figure 11. Kernel Density Plots for Matched Iceland Data

EM Bootstrap

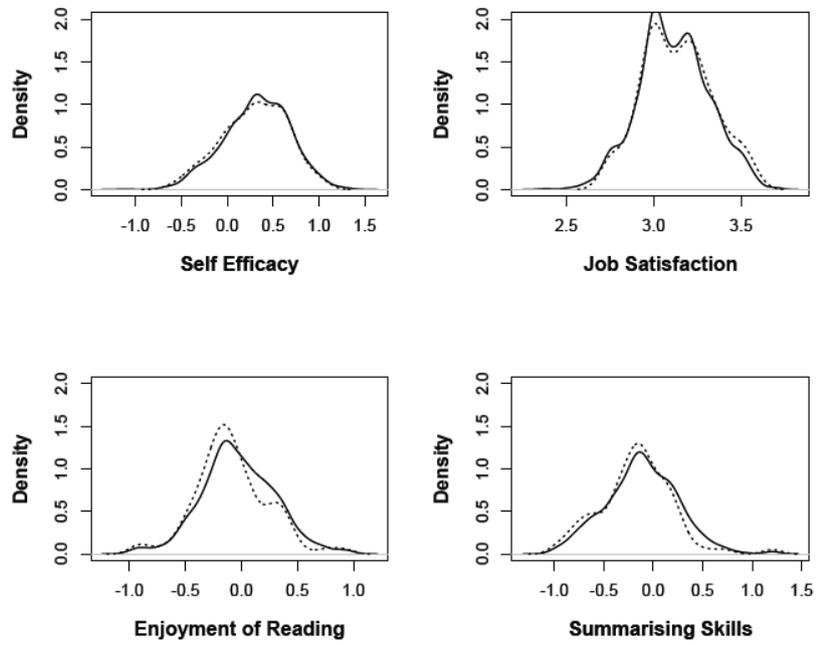
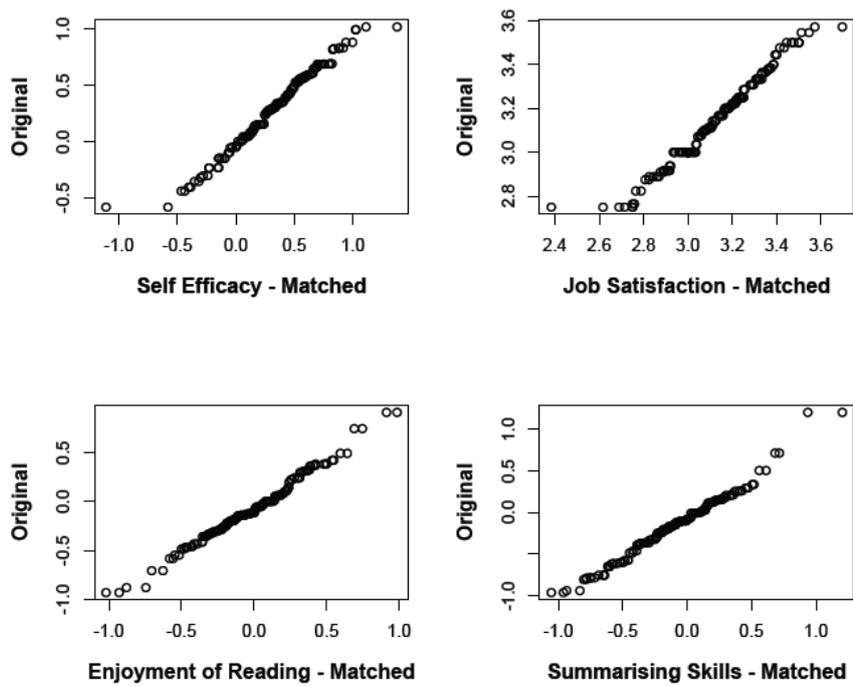


