Athens Journal of Education - Volume 6, Issue 2 – Pages 159-170

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

By Jianjun Wang[*] & Xin Ma[†]

This rejoinder keeps the original focus on statistical computing pertaining to the correlation of student achievement between mathematics and science from the Trend in Mathematics and Science Study (TIMSS). Albeit the availability of student performance data in TIMSS and the emphasis of the inter-subject connection in the Next Generation Science Standards in the U.S., the major TIMSS reports did not consistently include the correlational findings of student achievements, nor did the TIMSS macro, JACKREGPV per Mirazchiyski's recommendation, meet the need of correlation computing between each pair of plausible values (PV) in mathematics and science. In addition, an inflation of Type I error would be inevitable unless a feasible approach is taken to concurrently correlate the two sets of PVs from each subject. In this context, we adduced canonical correlation analysis (CCA) as a feasible approach that included considerations of both sampling weight and standard error adjustment due to complex sampling. More importantly, this rejoinder reveals two fundamental issues in the TIMSS literature: (1) Incompliance to the multiple imputation rule in reporting the TIMSS 1995 international findings (see Beaton et al., 1996a, b); (2) Need of substantially increasing the number of PV imputations as illustrated by the mainstream statistical computing in general (e.g., SAS Institute, 2015; STATA, 2017), and the recent international large scale assessments (ILSA) in particular.

Keywords: International Studies, TIMSS Data Analysis

It is our great honor to contribute this rejoinder to an article of our esteemed colleague, Plamen Vladkov Mirazchiyski, the former Deputy Head of the Research and Analysis Unit of the International Association for the Evaluation of Educational Achievement (IEA). We appreciate his effort on writing an extensive response (we shall brief his article name as "response") that nearly doubled the space of our original article. We sincerely accept his criticism on the briefness of our report that was originally designed for a 20-minute presentation at the ninth Annual International Conference on Statistics (Athens, Greece) in June, 2015.

In this rejoinder, we are given the opportunity to elaborate our research with a clear focus on the potential disagreements we seemed to have with Mirazchiyski. To keep it simple, we categorized this article into four parts. First, we will analyze the core issues that originated this track of in-depth inquiry. In the second section, we clarify what is not the issue. The third section is devoted to coverage of new literature beyond the foundation of multiple imputation method Mirazchiyski

[*]Professor, California State University, Bakersfield, USA.
[†]Professor, Department of Educational, School, and Counseling Psychology, University of Kentucky, USA.

https://doi.org/10.30958/aje.6-2-5 doi=10.30958/aje.6-2-5

reviewed from the period of 1970s – 1990s. We conclude this rejoinder in the last section to support the momentum of knowledge advancement in comparative education.

What Is the Issue?

Mirazchiyski raised issues on several fronts, including (1) the incorporation of sampling weights and the adjustment of result variance due to stratified, cluster sampling, (2) the ignorance of using TIMSS tools, particularly the macro of JACKREGPV.SPS, in the statistical computing that involves plausible values, (3) the use of design effect that, albeit its candidacy as a viable method for handling data with complex sampling in many decades (Kish, 1965), might produce crude estimates of the standard errors, (4) the inappropriateness of using the canonical correlation method in our original work for computing correlation coefficients.

All these issues appear to be handled with a simplistic question on what works. While it is still legitimate, a more precise question should be on "what works in which context." In reviewing the research context of the Trend in Mathematics and Science Study (TIMSS), we appreciate Mirazchiyski's attention on other [international] large scale assessments (ILSA)[1], such as the First International Mathematics Study (or FIMS, conducted in 1964), the First International Science Study (or FISS, conducted in 1970-1971), the Second International Mathematics Study (SIMS), and the Second International Science Study (SISS) in the 1980s. He also noted the National Assessment of Educational Progress (NAEP) conducted in the United States, which is not an international study. However, it was extended in the two phases of International Assessment of Educational Progress (IAEP) prior to TIMSS. Further, the NAEP method was largely adopted by TIMSS.

