Athens Journal of Education - Volume 6, Issue 2 – Pages 145-158

Response to "An Examination of Plausible Score Correlation from the Trend in Mathematics and Science Study"

By Plamen Vladkov Mirazchiyski[*]

This article is a response to an article written by Wang and Ma "An Examination of Plausible Score Correlation from the Trend in Mathematics and Science Study", published in the Athens Journal of Education. The purpose of this paper is to address issues with Wang's and Ma's suggestion to use analysis method for correlating plausible values from TIMSS. The paper reviews the design of international large-scale assessments, and TIMSS in particular, and its implications for data analysis in regard to the application of Canonical Correlation Analysis for calculating association between two sets of plausible values, as proposed by Wang and Ma. The conclusion is that, given the design of TIMSS and other large-scale assessments, the method proposed by Wang and Ma is not appropriate for correlating two sets of plausible values because they are not multivariate measures as the suggested method would assume. Some other methodological issues related to the overall analysis approach used by Wang and Ma is discussed as well.

Keywords: correlation, international large-scale assessments, methodology, multiple imputation, plausible values

Introduction

This paper is a response article that addresses some issues with an article by Wang and Ma (2016) published in the *Athens Journal of Education*. The original article published by the aforementioned authors represents an attempt to address methodological issues when correlating two sets of plausible values (PVs) from the Trends in International Mathematics and Science Study (TIMSS) that the authors claim to exist.

The main argument of the authors is that if two sets of PVs are used in correlation analyses of TIMSS data with the current approach the study uses, this can inflate the chance for making Type I error due to the non-additive nature of the correlation coefficients. The solution of this assumed issue the authors suggest is Canonical Correlation Analyses (CCA) that can accommodate for the multidimensionality of the signals and avoid the dependency on the coordinate system in which the variables are described (Wang & Ma, 2016). However, most of the assumptions the authors have do not reflect the design and the methodology TIMSS, as well as other large-scale assessments (ILSA), as well as the analytical strategy and methods it uses to produce unbiased estimates from any statistical analysis being employed.

The next section provides a review of the literature on generation of PV sin

[*]Chief Executive Officer, Researcher, International Educational Research and Evaluation Institute (INERI), Slovenia.

https://doi.org/10.30958/aje.6-2-4 doi=10.30958/aje.6-2-4

general and as they are used in TIMSS (as well as in other ILSA), what is the necessity for their derivation as proficiency scores, what they are, how they are derived and how they shall be used in analysis reflecting their nature and statistical theory that stands behind their derivation. In addition, the literature review presents the sampling procedures and the derivation of the sampling weights which must be included in the computation of any estimates using ILSA data. This is necessary because Wang and Ma (2016) do not address the important issues of the unequal probabilities of selection and the subsequent issue of sampling variance when computing estimates with TIMSS data. The literature review ends with an overview of the CCA, its assumptions and application in situations where other correlation methods are not appropriate. The purpose of this review is to add more clarity on the design of TIMSS and ILSA in general, and thus, justify the current approaches in using them in analysis. These provide the background against which the claims in Wang's and Ma's (2016) article will be discussed in the last, discussion, section.

Literature Review

TIMSS is the successor of other ILSA conducted prior 1990s: the First International Mathematics Study (or FIMS, conducted in 1964), the First International Science Study (or FISS, conducted in 1970-1971). These studies, conducted by the International Association for the Evaluation of Educational Achievement (IEA) had a follow-up cycles in the period 1980-1984, the Second International Mathematics Study (SIMS) and the Second International Science Study (SISS) (IEA, n.d.). TIMSS, initially named as Third International Mathematics and Science Study, was the third cycle of both mathematics and science studies where they were conducted jointly (Mullis et al., 1997). In 1999 the IEA replicated TIMSS and named it as TIMSS-R or TIMSS-Repeat (Mullis et al., 2000), and due to the decision of conducting the study in regular cycles, later "third" was changed to "trends in".

