Running head: PISA/TALIS LINK

Statistical Matching of PISA and TALIS

David Kaplan
Department of Educational Psychology
Alyn Turner
Department of Sociology
University of Wisconsin – Madison

Contract # EDU/JA00066381
Directorate For Education
Organization for Economic Cooperation and Development

Statistical Matching of PISA and TALIS

Statement of Problem

The OECD Program for International Student Assessment (PISA) and the Teaching and Learning International Survey (TALIS) constitute two of the largest ongoing international student and teacher assessments presently underway. Yet, each survey is missing an important component of the educational system in its survey design. For PISA, the survey design samples schools proportional to size, followed by a sample of the target student population within those schools – namely fifteen year olds. For TALIS, a two-stage stratified probability sample was employed with ISCED Level 2 teachers as second stage units were randomly selected from randomly selected schools.[1] Naturally, policy makers are interested in all three levels of the school system – students, teachers, and schools, in order to fully understand the inputs, processes, and outcomes of education.

One desirable approach to linking the PISA survey to the TALIS survey is to sample schools and administer both PISA and TALIS. However, this may not be feasible for many countries, and yet countries should have a mechanism whereby they can glean as much policy relevant information as possible from their involvement in both surveys, despite not having been conducted in the same schools or at the same point in time. One approach involves *statistical matching*, also referred to as *data fusion*. This proposal seeks to outline a comprehensive study of statistical matching in the context of PISA and TALIS. The intention is to provide a comparative investigation of statistical matching methods, highlighting their strengths and weaknesses, and to provide guidance as to the software needed to successfully fuse PISA and TALIS.

It should be noted that this proposal is quite technical, and yet it still lacks very specific statistical issues that will appear in publications emanating from this research. However, given increased interest in statistical matching of distinct data bases, a comprehensive study applied to PISA and TALIS is warranted. The results of this study should provide relatively simply guidelines and software necessary for key stakeholders to conduct a statistical matching of PISA and TALIS.

The organization of this proposal is as follows. In the next section, I provide a brief outline of the statistical matching problem. Next, I outline the proposed set of analyses that will be applied to PISA and TALIS. This is followed by a tentative timeline for the proposed research study and a description of the deliverables of the project.

Background on Statistical Matching

Simply put, statistical matching (also known as data fusion) is embedded in the larger problem of missing data analysis. Following the seminal work of (Rubin, 1976, see also; Little & Rubin, 2002), 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 income observed, but may be related to age. For example, perhaps older individuals do not report their incomes. This type of missing data is referred to as *missing at random* or *MCAR*. Under MAR, inferences will be valid, and there now exist many methods for handling missing data under the assumption of MAR.

For practical purposes, MCAR and MAR are fairly unrealistic. A more realistic situation is one in which the mechanism is non-ignorable. Taking our example of age and income, here missing data on income might be 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*. Here, inferences derived from conventional approaches are not valid, and what is required are models for the missing data process itself added to the substantive model of interest.

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 assumed 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 spiralling designs – such as the design for the cognitive outcome assessments in PISA. Another example, of concern to this research proposal is the case of statistical matching of different data sets. In the case of PISA and TALIS, the two data sets have no units in common but do have variables in common. Because there are no units in common across the two datasets, the missing data are reasonably considered to be MCAR.[2]

Approaches to Statistical Matching

In this section, I describe some common approaches to the problem of statistical matching from the conventional frequentist school of statistics and from the Bayesian school of statistics. Approaches within these two schools of statistics will be examined in terms of matching PISA and TALIS. Before beginning, however, it should be noted that the sampling designs of PISA and TALIS are not the same. As stated earlier, the PISA sampling design samples schools and then samples students in schools that satisfy the age requirements of PISA. For TALIS, schools are sampled, and then teachers are sampled within schools. Thus, the level of analysis common to both surveys is the school. This implies that student data in PISA and teacher data in TALIS will need to be aggregated to the school level and matching will take place among common variables at the school level.[3]

This section is organized as follows. First, I will describe two common approaches to statistical matching from the frequentist perspective (a) propensity score matching and (b) regression imputation with random residuals. Next, I will describe approaches based on data augmentation which rest on Bayesian ideas.