Throughout his response, Mirazchiyski did not pay any attention to the context so as to contrast TIMSS with other studies. One special feature of TIMSS is the opportunity of analyzing the correlation of student performance between mathematics and science that never existed in the past ILSA. Although FIMS, FISS, SIMS, SISS, and IAEP covered the subjects of mathematics and science, these assessments were conducted on different student samples, and thus, no attempt would be appropriate to compute correlation coefficients between mathematics and science scores.

Indeed, the inter-subject connection is important. For instance, the Next Generation Science Standards (NGSS) indicated strong needs to connect mathematics and science education in the United States (National Research Council, 2014). Similar inquiry-based interdisciplinary learning was emphasized in the previous version of the U.S. curriculum standards since the mid 1990s. TIMSS is unique and unprecedented for surveying mathematics and science performance from the same group of students concurrently.

Despite its 20-year history from TIMSS 1995 to TIMSS 2015, major TIMSS

[1]We took the liberty to add "International" because it was missing in Dr. Mirazchiyski's acronym definition.

reports are divided according to the subject boundary, one for mathematics (Beaton et al., 1996a; Mullis et al., 2000; Mullis, Martin, Gonzalez, & Chrostowski, 2004; Mullis, I.V.S., Martin, M.O., & Foy, 2008; Mullis, Martin, Foy, & Arora, 2012a; Mullis, Martin, Foy, & Hooper, 2016a) and the other for science (Beaton et al., 1996b; Martin et al., 2000; Martin, Mullis, Gonzalez, & Chrostowski, 2004; Martin, Mullis, & Foy, 2008; Mullis, Martin, Foy, & Arora, 2012b; Mullis, Martin, Foy, & Hooper, 2016b), as if there were no connections between them.

More importantly, TIMSS methods, particularly the JACKREGPV.SPS macro adduced by Mirazchiyski, are not prepared for supporting the correlation analyses of student achievement between mathematics and science. Because TIMSS did not produce a macro for correlation computing, it seems natural to first obtain the coefficient of determination (R^2) from a simple regression that contains one dependent variable and one independent variable. The correlation coefficient can be subsequently computed from the square root of the R^2 result. In his Discussion section, Mirazchiyski dictated,

> [T]he 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. (p. 14)

which concurred the paring approach of using one plausible value (PV) from mathematics and another PV from science. For five PVs in each subject, five rounds of repetitions are needed prior the result aggregation.

Nonetheless, to invoke the JACKREGPV.SPS macro from the TIMSS tools, researchers are expected to concurrently use all five plausible values (see NPV=5 below), instead of one PV from each subject[2]. As Foy, Arora, and Stanco (2013) acknowledged, "[t]he JACKREGP macro [sic] is used to perform a multiple linear regression between *a set of plausible values as the dependent variables* [italics added for emphasis] and a set of independent variables" (p. 52), which does not work for paring the PV for correlation computing in Mirazchiyski's afore-quoted suggestion.

[2]The screenshot here comes from page 163 of https://bit.ly/1sFVLw9.

```
Include "c:\jackregpv.sps".

Jackregpv
                    Infile   = temp                /
                    Cvar     = cntrid              /
                    Xvar     = regsex              /
                    Rootpv   = Prose               /
                    NPV      = 5                   /
                    Njkz     = 30                  /
                    Rpwt     = replic01 to replic30 /
                    Wgt      = popwt.
```

We further question Mirazchiyski's claim of support for JACKREGPV or JACKREGP [sic] application from Foy, Arora, and Stanco (2013) and Statistics Canada (2002). Although TIMSS has made a convention of using the macro with PV in the name for analyses that involve plausible values (Foy, Arora, & Stanco, 2013), the macro that can be used for the simple regression analysis with one PV as the dependent variable is JACKREG.SPS. In other words, the convention of Foy, Arora, and Stanco (2013) does not apply here when we handle correlation analyses with a pair of PVs, instead of all five PVs from both sides of mathematics and science. For that reason, the computation cannot be supported by the PV macros, such as JACKPV and JACKREGPV, from TIMSS.

In summary, the main issue seemed to hinge on the potential oversight of TIMSS colleagues on the unique opportunity to correlate student performance between mathematics and science that never occurred in previous ILSA projects. Although 20 years have lapsed between TIMSS 1995 and TIMSS 2015, no major reports were disseminated on the correlation outcomes, nor did the TIMSS team clarify the confusion on which macro to revoke for the computing. By choice, we picked TIMSS 1995 data in our study to help track the ongoing disengagement of TIMSS reporting on the correlation part since its inception.