TIMSS, as well as other studies, used the methodological developments originating in the National Assessment of Educational Progress (NAEP) conducted in US, extending these advancements (Rutkowski, Gonzalez, Joncas, & von Davier, 2010). Due to student fatigue, attrition and logistics of ILSA there was a need to find a solution to make possible carrying out assessment on a large scale to cope with the aforementioned limitations – not everyone can be tested using every single item. One such solution was the use of Multiple Matrix Sampling (MMS). MMS, used in many studies, is different than subjects or examinees sampling where the subjects (usually students) are selected from the population of interest. In MMS the measures on which the subjects are tested or surveyed on are sampled from a universe of interest, i.e. part (a sample) of the total assessment (Rutkowski, Gonzalez, von Davier, & Zhou, 2014). This facilitates the testing of the sampled subjects in broad content domains like mathematics or science where a large number of items are needed to have a reliable measure in the domain of interest. In addition, TIMSS has several content sub-domains in both mathematics

and science, but also cognitive sub domains in both areas. This would lead to estimated testing time as of more than 10 hours in TIMSS 2007, for example (Rutkowski et al., 2014). Thus, in TIMSS 2007 the total of 429 mathematics and science items were distributed in 14 blocks, each containing unique set of items, in each content domain (mathematics and science), rotated across 14 test booklets, each containing two mathematics and two science blocks. One of the mathematics blocks and one of the science blocks in each booklet was repeated in every next booklet, so that a link through common blocks across the booklets was ensured. This way, the testing time and logistic demands of the entire study are decreased, making it possible to conduct such a large study across a myriad of countries at the same time (Rutkowski et al., 2014). This way, no tested subject is taking every single item, but a sample of items which removes the burden from the examinees and is a cost-effective solution. Early developments of this technique have proven that group means appear to be more consistent than a sample of tested subjects taking all items. Currently there are different implementations of the MMS, one of them, also used in TIMSS is the Block Incomplete Booklet (BIB) set of designs (Rutkowski et al., 2014).

The use of MMS saves a lot of efforts and minimizes the testing time for the students participating in ILSA. However, this brings a serious challenge when estimating the student proficiency. Following the description of the MMS provided above,

> The relatively small number of items per block and the relatively small number of blocks per test booklet mean that the accuracy of measurement at the individual level of these assessments is considerably lower than is the level of accuracy common for individual tests used for diagnosis, tracking, and/or admission purposes(von Davier, Gonzalez, & Mislevy, 2009, p. 11).

Traditional methods for estimating the proficiency of tested subjects would yields "biased or inconsistent variance estimates of population parameters" (Rutkowski et al., 2010, p. 145). As stated previously, none of the sampled student takes all items, but (as for TIMSS 2007 example given above) two mathematics blocks and two science blocks containing unique items that do not appear in any other block; the students answer only the items presented to them and for the rest of the items student answers are missing by design. However, items differ in their characteristics, difficulty being the most important. Hence, percent correct for the items a student faced will not be the appropriate method because it limits comparability of results – the score will depend on the particular set of items a particular student receives (Mirazchiyski, 2013). The use of Item Response Theory (IRT) was more and more needed due to the use of MMS in assessments (Rutkowski et al., 2014). The traditional IRT approaches that use Marginal Maximum Likelihood (MML) and Expected a Posteriori (EP), however, are not appropriate solutions as well. These estimation techniques produce point estimates optimized for individual-level, but not group level estimation (von Davier et al., 2009). Several scholars developed group level models for estimating latent traits stemming from measurements using MMS. NAEP uses these models since its

assessment in 1983/84 to find a tractable solution for estimating the standard errors. The use of population models and their applications in cases of MMS together with IRT is commonly referred to "plausible values", although often other names are used as well (Rutkowski et al., 2014). The derivation of PVs from population models relies on Rubin's multiple imputation methods developed in the period between late 1970s to late 1980s. These models impute the tested subjects' scores for references on population level (Rutkowski et al., 2014).

Latent traits (such as intelligence or reading skills among many other) are not directly observable (von Davier, 2014) and in ILSA (and not only) are treated as missing (Rutkowski et al., 2010). Instead observing the latent traits directly, it is possible to observe the responses of examinees to tasks they face as indicators of these under lying traits. It is necessary to know, however, how these observable indicator variables relate to the latent trait. ILSA use IRT involving latent regression of the latent (unobservable) proficiency variable on number of predictor variables. This approach includes the information of all observable variables and follows the models of imputing data developed in the 1980s and 1990s. This means that the conditional distribution of the latent variable depends on the values of the observed variables, assuming that the missing values are missing at random (von Davier, 2014). As von Davier points out, "100% of the student proficiency data is imputed using a specialized imputation model based on statistical procedures that are tailored to incorporate both cognitive response data and student background data" (von Davier, 2014, p. 184). ILSA utilize latent regression models which provide Expected a Posteriori (EAP) estimates of posterior variance of the measured ability. These latent regression models are actually an extension of the multiple group IRT model. What they provide is a "different conditional prior distribution for each respondent's proficiency based on a set of predictor variables" (von Davier, 2014, p. 184). Although the total number of achievement items in each ILSA is large, it is still limited for each student due to MMS. Thus, it is complemented with the items from the background questionnaires, applying Rubin's multiple imputation approach to impute the answers of the items the student did not face and create student ability distribution for the entire population or sub-populations of interest (Rutkowski et al., 2010). It may not be immediately obvious, but PVs "add exactly the right amount of variability to make the distribution of the PVs in the group match the distribution of the true values in the group" (von Davier et al., 2009, p. 35). The foundations of the PVs methodology, its theoretical rationale, foundations and mathematical proof are laid out in Mislevy (1991) and Mislevy, Beaton, Kaplan and Sheehan (1992). An overview of what PVs are for the non-technical reader is provided by von Davier et al. (2009).