Frequentist Approaches to Statistical Matching

Following the work of Rässler (2002, see also; D'Orazio, Di Zio, & Scanu, 2006) the general problem of statistical matching of two data sets such as PISA and TALIS is akin to the problem of file concatenation. Taking PISA and TALIS as the example, let $\mathbf{X}$ denote all of the variables unique to PISA, let $\mathbf{Y}$ denote all the variables unique to TALIS, and let $\mathbf{Z}$ denote all of the variables common to PISA and TALIS. Clearly, $\mathbf{X}$ is missing in TALIS and $\mathbf{Y}$ is missing in PISA, and we will assume that the missing data are missing completely at random. In cases where the common variables in $\mathbf{Z}$ are in different metrics, it is advisable to standardize them –

including categorical variables.

Continuing, let $\mathbf{A} = (\mathbf{X}, \mathbf{Z})$ represent the PISA data set and let $\mathbf{B} = (\mathbf{Y}, \mathbf{Z})$ be the TALIS data set. Using the terminology of statistical matching, let PISA be the *recipient* sample, and let TALIS be the *donor* sample. We choose this particular designation of the donor and recipient sample because it advised that the larger sample be the recipient sample (Rässler, 2002). The goal at this point is to find donor units from TALIS that are to be matched as closely as possible to recipient units in PISA.

Propensity Score Matching. One relatively straightforward approach to statistical matching is based on the theory of propensity scores (Rosenbaum & Rubin, 1983). A discussion of propensity scores is beyond the scope of this proposal. Suffice to say, however, that the propensity score is the estimated conditional probability of assignment to a group given a set of covariates. The propensity score has been productively used in dealing with non-equivalence of treatment and control groups in quasi-experimental designs. The conventional approach to propensity score estimation utilizes logistic regression to yield the estimated probability of treatment assignment conditional on covariates. Then, a variety of different procedures can be implemented to obtain the adjusted treatment effect − adjusted for non-equivalence. These include (a) subclassification on the propensity score (b) inverse probability weighting using the propensity score, and (c) optimal matching on the propensity score.

In the context of statistically matching PISA and TALIS, the propensity score approach would proceed as follows.

1. Extend the PISA and TALIS samples by adding a new survey indicator variable D to both samples, where $D_i = 1$ if the school is observed in PISA and

$D_i = 0$ if the school is observed in TALIS.

2. Conduct a logistic (or probit) regression of the survey indicator D using the common variables $\mathbf{Z}$ as the covariates.

3. Calculate the estimated propensity score for each school based on the estimates obtained from the logistic regression. Note that the propensity score reflects the probability that a school belongs to PISA.

4. For each recipient school in PISA, find a school in TALIS that has a similar propensity score. If a TALIS school is found for every PISA school, then its TALIS $\mathbf{Y}$ variables are added to PISA.

The last item in the above steps rests on the type of algorithm that matches donor units to recipient units. There are a number of algorithms that are available to match donor units to recipient units based on "nearest-neighbor" algorithms or "greedy-matching" algorithms. One approach that we will examine is based on the notion of optimal matching (see e.g. Hansen, 2004; Hansen & Klopfer, 2006; Rässler, 2002; Rosenbaum, 1989). Following Rosenbaum (1989), consider the problem of matching a TALIS school to a PISA school on $\mathbf{Z}$. A *matched pair* is an ordered pair (i, j), with $1 \leq i \leq N$ and $1 \leq j \leq M$ denoting that the i^{th} PISA school is matched with the j^{th} TALIS school. As defined by Rosenbaum (1989) "[A] *complete matched pair* is a set $\Im$ of N disjoint matched pairs, that is N matched pairs in which each PISA school appears once, and each TALIS school appears either once or not at all".