What Is Not the Issue?

The previous section addressed the issues that are important to us but was overlooked in Mirazchiyski's response. In this section, we intend to reciprocally discuss topics that have been considered as issues by him, but not us.

First, we agree that TIMSS data analyses need to consider both sampling weight due to unequal rates of sample selection and variance inflation for using a stratified, cluster sample structure. We also agree that canonical correlation analysis (CCA) is a method not solely developed to serve TIMSS researchers. For the sampling weight part, CCA does not have the issue because standard statistical software packages such as SAS allow incorporation of sampling weight in computing the coefficient from canonical correlation[3]. We were curious on why Mirazchiyski argued that our method "ignores another important design issue, the

[3]p. 11 of https://bit.ly/2StobHJ.

complex sampling design of TIMSS 1995, and the necessary use of sampling weights" (p. 18).

The adjustment of variance to address the impact of complex sampling also needs to be done. However, it can be considered in multiple ways. For instance, SAS PROC SURVEYMEANS offered both design effect and Jackknife options for researchers to choose because supporters can be found in either aisle[4]. While the variance could be inflated by a factor of k (the design size) due to complex sampling (Kish, 1965), our article also clarified the cancellation of the k factor in correlation computing because of its involvement at both numerator and denominator of the formula. Mirazchiyski did not dispute our special attention for the correlation computing. Instead, he seemed to come from the Jackknife aisles of practitioners and wrote: "As per the use of design effect itself for estimating the sampling variance, it is largely discouraged in the recent years" (p. 17). In terms of the reason, he argued without an independent study that it produced "rather crude estimates of the standard errors" (p. 17). Nonetheless, even with a crude k value estimation as he alleged, little impact could have occurred in our correlation computing due to the numerator and denominator parts of the correlation configuration noted in our article.

> In reviewing the past TIMSS practice, Mirazchiyski acknowledged,
> Following the theoretical developments of Rubin (1987) and Little and 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). (p. 8)

Indeed, Rubin's (1987) breakthrough has offered several desirable features, including (1) introducing appropriate random errors beyond any deterministic single imputation and (2) offering concise rules for combining the results from multiple imputations for statistical inference.

Despite Mirazchiyski's respect for Rubin's (1987) requirement on the results aggregation, a single PV was used to represent student achievement score in TIMSS 1995 result reporting (see Beaton et al., 1996a, b). According to Gonzalez and Smith (1997, ch. 6, p. 3), the essential step of Rubin's (1987) was not taken because of a decision to ignore the imputation error. When the imputation error was not considered, what was the purpose of spending the precious resources for multiple imputation (MI)? Arguing that one set of the imputed plausible scores can be considered as good as another (Gonzalez & Smith, 1997, ch. 6, p. 3), the TIMSS report fell back to the single imputation result to generate the findings from the first PV. Consequently, the choice of other imputed plausible scores may result in alternative findings different from those in the released TIMSS reports (Wang, 2001).

Mirazchiyski did hint on the detailed differences between TIMSS 1995 and the subsequent cycles of TIMSS in result reporting:

[4]https://bit.ly/2qhV4Tr.

> 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. (p. 6)

Without his elaboration on the details of the differed steps, we take the liberty to clarify one difference in the major international reports between TIMSS 1995 and its subsequent waves of trend studies. The reporting issue should be treated seriously according Mirazchiyski's following statement,

> 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. (p. 18)

Despite the widespread negative impact of the TIMSS 1995 reports (Beaton et al., 1996a, b) against modeling the professional practice in MI application, we no longer treat this as a long-lasting issue. Mirazchiyski's following notes show his agreement with us on the PVs as different variables:

> 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. (p. 8)

More importantly, the incompliance did not repeat itself in other waves of TIMSS reporting (e.g., Mullis et al., 2000; Mullis, Martin, Gonzalez, & Chrostowski, 2004; Mullis, I.V.S., Martin, M.O., & Foy, 2008; Mullis, Martin, Foy, & Arora, 2012a; Mullis, Martin, Foy, & Hooper, 2016a, b).