When it comes to the actual application of the population modeling and imputing the missing data along with the latent regression models used in ILSA, each study has its own specifics, although many things in common as well. The presentation here continues with a description of the TIMSS 1995 proficiency scaling methodology because this is the study and cycle the authors of the original article (Wang & Ma, 2016) used. The subsequent cycles of TIMSS use the same approach and steps for scaling the cognitive data, although some details may differ.

In TIMSS 1995 first a subsample of 600 students is drawn from each country, forming an "international calibration samples" that are about equal in number of selected students. These samples were drawn systematically with Probability Proportional to the Size (PPS) of the Primary Selection Units (PSU) (i.e. schools) using their overall weights as measure of size. This led to equal selection probabilities within the national samples to draw the calibration samples within each country, thus each country was given an equal weight in estimating the item parameters in the next step(Adams, Wu, & Macaskill, 1997). The scaling model applied was a generalization of the more basic unidimensional model. In addition, a multivariate linear model imposed on the population distribution. The item parameters were estimated using the international sample. Then the model was fit in each country using fixed item parameters obtained in the previous step (Adams et al., 1997). The population model uses the item response model which is a conditional model describing the process of generating the responses conditional on the latent variable. The derivation of conditioning variables is done using background information to form a vector included in the latent regression model as a predictor. Then the five PVs are derived for each student by making five random draws from the formed marginal posterior latent distribution (Adams et al., 1997). The chapter on the scaling methodology and procedures in the TIMSS 1995 Technical Report (Adams et al., 1997) provides more comprehensive technical and detailed information. The corresponding chapter in the TIMSS 2007 Technical Report (Foy, Galia, & Li, 2008) provides more reader-friendly description of the scaling procedures.

The important detail from the review of the methodology and procedures of obtaining the PVs as proficiency scores that has to be stressed is that a set of PVs for each subject area (e.g. overall science) or content sub-domain (e.g. chemistry) or cognitive sub-domain (e.g. applying) is a set of variables representing unidimensional measure of the same construct of interest. As explained above, a set of PVs represent five random draws from a marginal posterior latent distribution. That is, each one of the PVs in a set (e.g. overall science score) represents a measure on the same construct that includes information from the same cognitive items and the same background variables that are used to construct the marginal posterior distribution they are drawn from.

As stated earlier, the procedures of obtaining PVs follow the theory of multiple imputation using the information from the background variables as predictors in a latent regression model. Rubin (1987) provides a theoretically and methodologically sound background on the imputation techniques, models and analysis with the resulting imputed data sets. As Mislevy (1993) notes, Rubin’s approach of imputing missing data multiple times creates data sets where "each missing variable is replaced by a draw from its predictive distribution, conditional on the observed data" (Mislevy, 1993, p. 79). The generation of PVs follows the same logic, as already described in the previous paragraphs. The methods for analysis of imputed data sets were specified and complemented in subsequent publications (e.g. Little & Rubin, 1987, 2002). The five randomly drawn PVs for each student vary in their values as a result of the multiple imputation. When it comes to analysis of PVs, five estimates of any statistics are computed with each