Rosenbaum suggests two aspects of a "good" match. Close matching in terms of a distance measure on the vector of covariates – for example, nearest neighbor. Obtaining close matches becomes more difficult as the number of covariates increases. Good matching, in contrast, is based on covariate balance, for example, obtained on the propensity score. If within-matched sample distributions on the propensity score are similar, then there is presumed to be balanced matching on the covariates.

For this study, we consider *optimal matching*: an improvement on so-called *greedy matching*. Greedy matching finds a TALIS school to be matched to a PISA school on the basis of the distance between those schools alone. The greedy algorithm works sequentially, starting with a match of minimum distance, and then removes the TALIS school from further consideration. Once a PISA school is matched with two TALIS schools, the match is then removed. It is important to point out that greedy matching does not revisit the match, and therefore does not attempt to provide the lowest overall "cost" for the match.

Optimal matching, in contrast, proceeds much the same way as greedy matching. However, rather than simply adding a match, and removing the control (and treatment) from further consideration, optimal matching might reconsider a match if the total distance across matches is less than if the algorithm proceeded. According to Rosenbaum (1989), optimal matching is as good and often better than greedy matching. Indeed, although greedy matching can provide a good answer, there is no guarantee that the answer will be tolerable – and often it can be quite bad. However, when "calipers" are placed on the propensity score, certain matches will be forbidden unless they are within the calipers. Optimal matching on the propensity score, along with calipers, provides perhaps the best balance. For this study, we use the optimal full matching algorithm discussed in Hansen and Klopfer (2006) and implemented in their R package *optmatch*. Again, once an optimal match is found, the missing values are added to PISA. Propensity score matching is, arguably, the simplest approach to fuse PISA and TALIS.

Regression Imputation with Random Residuals. In the context of usual missing data problems, one approach is based on simple regression imputation. Taking the age and income example, with missing data on income, we first regress the complete income data on age via the regression $income = a + b(age) + e$. Next, we obtain the predictive value of income, that is $\widehat{income} = \hat{a} + \hat{b}(age)$. Finally, we impute the estimated income value for individuals whose incomes are missing but where their ages are observed. This approach has also been referred to as *predictive mean imputation.* Extensions of regression imputation now account for uncertainty in the prediction by adding a random residual drawn from a normal (0, 1) dstribution to account for possible regression-to-the-mean effects.

How might regression imputation with random residuals work for statistically matching PISA and TALIS? Consider two regression models for PISA and TALIS.

$$PISA: \qquad \mathbf{X} = \mathbf{Z}\boldsymbol{\beta}_{xz} + \mathbf{U}_P, \tag{1}$$

$$TALIS: \qquad \mathbf{Y} = \mathbf{Z}\boldsymbol{\beta}_{yz} + \mathbf{U}_T, \tag{2}$$

where, again, $\mathbf{X}$ and $\mathbf{Y}$ are variables unique to PISA and TALIS, respectively, and $\mathbf{Z}$ are the variable common to PISA and TALIS. The terms $\mathbf{U}_P$ and $\mathbf{U}_T$ are regression disturbance terms for the PISA and TALIS equations respectively.

With these questions in hand, statistical matching of PISA and TALIS would take place as follows (see Rässler, 2002, for details):

1. Estimate equations (1) and (2) using either ordinary least squares or maximum likelihood estimation to obtain $\hat{\boldsymbol{\beta}}_{xz}$ and $\hat{\boldsymbol{\beta}}_{yz}$

2. Using $\hat{\boldsymbol{\beta}}_{xz}$ and $\hat{\boldsymbol{\beta}}_{yz}$, obtain estimates of the residual matrices.

3. Create the following new regression models

$$PISA: \qquad \mathbf{X} = \mathbf{Z}\boldsymbol{\beta}_{xz.y} + \mathbf{Y}\boldsymbol{\beta}_{xy.z} + \mathbf{V}_P, \tag{3}$$

$$TALIS: \qquad \mathbf{Y} = \mathbf{Z}\boldsymbol{\beta}_{yz.x} + \mathbf{X}\boldsymbol{\beta}_{yx.z} + \mathbf{V}_T, \tag{4}$$

where $\boldsymbol{\beta}_{xz.y}$, $\boldsymbol{\beta}_{yz.x}$, and $\boldsymbol{\beta}_{xy.z}$ are partial regression coefficients.