Figure 12. Quantile-Quantile Plots for Matched Iceland Data

EM Bootstrap



REFERENCES

- Ashton, P. and N. Webb (1986). *Making a Difference: Teacher Efficacy and Student Achievement*. Monogram, Longman, White Plains, New York.
- Barr, R., & Dreeben, R. (1983). *How schools work*. Chicago: University of Chicago Press.
- van Buuren, S., & Groothuis-Oudshoorn, K. (2010, January). Multivariate imputation by chained equations, version 2.3. <http://www.multiple-imputation.com/>.
- van Buuren, S., & Groothuis-Oudshoorn, K. (forthcoming). Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*.
- 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 jackknife. *The Annals of Statistics*, 7, 1-26.
- Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.
- Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford.
- Gamoran, A., & Dreeben, R. (1986). Coupling and control in educational organizations. *Administrative Science Quarterly*, 31, 612-632.
- Gamoran, A., Secada, W. G., & Marrett, C. B. (2000). The organizational context of teaching and learning. In M. T. Hallinan (Ed.), *Handbook of the sociology of education* (pp. 37-63). New York: Kluwer Academic/Plenum Publisher.
- Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Trans. Pattn. Anal. Mach. Intel.*, 6, 721-741.
- Hanushek, E. A., & Lindseth, A. A. (2009). *Schoolhouses, courthouses, and statehouses: Solving the funding-achievement puzzle in America's public schools*. Princeton, NJ: Princeton University Press.
- 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).
- Jencks, C., & Tach, L. (2006). Would equal opportunity mean more mobility? In S. Morgan, D. Grusky, & G. Fields (Eds.), *Mobility and inequality: Frontiers of research in sociology and economics (studies in social inequality)* (pp. 26-57). Palo Alto: Stanford University Press.
- Kaplan, D. (1995). The impact of BIB spiraling-induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioural Statistics*, 20, 69-82.

Kaplan, D., & Depaoli, S. (in press). Bayesian statistical methods. In T. D. Little (Ed.), Oxford handbook of quantitative methods. Oxford: Oxford University Press.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data (2nd. ed.). New York.

Meinfelder, F. (2011). BaBooN: Bayesian bootstrap predictive mean matching for 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)

OECD. (2009). Creating effective teaching and learning results: First results from TALIS.

OECD. (2010a). PISA 2009 Results: Executive summary.

OECD. (2010b). PISA 2009 Results: Learning to learn from student engagement, strategies and practices (Vol. 3).

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).

Rassler, S. (2002). Statistical matching: A frequentist theory, practical applications, and alternative Bayesian approaches. New York: Springer.

Ross, J.A. (1998), "The Antecedents and Consequences of Teacher Efficacy", in J. Brophy (ed.) *Advances in Research on Teaching*, Vol. 7, pp. 49-74. JAI Press, Greenwich, Connecticut.

Rubin, D. B. (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 non-response surveys*. Hoboken, NJ: Wiley.

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Tanner, M. H., & Wong, W. A. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association*, 82, 528-550.

Werfhorst, H. Van de, & Mijs, J. M. (2010). Achievement inequality and the institutional structure of educational systems: A comparative perspective. *Annual Review of Sociology*, 36, 407-428.