of the five PVs (or any measure that has been imputed multiple times) and they are all different. This is a result of what is called "imputation variance" or "imputation error" (Foy et al., 2008) which reflects the measurement error stemming from the use of MMS (Foy et al., 2008). All analyses that include PVs have to follow the approach of performing analyses with multiple imputation variables. Rubin (1987) and (Little & Rubin, 1987, 2002) provide a set of rules in order to combine the parameter estimates and compute the variance associated with the multiple imputation of measures. These rules of combining the estimates and compute the imputation variance have found strong empirical support in various studies as in Schafer (1999) and in the case of PVs in various papers, as in von Davier et al. (2009) and Rutkowski et al. (Rutkowski et al., 2010). Following the theoretical developments of Rubin (1987) and Little an Rubin (1987, 2002), any analysis of TIMSS 1995 involving PVs will perform the computations five times (once with each PV) and the results of these computations will be averaged to obtain an unbiased estimate of student performance (Gonzalez, 1997). Formula for computing the imputation variance in TIMSS and ILSA in general are provided by Foy et al. (2008) and (von Davier et al., 2009), to name just few, and the technical report of each ILSA provides such formulas reflecting the specifics of the study. The same approach is used not only in TIMSS, but in other ILSA, such as PIRLS, ICCS, ICILS and PISA, and are implemented in statistical software which will be discussed later in the paper.

The imputation variance, however, is not the only source of error in ILSA. Besides the sample of items from the universe of all possible items that can measure given construct, ILSA also use sample of students from the target population for which the construct is measured. The sampling design of TIMSS 1995 is a two-stage stratified cluster sampling design and is done separately for each population of interest. In the first stage in each participating country 150 schools where students in a particular population of interest study are sampled with PPS. Schools are the PSUs. The second stage of sampling picks intact classrooms within the sampled schools (clusters). Usually one intact classroom is sampled, although some participated countries preferred two. In some countries a third stage (sampling students within the selected classrooms) was added, but these were exceptional cases. Due to the clustering effect of selecting intact classrooms (students in the same classroom tend to be more alike), intraclass correlation (ICC, a measure of similarity within a cluster) and the size of the classroom, along with the desired standard error from the sample, were taken into account when calculating the desired sample size within each country (Foy, Rust, & Schleicher, 1996). A stratification (grouping of schools according an attribute) was applied in most countries to improve sampling efficiency, making estimates more reliable, to apply different sample design to specific groups of schools or regions, and to ensure adequate representation of specific groups in the target population. Explicit stratification would construct independent lists of schools on an attribute. Implicit stratification would use the same list of schools where schools are sorted by the attribute (Foy et al., 1996). In the first stage of sampling, schools in the sampling frame (or frames, in case of explicit stratification) are sorted by their measure of size (MOS) (number of students in the target population). A sampling interval is

defined by dividing MOS by the number of schools to be sampled. The first school is selected by choosing a random number between 1 and the number representing the sampling interval. Thus, the number obtained represents the MOS of the first school being selected. Adding the sampling interval to this number would give the number of the second sampled school, and so on. If an implicit stratification is applied, the sampling of schools will reflect the implicit strata within which the schools are sorted by their MOS too. At the second sampling stage one or two (depending on the countries preferences) intact classes were sampled (Foy et al., 1996).

Not only TIMSS, but all other ILSA as well, follow the same or similar sampling strategy. This kind of sampling (PPS) is rather different than the Simple Random Sampling (SRS). The difference is that with SRS every student is selected with probability that is equal to the probability of selecting any other student (Rutkowski et al., 2010). On the contrary, as presented above, in TIMSS (as well as all other ILSA), do not use SRS. The application of PPS sampling means that each sampled school, hence classroom and student within it, are sampled with different probability that depends on the number of students in the target grade attending the schools. Then different students in the sample will not represent the same number of students in the population they were selected from. In addition, different countries chose different stratification variables in the sampling process to satisfy their research demands (Foy, 1997). This is an additional challenge for analyzing data that stems from TIMSS or other ILSA. Sampling weights in ILSA are introduced to accommodate for the sampling with different probabilities to ensure that certain groups in the population of interests are not overrepresented in the sample (Rutkowski et al., 2010). The TIMSS 1995 weighs calculation was done in three steps. First, calculating the school weights, adjusting for the school non-response independently for each design domain or explicit stratum. Second, calculating the classroom weight adjusting for the non-response of the classroom. When only one classroom was sampled, no classroom adjustment was necessary. Third, computing the sampling weight of the participating students adjusting for their non-response. The final weight for each student was added as product of the three intermediate weights from the previous steps. The weights computed in the first three steps before their non-response adjustments are computed as the inverse probabilities of selection (for the school, class and student) (Foy, 1997).