Although Mirazchiyski agreed on the "imputation variance" or "imputation error" that caused the differences among five PVs, he stressed dimensionality considerations that treat PVs as "a set of variables representing unidimensional measure of the same construct of interest" (p. 7). In particular, Mirazchiyski criticized us for using CCA because "a set of Plausible Values (PVs) does not contain multiple different measures on multiple different latent traits as CCA would assume" (p. 13). Unfortunately, we still do not think this as an issue because CCA does not have the assumption on multiple different latent traits. As Borga (2001) pointed out,

> Canonical correlation analysis (CCA) is a way of measuring the linear relationship between two multidimensional variables. It finds two bases, one for each variable that are optimal with respect to correlations and, at the same

> time, it finds the corresponding correlations. In other words, it finds the two bases in which the correlation matrix between the variables is diagonal and the correlations on the diagonal are maximized. *The dimensionality of these new bases is equal to or less than the smallest dimensionality of the two variables* [italics added for emphasis]. (p. 2)

In theory, if the five PVs have a latent dimensionality equal to 1 for either science or mathematics performance, CCA can still generate statistical findings on these new bases of the latent traits with unidimensionality. In practice, the feasibility was reconfirmed by CCA computing in SAS without any warning or error messages pertaining to the dimensionality checking. Built on the relation between CCA and factor analysis, Levine (1977) also confirmed that just as one (general) factor can occur in factor analysis, so can unidimensionality happen in CCA.

In summary, we agree with Mirazchiyski on the importance of incorporating sampling weight to ensure that certain groups are not overrepresented in the sample for each education system in ILSA. We also agree that additional consideration should be given to the adjustment of variability index due to complex sampling. These methods, including the use of design effect and Jackknife approaches, are included in standard software such as SAS for survey data analyses. Further, the procedure for CCA was well-established in SPSS or SAS to allow dimensionality of the latent bases to range from 1 to another integer, depending on the vector settings. Therefore, we have clarified the issues that seemed important to Mirazchiyski, but not us.

In summary, our original analyses of the TIMSS data were built on in-depth understanding of the sample survey, data imputation, and CCA computing with support from standard software packages in the market. Meanwhile, even with the identification of an indisputable issue in the TIMSS 1995 reports (Beaton et al., 1996a, b) for incompliance to the MI application rule established by Rubin (1987), we still do not treat it as a permanent issue because the compliance problem has been corrected in the other waves of TIMSS reports. From the perspective of supporting ongoing improvement for the benefit of our professional community, we reviewed additional literature below to sustain the momentum of methodology advancement in TIMSS.

What Does the Current Literature Say?

The literature indicates that the employment of Multiple Matrix Sampling design (as in TIMSS) results in "[t]he relatively small number of items per block and the relatively small number of blocks per test booklet" (von Davier, Gonzalez, & Mislevy, 2009, p. 11). Because the total number of achievement items in each ILSA is large, the task of missing value imputation is not small. The imputation model also involves information from background questionnaire to compose predictors. Mirazchiyski relied on the past practice to acknowledge that "[t]he derivation of PVs from population models relies on Rubin's multiple imputation

methods developed in the period between late 1970s to late 1980s" (p. 5).

At that time, Rubin (1987) convinced the research community that five imputations were adequate for treating missing data in general. Software packages, such as SPSS and SAS, accepted this convention to set the number of default imputation to 5 in their MI procedures. Allison (2001) applauded that "Like maximum likelihood, multiple imputation estimates are consistent and asymptotically normal. They are close to being asymptotically efficient" (p. 81).

More recently, Bodner (2008) illustrated that important quantities (e.g., p values, confidence intervals, and estimated fractions of missing information) suffer from substantial imprecision with a small number of imputations. Sullivan, Salter, Ryan, and Lee (2015) concurred that "increasing the number of imputations entails greater precision" (p. 553). Practices of MI computing since the 1990s have led to an increase on the number of imputations in major standard software packages. For instance, SAS Institute (2015) announced that "the default number of imputations in PROC MI has been changed from NIMPUTE=5 to NIMPUTE=25 in SAS/STAT 14.1" (p. 5921). The STATA software manual also indicated that "we recommend using at least 20 imputations to reduce the sampling error due to imputations" (STATA, 2017, p. 5).