4. Obtain estimates of these partial regression coefficients and calculate predicted values from the new regression equations

$$\begin{aligned} PISA: \quad & \hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\beta}}_{xz.y} + \mathbf{Y}\hat{\boldsymbol{\beta}}_{xy.z} && (5) \\ PISA: \quad & \hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}}_{yz.x} + \mathbf{X}\hat{\boldsymbol{\beta}}_{yx.z} && (6) \\ TALIS: \quad & \hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}}_{yz.x} + \mathbf{X}\hat{\boldsymbol{\beta}}_{yx.z}, \text{ and} && (7) \\ TALIS: \quad & \hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\beta}}_{xz.y} + \mathbf{Y}\hat{\boldsymbol{\beta}}_{xy.z}. && (8) \end{aligned}$$

5. Conditional residual variances can be obtained from equations (5) – (8), from which we can draw random residuals and add to our imputation model

$$\begin{aligned} PISA: \quad & \hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\beta}}_{xz.y} + \mathbf{Y}\hat{\boldsymbol{\beta}}_{xy.z} + \hat{\mathbf{V}}_P, \text{ and} && (9) \\ TALIS: \quad & \hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}}_{yz.x} + \mathbf{X}\hat{\boldsymbol{\beta}}_{yx.z} + \hat{\mathbf{V}}_T. && (10) \end{aligned}$$

Bayesian Approaches to Statistical Matching

The above approaches lie within the conventional frequentist framework of statistical matching. In what follows, we describe three approaches to statistical matching that rest on Bayesian inference.

Multiple Imputation. As seen in the regression imputation approach, a concern with imputing missing data is expressing the uncertainty in the missing data process. That is, imputing a single missing value for an individual and assuming that this is the value that would have been observed had it not been missing appears, on the surface, to be unrealistic. Multiple imputation is an approach that accounts for uncertainty in the imputation process. In contrast to conventional approaches to handling missing data, the multiple imputation approach usually results in wider confidence intervals around parameter estimates.

The general approach to multiple imputation proceeds as follows:

1. Again, let $\mathbf{X}$ be the variables unique to PISA, $\mathbf{Y}$ be the variables unique to TALIS, and $\mathbf{Z}$ be the variables common to PISA and TALIS.

2. Instead or imputing a single value for the missing data, multiple imputation yields k (say 5 or more) different values for each missing datum based on the posterior predictive distribution given the observed data. This yields k versions of the PISA and TALIS concantenated files.

3. Standard statistical procedures of interest are applied k times for the k different files, and parameter estimates are obtained by averaging over the k files. Standard errors for the parameter estimates are obtained as weighted combinations over the k files. Most major software programs such as *SPSS*, *SAS*, *STATA*, *Mplus*, and *R* allow for the analysis of multiple imputed data sets.

NIBAS. Another approach to addressing the statistical matching of PISA and TALIS is based on the regression imputation approach described earlier, but rests on Bayesian logic. It is referred to as *non-iterative Bayesian-based imputation* or *NIBAS* Rässler (2002). The NIBAS procedure can be outlined as follows:

1. The common variables $\mathbf{X}$ and $\mathbf{Y}$ are assumed to be at least univariate normal.[4]

2. Assume that a linear model holds as in equations (1) and (2).

3. Estimate equations (1) and (2) using either ordinary least squares or maximum likelihood estimation to obtain $\hat{\boldsymbol{\beta}}_{xz}$ and $\hat{\boldsymbol{\beta}}_{yz}$

4. Calculate sample residual covariance matrices $\mathbf{S}_P$ and $\mathbf{S}_T$.

5. Choose values for the correlation matrix $\mathbf{R}_{XY|Z}$. These can be somewhat

arbitrary.