The TIMSS, or any other ILSA, sampling design provides country samples in ILSA representative for the population they have been drawn from (Foy, 1997). Not using the sampling weights in analyzing ILSA data leads to giving more importance on some students due to the sampling design. Relevant example in this regard is provided by Rutkowski et al. (2010) who demonstrate how the sampling of students from different school types biases the results on population level when weights are not used to adjust for the number of students in the population each one of the sampled students represents.

An important issue in ILSA is the variance estimation to compute the standard errors due to the stratified multi-stage sampling. The standard procedures and formulas for computing the standard errors do not apply to ILSA because the

sampling strategy does not rely on SRS. The IEA and OECD studies rely on replication techniques to estimate the variance. TIMSS, in particular, uses Jackknife Repeated Replication (JRR). When analyzing ILSA data, replication techniques should be used for all sampling variability estimates, which are the sampling errors, to obtain unbiased estimates (Rutkowski et al., 2010). The necessity of using replication methods for sampling variance estimation stems from the sampling strategy that uses unequal probability to sample schools teaching students in a target population to obtain efficient and cost-effective samples. The JRR variation in TIMSS 1995 assumes that PSUs (schools) can be paired according to the sampling design, forming pseudo-stratums (pairs of schools) to estimate the sampling variance. This approach appropriately reflects the combined effect of the within- and between-school contribution to the sampling variance (Gonzalez & Foy, 1997). The procedure is as follow. First, sampling zones (paired schools) are formed. The sampling zones are formed in the same order in which the schools were sampled. With 150 sampled schools, 75 zones are formed. When more schools are sampled, sometimes schools were combined before forming a sampling zone. Second, the variance is estimated in each sampling zone by setting the weight of one of the paired schools to 0 and doubling the weight of the other school. The estimation is done 75 times plus once with the full weight. At the end, the variance is estimated by combining the results. More details and the formula for combining the estimates to produce the sampling variation of an estimate is provided by Gonzalez and Foy (1997). This procedure is rather different than standard methods of estimating the error under SRS. Rutkowski et al. (2010) provide a clear example what are the consequences when replication is not applied using TIMSS 2007 data.

When using PVs, each estimate is computed five times (once with each PV) and within each JRR zone. The standard error of an estimate using PVs is computed using both the sampling and imputation variance components. Both of these components are important, omitting any of them can produce biased result. Pedagogical examples are provided by Rutkowski et al. (2010). Formulas for combining the estimated imputation and sampling variance, as well as the total standard error of an estimate, are provided for each ILSA reflecting the specifics of the study and even the study's cycle. Such can be found in Foy et al. (2008) and Schulz (2011), for example.

The systematic publications on correlating two sets of variables begin in 1936 with a publication of Hotelling (1936), although some elements were developed earlier by Bravais, Galton, Pearson, Yule and others (Hotelling, 1936). Hotelling (1936) was concerned with issues in correlating two sets of variables representing multidimensional measures and suggests the name "canonical correlations". Later, this kind of analysis was referred to mainly as Canonical Correlation Analysis (CCA). Borga (2001), also cited by Wang and Ma (2016), defines CCA as "a way of measuring the linear relationship between two multidimensional variables" (Borga, 2001, p. 2), that is, each set of variables would represent different measures, directly observable or not. Härdle and Simar (2007) define it as technique for analyzing the association of two data sets based on projections where "an index (projected multivariate variable) that maximally correlates with the

index of the other variable" (Härdle & Simar, 2007, p. 321). Similarly, Borga (2001) defines the CCA as "finding two sets of basis vectors, one for **x** and the other for **y**, such that the correlations between the projections of the variables onto these basis vectors are mutually maximized" (Borga, 2001, p. 2). Thus, CCA "is based on linear indices, i.e., linear combinations of the random variables" (Härdle & Simar, 2007, p. 321). The most important thing to note about CCA is that it is a method of correlating two multidimensional sets of variables, i.e. the variables in each set are measures that quantify different properties of the objects. The next section will reiterate this when discussing the CCA as a method for correlating PVs in TIMSS, suggested by Wang and Ma (2016).