Like TIMSS, the MI method implemented by SAS, SPSS, and STATA are all grounded on Rubin's (1987) classic work. However, Wang and Johnson (2018) compared the results of SPSS and SAS computing when the number of imputations was kept at 5. They found that a predictor variable could be claimed both significant and not significant depending on the software being used. As Von Hippel (2016) noted, "[n]on-replicable results reduce scientific openness and transparency, and the possibility of changing results by re-imputing the data offers researchers an opportunity to capitalize on chance by imputing and re-imputing the data until a desired result, such as $p<.05$, is obtained" (p. 2).

More critically, when more variables are involved in the model setting and a large portion of missing information exists in the data set, researchers may encounter a high risk of obtaining conflicting MI results because the number of imputations is set too small in computer-based analyses (Wang & Johnson, 2018). To overcome this kind of problems, Allison (2012) recommended that "if 27% of the cases in your data set have missing data on one or more variables in your model, you should generate about 30 imputed data sets" (p. 1). Hence, after more than 20 years since the completion of TIMSS 1995, the research literature calls into question the optimism of the TIMSS guideline to delimit the number of PV imputations at 5. Finally, we note that PISA (Program for International Student Assessment), another important ILSA project, has already increased the number of PVs from 5 to 10 in its recent cycles[5].

If more PVs were imputed according to the latest wisdom of the research community, a much larger number of PVs would be released in the TIMSS database. Consequently, the risk of Type I error would substantially increase in computing correlation coefficients from more pairs of PVs between mathematics and science. Although this inflation of Type I error is a totally separate issue from the non-additive nature of the correlation coefficients, Mirazchiyski insisted that

[5]https://bit.ly/2qjBs8B.

"The main argument of the authors [i.e., Wang & Ma, 2016] 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* [italics added for emphasis]" (p. 1-2).

In retrospect, this section is built on the latest research literature to urge for imputing more PVs in TIMSS and make it conform to the current professional practice in statistical computing. Meanwhile, more exploration, such as CCA in our original paper, should be encouraged to help control Type I error under the new circumstances. As the number of PVs is on the rise in current ILSA, the CCA approach may offer more advantage to correlate multiple outcome measures for the benefit of computational efficiency and control of Type I error.

Conclusion

TIMSS produces rich comparative information, and it is a daunting task to comprehensively document correct analytical methods for all data analysts. Nonetheless, colleagues in charge of TIMSS data dissemination confronted the challenge by providing a user guide for secondary data analysis (e.g., Foy, Arora, & Stanco, 2013). It should not come as a surprise that researchers find some analytical circumstances that are not clearly articulated by the TIMSS user guide. For instance, researchers are advised to perform computations five times (once for each PV) and average the results (Gonzalez, 1997). The average mechanism, per instruction of Little and Rubin (2002), was an arithmetic average without referring to any transformations. TIMSS researchers never cited StatSoft (2000, p. 10) on the need for converting correlation coefficients into additive measures. We are glad that Mirazchiyski followed a citation from our original article (i.e., StatSoft, 2000) to amend the part not covered by the TIMSS user guide. As we deal with the ongoing knowledge development, there is no need for assuming panacea on what worked. Instead, more attention should be given on what works in a specific context.

Accommodating more imputations is another front that imposes technical challenges for TIMSS data producers to catch up with the recent advancement in statistical computing. The existing tools, including macros such as JACKREGPV and the IDB Analyzer (IEA, 2016), are not flexible for supporting correlation of student performance between mathematics and science with an adequate number of imputations much larger than 5. While we have completely addressed our rationale for using CCA and design effect, TIMSS researchers should be allowed to continue using the methods of their choice, such as Jackknife computing to adjust standard errors that are not delimited to the correlation analyses. Furthermore, ongoing exploration is needed to control the inflation of Type I error should the number of imputations be increased to a default of 25, as accomplished by a standard software package like SAS.

Finally, we wish to acknowledge that Mirazchiyski' response demonstrated his rich knowledge about TIMSS practices. The entire research community owes

him and his colleagues for their persistent effort in producing large-scale international data that are important in comparative studies. In terms of the value of TIMSS, our rejoinder might have inadvertently made his responses like preaching to the converted. Still, his article clearly offered a great opportunity for clarifying the details of TIMSS methodology. It has been our privilege to interact with this exemplary scholar in the field of ILSA.