ANNEX A

R Scripts for PISA-TALIS Matching Project
 # David Kaplan and Alyn Turner
 # University of Wisconsin-Madison

```
#####
# Amelia (EM Bootstrap)
#####

#Step 1: Read in data file that includes PISA and TALIS schools
pisatalis <- read.csv("datafile.csv",header=T)

#Step 2: Set bounds on variables to be imputed. These should be determined by the actual distributions of each
variable.
bds <- matrix(c(1,-3.5,3.5, 2,1,4, 3,-3.5,3.5, 4,-3.5,3.5), nrow = 4, ncol=3, byrow=TRUE)

#Step 3: Run the AMELIA program, specifying the data object, the number of imputed data sets desired, and the
bounds for imputed variables
amelia <- amelia(x=pisatalis,m=5, bounds=bds)

#Step 4: Save imputed datasets
write.amelia(amelia,file.stem="amelia",extension=".csv")

#####
# BaBooN (Bayesian Predictive Mean Matching)
#####

#Step 1: Read in data file that includes PISA and TALIS schools
pisatalis <- read.csv("datafile.csv",header=T)

#Step 2: Run BaBooN program, specifying the data object, number of iterations desired, the name of the output
file, and the number of imputed data sets desired.
pisatalis.bbpm <- BBPMM(pisatalis, nIter=5, outfile="BaBooN.csv", M=5)

#####
# MICE pmm (Predictive Mean Matching)
#####

#Step 1: Read in data file that includes PISA and TALIS schools
pisatalis <- read.csv("datafile.csv",header=T)

#Step 2: Prepare program to run "PMM", and specify bounds for imputed variables based on variable distributions
ini <- mice(pisatalis,max=0,pri=F)
meth <- ini$meth
meth["x1"] <- "pmm"
meth["x2"] <- "pmm"
meth["y1"] <- "pmm"
meth["y2"] <- "pmm"
post <- ini$post
post["x1"] <- "imp[[j]][i] <- squeeze(imp[[j]][i],c(-3.5,3.5))"
post["x2"] <- "imp[[j]][i] <- squeeze(imp[[j]][i],c(1,4))"
post["y1"] <- "imp[[j]][i] <- squeeze(imp[[j]][i],c(-3.5,3.5))"
post["y2"] <- "imp[[j]][i] <- squeeze(imp[[j]][i],c(-3.5,3.5))"

#Step 3: Run the MICE program, specifying the data object, number of desired imputed datasets, the method (defined in Step 2), and the desired number of iterations,
and bounds (defined in Step 2)
pisatalis.pmm <- mice(pisatalis, m = 5, meth = meth, maxit = 5, post=post)

#Step 4: Write the imputed data sets to a file
pisatalis.complete.pmm <- complete(pisatalis.pmm, "long", inc=T)
write.table(pisatalis.complete.pmm,file="pmm.csv",sep=",")

#####
# MICE norm (Bayesian Linear Regression)
#####

#Step 1: Read in data file that includes PISA and TALIS schools
pisatalis <- read.csv("datafile.csv",header=T)

#Step 2: Prepare program to run "NORM", and specify bounds for imputed variables based on variable distributions
ini <- mice(pisatalis,max=0,pri=F)
meth <- ini$meth
meth["x1"] <- "norm"
meth["x2"] <- "norm"
```

```

meth["y1"] <- "norm"
meth["y2"] <- "norm"
post <- ini$post
post["x1"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-3.5,3.5))"
post["x2"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(1,4))"
post["y1"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-3.5,3.5))"
post["y2"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-3.5,3.5))"

```

#Step 3: Run the MICE program, specifying the data object, number of desired imputed datasets, the method(defined in Step 2), and the desired number of iterations, and bounds (defined in Step 2)

```

pisatalis.norm <- mice(pisatalis, m = 5, meth = meth, maxit = 5, post=post)
pisatalis.complete.norm <- complete(pisatalis.norm, "long", inc=T)

```

#Step 4: Write the imputed data sets to a file
write.table(pisatalis.complete.norm,file="norm.csv",sep=";")

```

#####
# MICE norm.nob (Non-Bayesian Linear Regression)
#####

```

#Step 1: Read in data file that includes PISA and TALIS schools
pisatalis <- read.csv("datafile.csv",header=T)

#Step 2: Prepare program to run "NORM.NOB", and specify bounds for imputed variables based on variable distributions

```

ini <- mice(pisatalis,max=0,pri=F)
meth <- ini$meth
meth["x1"] <- "norm.nob"
meth["x2"] <- "norm.nob"
meth["y1"] <- "norm.nob"
meth["y2"] <- "norm.nob"
post <- ini$post
post["x1"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-3.5,3.5))"
post["x2"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(1,4))"
post["y1"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-3.5,3.5))"
post["y2"] <- "imp[[j]][,i] <- squeeze(imp[[j]][,i],c(-3.5,3.5))"

```

#Step 3: Run the MICE program, specifying the data object, number of desired imputed datasets, the method(defined in Step 2), and the desired number of iterations, and bounds (defined in Step 2)

```

pisatalis.normnob <- mice(pisatalis, m = 5, meth = meth, maxit = 5, post=post)