Discussion

Wang and Ma (2016) raised concerns about the calculation of correlation coefficients as computed in TIMSS when two sets of PVs are used. The concerns were raised due to the non-additive nature of correlation coefficients and suggested the use of canonical correlation instead to reduce the risk of making Type I error. However, as already clarified in the literature review, a set of Plausible Values (PVs) does not contain multiple different measures on multiple different latent traits as CCA would assume. The important details that needs to be reiterated here is that set of five PVs in any subject area, content or cognitive subdomain is derived as five random draws from the same marginal posterior distribution (Adams et al., 1997), i.e. each PV in a set of PVs (e.g. overall mathematics) represents the same measure, carrying out the information obtained using the same items, their IRT parameters, conditioned on the same background variables. Hence, it is the same trait presented as five different variables (PVs), and not five different latent traits in five different variables. Being imputed variables, all rules that apply to analysis of multiply imputed data sets (see Little & Rubin, 2002) apply when working with PVs regardless of the analysis type. As Mislevy (1993) notes, this situation resembles the situation of having multiple unbiased and conditionally independent indicator variables, but only on the surface. Further, he demonstrates the pitfalls of using PVs as unbiased and conditionally independent indicators of a latent variable, producing incorrect estimates. Given the answer of the third research question Wang and Ma (2016) have, the correct approach for correlating the two sets of mathematics and science achievement PVs would be to correlate the first PV in mathematics with the first PV in science, the second PV in mathematics with the second PV in science, and so on, then averaging the obtained estimates to derive the final estimate of the correlation. The computation of the sampling and imputation variance and the final standard error of this coefficient would follow formulas which can be found in Foy et al. (2008). This kind of pairing will not allow underestimation of the correlation between the subjects. There are software products that are capable to do this tedious work with minimal effort from the side of the analyst. One such product is the IDB Analyzer, freely available from the IEA (IEA, 2016) where the correlation of two sets of PVs follows the routine described above. The standard error, in turn, is necessary to test

the statistical significance of the estimates. Another issue that deserves attention in regard to the sampling variance as component of the standard error is that there is no indication if the authors used any weights. As mentioned above, the sampling weights must be used to properly estimate the sampling variance. This issue is discussed further.

The aforementioned issue is the main issue with Wang's and Ma's (2016) publication: CCA is not the appropriate method for analyzing ILSA data stemming from any study that uses the PVs methodology because each PV represents the same unidimensional latent trait, and not multiple dimensions of a latent trait or multiple traits as CCA assumes. Information on what PVs are and how to use them in analysis of student achievement, along with the literature sources was already provided by the literature review.

There are other points of concern with the Wang's and Ma's (2016) article, as outlined below.

First, on page 305 the authors provide a citation from Garcia (2010) saying that "One cannot add raw r values to compute an arithmetic average an r" (Garcia, 2010, p. 2), but missed important information located few lines below where the author of the original publication adds that,

> "in order to compute $\bar{r}$ [that is, average correlation] individual r values have to be converted into additive quantities. Several techniques, each with their own assumptions and drawbacks, can be used: transcendental transformations, numerical expansions, weighted averages, or combination of techniques" (Garcia, 2010, p. 2).

Further, the authors added a citation from Statsoft (2000) to reconfirm that,

> "Because the value of the correlation coefficient is not a linear function of the magnitude of the relation between the variables, correlation coefficients cannot simply be averaged" (StatSoft, 2000, p. 10).

It is very strange, however, that the authors missed the next sentence that says that,

> "In cases when you need to average correlations, they first have to be converted into additive measures. For example, before averaging, you can square them to obtain coefficients of determination which are additive (as explained before in this section), or convert them into so-called Fisher z values, which are also additive." (StatSoft, 2000, p. 10).

There is one important thing to note in the above citations, as well as with the sources as a whole: they consider a general case and do not pertain to analyzing multiple imputation data sets.

Second, on page 307 the authors wrote that "After completing canonical correlation analyses, the results are merged with mathematics and science performance scores at the country level to address Question 3. The combined data set is attached in Appendix 1."

Unfortunately, they do not provide any detail on how the CCA results were computed and merged to the TIMSS 1995 mathematics and science scores. The methodology of extracting information from variables to add it later to the same variables shall be explained: 1) why was it done, what was the purpose; and 2) how was it done. Also, besides the data, the aforementioned appendix should contain combined data and tables with the canonical correlations, as mentioned on page 309. However, there is no appendix to this article and these are not possible to inspect.

Third, on page 307 the authors wrote that "The inclusion of canonical correlation as a predictor automatically assumes co-existence of mathematics achievement as an explanatory variable". It is quite unclear what is meant by this, especially how a statistical method can be included as a predictor. Probably they meant that the results of CCA were included as a predictor. However, as mentioned above, there is no information how the CCA results were merged with the scores.