6. Perform random draws for the parameters from their observed-data posterior distribution (see Rassler, 2003, for details).

7. Repeat the process k times to obtain k multiply imputed data sets and analyze via conventional statistical methods.

The NIBAS approach, is quite similar to multiple imputation. First, the observed-data posterior distribution is used to obtain values for the parameters. Second, the posterior predictive distribution is used to fill in missing data. In large samples, this approach will yield multiply imputed data sets for which standard statistical procedures can be applied.

Data Augmentation. The last approach we will consider and also based on Bayesian logic is referred to as *data augmentation.* Data augmentation utilizes a variant of the so call Gibbs-sampler which is popularly used in Markov chain Monte Carlo estimation. The procedure of statistical matching via data augmentation works as follows.

1. For PISA and TALIS separately, we obtain estimates of the missing data in each file, given the common variables and starting values for model parameters. This is referred to as the imputation step and is based on the predictive distribution of the missing data.

2. Given values of the missing data in PISA and TALIS separately, new parameter values are obtained from the complete posterior distribution.

Although these steps appear extremely similar to multiple imputation, the difference is that the imputation step and the posterior step yield a so-called Markov chain that can be iterated until it converges to a stationary distribution of both the missing values and the parameters. Issues regarding the convergence of the

Markov chain need to be considered.

Proposed Analyses and Approximate Timeline

Having outlined the various general methods of statistical matching, the intention of this study is to compare these approaches, and others, to the problem of matching PISA and TALIS. It is anticipated that this project will take approximately three person-months to complete. The proposed steps of analysis including a timeline given in person-days are as follows.

1. Download PISA and TALIS data for relevant countries. Analyses will compare the above statistical matching approaches on at least two relevant countries. (1 day)

2. Aggregate PISA and TALIS data to the school level and determine which school level variables are common to both PISA and TALIS. (1 day)

3. Conduct necessary scaling of common PISA and TALIS school variables in preparation for applying statistical matching methods. (4 days)

4. Conduct statistical matching methods – including (a) propensity score matching, (b) regression imputation with random residuals, (c) multiple imputation, (d) non-iterative Bayesian statistical matching, and (e) data augmentation. (30 days[5]).

5. Evaluate the quality of the matching procedures. Draw comparisons based on the analysis of matched data using descriptive and inferential statistical procedures. In the latter case, several relatively simple models, including regression and structural equation models will be specified. The goal is to compare inferences across statistical matching methods. (12 days)

6. Prepare report and submit to OECD secretariat. The project deliverables will include a complete report of publishable quality.[6] The completed report will

include annotated software code used for all analyses. (12 days)

References

D'Orazio, M., Di Zio, M., & Scanu, M. (2006). *Statistical matching: Theory and practice*. New York: Wiley.

Hansen, B. B. (2004). Full matching in an observational study of coaching for the SAT. *Journal of the American Statistical Association*, *99*, 609–618.

Hansen, B. B., & Klopfer, S. O. (2006). Optimal full matching and related designs via network flow. *Journal of Computational and Graphical Statistics*, *15*, 609–627.

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

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

Rassler, S. (2003). A non-iterative bayesian approach to statistical matching. *Statistica Neerlandica*, *57*, 58–74.

Rosenbaum, P. R. (1989). Optimal matching for observational studies. *Journal of the American Statistical Association*, *84*, 1024–1032.

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

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*, 581–592.

Footnotes

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

[2] Of course, within a data set, 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 within PISA or TALIS on variables in common are MAR.

[3] For countries in PISA that participated in the international teacher questionnaire, those data would need to be aggregated to the school level as well. However, this also provides for additional common variables on which to conduct matching.

[4] This is admittedly a heroic assumption. Transformations to normality are permitted.

[5] The exact amount of time taken for this step will depend primarily on the speed of the MCMC algorithms that are employed.

[6] It is my intention, contingent on agreements with the Secretariat, that the results of this report be published in a relevant methodological journal.