```

#Step 4: Write the imputed data sets to a file
pisatalis.complete.normnob <- complete(pisatalis.normnob, "long", inc=T)
write.table(pisatalis.complete.normnob,file="normnob.csv",sep=";")

```

#####
# StatMatch (Hot Deck Distance Matching)
#####

```

#Step 1: Read in data file that includes PISA and TALIS schools
pisatalis <- read.csv("datafile.csv",header=T)

#Step 2: Identify TALIS schools
talisrow <- c(1:78)

#Step 3: Identify PISA schools
pisarow <- c(79:156)

#Step 4: Set donor and recipient data frames
pisa.don <- pisatalis[pisarow, c(3:4,5:10)] # donor data.frame
talis.rec <- pisatalis[talisrow, c(1:2,5:10)] # recipient data.frame

#Step 5: Run Hot deck Matching program using the Euclidean distance function, specifying the columns that include the variables on which the match is to be based
out.NND <- NND.hotdeck(data.rec=talis.rec, data.don=pisa.don,dist.fun="Euclidean", match.vars=c(3:8))

#Step 6: Create synthetic data.set, without the duplication of the matching variables
fused.1 <- create.fused(data.rec=talis.rec, data.don=pisa.don, mtc.ids=out.NND\$mtc.ids, z.vars=c("y1","y2"))

#Step 7: Write the dataset to a file
write.table(fused.1,file="hotdeck.csv",sep=";")

ANNEX B

```

## Script for calculating marginal distributions and conditional covariance matrix
## Needed to check third and fourth order validity

require(MASS)
require(psych)
require(graphics)
require(xtable)
require(Zelig)

##Set working directory
#getwd()
#setwd("~/Desktop/OECD")

pisatalis <- read.csv("~/Documents/Data Fusion/OECD/Analysis/Iceland/Data/Iceland78.csv",header=T)
pisatalis <- pisatalis[c(-11:-13)]
pisatalis
#move header over, delete column, delete original NA's
hotdeck <- read.csv("~/Documents/Data Fusion/OECD/Analysis/Iceland/Data/Match/Matched/hotdeck.csv",header=T)
hotdeck

#####
## Marginal distributions and plots for original and fused files

summary.matched <- describe(hotdeck[,1:4])
summary.matched
summary.matched.xtable <- xtable(summary.matched,caption='Summary Statistics on Matched Iceland Data: Hot
Deck Distance Matching.')
print(summary.matched)
print(summary.matched.xtable)

#####
## Calculate conditional covariance matrix of x and y given z. Values should be close to zero

fused <- read.csv("~/Documents/Data Fusion/OECD/Analysis/Iceland/Data/Match/Matched/hotdeck.csv",header=T)
fused1 <- fused
print(fused1)

fusedcov <- cov(fused1)
fusedcov
fusedcor <- cov2cor(fusedcov)
fusedcov

fusedcovxy <- fusedcov[1:2,3:4]
fusedcovxy
fusedcorxy <- fusedcor[1:2,3:4]
fusedcorxy

sigmaxy <- print(fusedcov[1:2,3:4])
sigmaxz <- print(fusedcov[1:2,5:10])
sigmazzin <- print(solve(fusedcov[5:10, 5:10]))
sigmazzy <- print(fusedcov[5:10,3:4])
sigmaxx <- print(fusedcov[1:2,1:2])
sigmayy <- print(fusedcov[3:4,3:4])
sigmayx <- print(t(sigmaxy))
sigmayz <- print(t(sigmazy))
sigmazx <- print(t(sigmaxz))

condcovxy <- print(sigmaxy - (sigmaxz%%sigmazzin%%sigmazzy))
condvarx <- print(sqrt(sigmaxx-(sigmaxz%%sigmazzin%%sigmazzy))
condvary <- print(sqrt(sigmayy-(sigmayz%%sigmazzin%%sigmazzy))

a <- rbind(condvarx,t(condcovxy))
b <- rbind(condcovxy, condvary)

condcovfull <- cbind(a,b)
condcovfull

condcorrfull <- print(cov2cor(condcovfull))

condcorrxy <- print(condcorrfull[1:2,3:4])
condcorrxy
corrxy_mean <- xtable(condcorrxy,caption='Conditional Correlation for Matrix Matched Iceland Data: Hot Deck
Distance Matching.')
```

```

print(corrxy_mean)

#####
## Create density plots for fused and original variables

par(mfrow=c(2,2),oma=c(5,0,0,0),font=2,font.lab=2,cex.main=1.2,cex.lab=1.2,cex.sub=1)

plot(density(fused1$x1, na.rm=TRUE),main="",ylim=c(0,2),xlab='Self Efficacy')
lines(density(pisatalis$x1,na.rm=TRUE),lty=3)

plot(density(fused1$x2, na.rm=TRUE),main="",ylim=c(0,2),xlab='Job Satisfaction')
lines(density(pisatalis$x2,na.rm=TRUE),lty=3)

plot(density(fused1$y1, na.rm=TRUE),main="",ylim=c(0,2),xlab='Enjoyment of Reading')
lines(density(pisatalis$y1,na.rm=TRUE),lty=3)

plot(density(fused1$y2, na.rm=TRUE),main="",ylim=c(0,2),xlab='Summarising Skills')
lines(density(pisatalis$y2,na.rm=TRUE),lty=3)

mtext("Kernel density plots for Matched Iceland Data: Hot Deck Distance
Matching.",cex.main=1,side=1,outer=TRUE)

#####

# Compare imputed values to original with qqplot
# This assesses 4th level validity

par(mfrow=c(2,2),oma=c(5,0,0,0),font=2,font.lab=2,cex.main=1.2,cex.lab=1.2,cex.sub=1)
qqplot(fused1$x1,pisatalis$x1,plot.it=TRUE,ylab='Original',xlab='Self Efficacy - Matched')
qqplot(fused1$x2,pisatalis$x2,plot.it=TRUE,ylab='Original',xlab='Job Satisfaction - Matched')
qqplot(fused1$y1,pisatalis$y1,plot.it=TRUE,ylab='Original',xlab='Enjoyment of Reading - Matched')
qqplot(fused1$y2,pisatalis$y2,plot.it=TRUE,ylab='Original',xlab='Summarising Skills - Matched')
mtext("qqplots plots for Matched Iceland Data: Hot Deck Distance Matching.",cex.main=1,side=1,outer=TRUE)

#####