Fourth, on page 307 the authors added SPSS syntax to compute regression analysis, noting that it is a "simple SPSS application without involvement of complex Macro syntax on the variable dimension" (Wang & Ma, 2016, p. 309). As mentioned above, it is not clear how the canonical correlations were computed and included in the analysis this syntax uses. The provided syntaxis not only simple, it is oversimplified and does not take into account any of the design issues TIMSS 1995 has, as the JACKREGPV. SPSmacro provided by Foy, Arora and Stanco (2013) and recommended by Statistics Canada (2002) does. The syntax provided by Wang and Ma (2016) does not include any statement that weights the data using any of the weighting variables available in the TIMSS 1995 data and do not use the variance estimation methods as used in TIMSS (see Foy et al., 2008) to estimate the sampling variance and take it into account when computing the standard errors. In the discussion part of the paper, on page 309, the authors mention for the first time that the TIMSS uses multistage sampling. TIMSS, as well as other ILSA, uses sampling with probability proportional to the size of the schools. Thus, schools, and their students respectively, are sampled with different probabilities. To be able to produce estimates on population level, which is the purpose of TIMSS and any other ILSA, weights that account for the unequal probability of selection must be used. Due to the PPS sampling, if no weights are used, some students can have a disproportional impact on the estimates and the analysis can provide rather biased results, as demonstrated by Rutkowski et al. (2010). An additional issue is that with large samples, as in TIMSS, not using appropriate methods of variance estimation will underestimate the SE. This is probably the case with the strong and significant results the authors obtained in their study, not using any weight and their replication in the analysis. Given all of the above, the conclusion the authors make on page 309 that "the influence from complex sampling is washed out" (Wang & Ma, 2016, p. 309) along with the arguments is incorrect simply because weighting matters for all types of statistical estimates with ILSA data, including TIMSS. As per the use of design effect itself for estimating the sampling variance, it is largely discouraged in the recent years. There are different methods to compute the standard errors due to the clustering

effect of the sampling used in large-scale assessments, one of them being the design effect. It is an estimate of how large the effects of dependency among observations are, including clustering, on the sampling variance. Despite the ease of calculating design effects, they appear as rather crude estimates of the standard errors which actually appear to be inflated (Barron, 2000) and conservative compared to other methods (Stapleton, 2008). In addition, an empirical study conducted by Stapleton (2008) concludes that design effect are only appropriate for univariate statistics, when applied accurately. In contrast, using jackknifing to compute standard errors is much more precise (Barron, 2000). As Barron (2000) notes, the computation of jackknife estimates of standard errors has been difficult for secondary analysts. However, in recent years computers have become more powerful and software for using jackknifing is available.

Conclusions

The method for correlating sets of PVs suggested by Wang and Ma (2016) is an attempt to solve an alleged problem in analyzing TIMSS 1995 data. However, as this paper demonstrates, the authors do not take into account the assessment design of TIMSS 1995. This resulted in recommendation of analysis method that has assumptions and applications not relevant to analysis of PVs due to their specifics. Further, the actual application of the method ignores another important design issue, the complex sampling design of TIMSS 1995, and the necessary use of sampling weights. This is proven to have negative effects when conducting analyses with ILSA data regardless of the analysis method.

The contemporary ILSA are tools for policy making in education. The decisions made from analysis results have an impact on the on the implementation of policies and reforms in education. It is a great responsibility of researchers using these data to apply appropriate analysis methods, taking into consideration the study design and nature of the measures. Otherwise, biased results presented to policy makers may lead to ineffective policies.

References

Adams, R. J., Wu, M. L., & Macaskill, G. (1997). Scaling Methodology and Procedures for the Mathematics and Science Scales. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study: Technical Report* Vol. 2, pp. 111–146. Chestnut Hill, MA: Boston College.

Barron, S. (2000). Difficulties Associated with Secondary Analysis of NAEP Data. In N. S. Raju, J. W. Pellegrino, M. W. Bertenthal, K. J. Mitchell, L. R. Jones, & National Research Council (U.S.) (Eds.), *Grading the Nation's Report Card: Research from the Evaluation of NAEP*, pp. 172–194. Washington, D.C: National Academy Press.

Borga, M. (2001, January 12). Canonical Correlation: A Tutorial. Retrieved from https://bit.ly/2OfnsX0.

Foy, P. (1997). Calculation of Sampling Weights. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study: Technical Report* Vol. 2, pp.

71–79.
Foy, P., Arora, A., & Stanco, G. M. (Eds.). (2013). *TIMSS 2011 User Guide for the International Database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College and the IEA.
Foy, P., Galia, J., & Li, I. (2008). Scaling the Data from the TIMSS 2007 Mathematics and Science Assessments. In J. F. Olson, M. O. Martin, & I. V. Mullis (Eds.), *TIMSS 2007 Technical Report* , pp. 225–280. Chestnut Hill, MA: Boston College.
Foy, P., Rust, K., & Schleicher, A. (1996). Sample Design. In M. O. Martin & D. L. Kelly (Eds.), Vol. 1, pp. 4-1-4–17. Chestnut Hill, MA: Boston College.
Garcia, E. (2010). *A tutorial on correlation coefficients*. Retrieved from https://bit.ly/2Q6 SYbC.
Gonzalez, E. J. (1997). Reporting Student Achievement in Mathematics and Science. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study: Technical Report*, Vol. 2, pp. 147–174. Chestnunt Hill, MA: Boston College.
Gonzalez, E. J., & Foy, P. (1997). Estimation of Sampling Variability, Design Effects, and Effective Sample Sizes. In M. O. Martin & D. L. Kelly (Eds.), *Third International Mathematics and Science Study: Technical Report*, Vol. 2, pp. 81–100. Chestnut Hill, MA: Boston College.
Härdle, W., & Simar, L. (2007). *Applied multivariate statistical analysis* (2nd ed.). Berlin: Springer.
Hotelling, H. (1936). Relations between Two Sets of Variates. *Biometrika, 28*(3/4), 321–377.
IEA. (2016). *IEA IDB Analyzer* (Version 4.0) [Computer software]. Hamburg, Germany: IEA.
IEA. (n.d.). *Other IEA studies.* Retrieved January 3, 2018, from https://bit.ly/2OeApRa.
Little, R. J. A., & Rubin, D. B. (1987). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.
Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.
Mirazchiyski, P. (2013). *Providing School-Level Reports from International Large-Scale Assessments: Methodological Considerations, Limitations, and Possible Solutions*. Amsterdam: International Association for the Evaluation of Educational Achievement.
Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, (56), 177–196.
Mislevy, R. J. (1993). Should "Multiple Imputations" be Treated as "Multiple Indicators"? *Psychometrika*, *58*(1), 79–85.
Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161.
Mullis, I. V. S., Martin, M. O., Beaton, A. E., Gonzalez, E. J., Kelly, D. L., & Smith, T. A. (1997). *Mathematics Achievement in the Primary School Years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS International Study Center, Boston College.
Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., Gregory, K. D., Garden, R. A., O'Connor, K. M., … Smith, T. A. (2000). *TIMSS 1999 International Mathematics Rreport: Findings from IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade*. Chestnut Hill, MA: International Study Center, Boston College, Lynch School of Education.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International Large-

Scale Assessment Data: Issues in Secondary Analysis and Reporting. *Educational Researcher, 39*(2), 142–151.

Rutkowski, L., Gonzalez, E., von Davier, M., & Zhou, Y. (2014). Assessment Design for International Large-Scale Assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessments: Background, Technical Issues, and Methods of Data Analysis*, pp. 75–96. Boca Raton, FL: CRC Press.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research, 8*, 3–15.

Schulz, W. (2011). The reporting of ICCS results. In W. Schulz, J. Ainley, & J. Fraillon (Eds.), *ICCS 2009 Technical Report*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Stapleton, L. M. (2008). Analysis of data from complex surveys. In E. D. Leeuw, J. J. Hox, & D. A. Dillman (Eds.), *International Handbook of Survey Methodology*, pp. 342–369. New York, NY: Lawrence Erlbaum Associates.

Statistis Canada. (2002). *The International Adult Literacy and Skills Survey, 2003*. Ottawa, ON: Statistics Canada.

StatSoft. (2000). Basic Statistics. Retrieved August 20, 2017, from https://bit.ly/2AyXA4Q.

von Davier, M. (2014). Imputing Proficiency Data under Planned Missingness in Population Models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of International Large-Scale Assessments: Background, Technical Issues, and Methods of Data Analysis*, pp. 175–202. Boca Raton, FL: CRC Press.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series, 2*(1), 9–36.

Wang, J., & Ma, X. (2016). An Examination of Plausible Score Correlation from the Trend in Mathematics and Science Study. *Athens Journal of Education, 3*(4), 301–312.