```

ANNEX C

RECENT OECD PUBLICATIONS OF RELEVANCE TO THIS WORKING PAPER

Jensen, B., et al. (2012), *The Experience of new Teachers: Results from TALIS 2008*, OECD Publishing.

OECD (2012), *Equity and Quality in Education: Supporting Disadvantaged Students and Schools*, OECD Publishing.

OECD (2010), *TALIS 2008 Technical Report*, OECD Publishing.

OECD (2010), *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I)*, OECD Publishing.

OECD (2010), *PISA 2009 Results: Overcoming Social Background: Equity in Learning Opportunities and Outcomes (Volume II)*, OECD Publishing.

OECD (2010), *PISA 2009 Results: Learning to Learn – Student Engagement, Strategies and Practices (Volume III)*, OECD Publishing.

OECD (2010), *PISA 2009 Results: What Makes a School Successful? – Resources, Policies and Practices (Volume IV)*, OECD Publishing.

OECD (2010), *PISA 2009 Results: Learning Trends: Changes in Student Performance since 2000 (Volume V)*, OECD Publishing.

OECD (2009), *Creating Effective Teaching and Learning Environments: First Results from TALIS*, OECD Publishing.

THE OECD EDUCATION WORKING PAPERS SERIES ON LINE

The OECD Education Working Papers Series may be found at:

- The OECD Directorate for Education website: www.oecd.org/edu/workingpapers
- The OECD's online library, SourceOECD: www.sourceoecd.org
- The Research Papers in Economics (RePEc) website: www.repec.org

If you wish to be informed about the release of new OECD Education working papers, please:

- Go to www.oecd.org
- Click on "My OECD"
- Sign up and create an account with "My OECD"
- Select "Education" as one of your favourite themes
- Choose "OECD Education Working Papers" as one of the newsletters you would like to receive

For further information on the OECD Education Working Papers Series, please write to: edu.contact@oecd.org.