Acknowledgement

The authors wish to thank Drs. Cosimano, Szolowicz, and Wisman for proofreading this rejoinder and providing valuable comments

References

Allison, P. (2001). *Missing data*. Retrieved from https://bit.ly/2AyhVHG.

Allison, P. (2012). *Why you probably need more imputations than you think*. Retrieved from https://bit.ly/2Q30mot.

Beaton, A.E., Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., Kelly, D.L., & Smith, T.A. (1996a). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Beaton, A.E., Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Smith, T.A., & Kelly, D.L. (1996b). *Science achievement in the middle school years: IEA's Third International Mathematics and Science Study (TIMSS)*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Bodner, T.E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling, 15*, 651-675.

Borga, M. (2001, January 12). Canonical correlation: A tutorial. Retrieved from https:// bit.ly/1XWgwB9

Foy, P., Arora, A., & Stanco, G.M. (1 Eds.) (2013). *TIMSS 2011 user guide for the international database*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College and the IEA.

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., & Smith, T.A. (1997). *Users guide for the TIMSS international database*. Chestnut Hill, MA: TIMSS International Study Center.

IEA. (2016). IEA IDB Analyzer (Version 4.0) [Computer software]. Hamburg, Germany: IEA.

Levine, M.S. (1977). *Canonical analysis and factor comparison*. Newbury Park, CA: Sage.

Kish, L. (1965). *Survey sampling*. New York: John Wiley & Sons.

Little, R.J.A., & Rubin, D.B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., Gregory, K.D., Smith, T.A., Chrostowski, S.J., Garden, R.A., & O'Connor, K.M. (2000). *TIMSS 1999 international science report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston

College.
Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 international science report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Martin, M.O., Mullis, I.V.S., & Foy, P. (with Olson, J.F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades. Chestnut Hill, MA: TIMSS & PIRLS 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., Chrostowski, S.J., & Smith, T.A. (2000). *TIMSS 1999 international mathematics report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Mullis, I.V.S., Martin, M.O., Gonzalez, E.J., & Chrostowski, S.J. (2004). *TIMSS 2003 international mathematics report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Mullis, I.V.S., Martin, M.O., & Foy, P. (with Olson, J.F., Preuschoff, C., Erberber, E., Arora, A., & Galia, J.). (2008). TIMSS 2007 international mathematics report. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012a). *TIMSS 2011 international results in mathematics*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Mullis, I.V.S., Martin, M.O., Foy, P., & Arora, A. (2012b). *TIMSS 2011 international results in science*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2016a). *TIMSS 2015 international results in mathematics*. Retrieved from https://bit.ly/2nAWGOk.
Mullis, I.V.S., Martin, M.O., Foy, P., & Hooper, M. (2016b). *TIMSS 2015 international results in science*. Retrieved from https://bit.ly/2PtZcoX.
National Research Council (2014). *Developing assessments for the Next Generation Science Standards*. Washington, D.C.: The National Academies Press.
Rubin, D. B. (1987) Multiple imputation for nonresponse in surveys. New York: Wiley.
SAS Institute (2015). *SAS/STAT 14.1 user's guide*. Gary, NC: Author.
STATA (2017). *Stata multiple imputation reference manual (Release 15)*. College Station, TX: Author.
StatSoft (2000). *Basic statistics*. Retrieved from http://bit.ly/1rpeZFf
Sullivan, T., Salter, A., Ryan, P., & Lee, K. (2015). Bias and precision of the "Multiple Imputation, Then Deletion" method for dealing with missing outcome data. *American Journal of Epidemiology, 182*(6), 528-534. doi:10.1093/aje/kwv100.
Von Hippel, P. (2016). *The number of imputations should increase quadratically with the fraction of missing information.* Retrieved from https://arxiv.org/ftp/arxiv/papers/1608/1608.05406.pdf.
Wang, J. (2001). TIMSS primary and middle school data: Some technical concerns. *Educational Researcher, 30*(6), 17-21.
Wang, J., & Johnson, D. (2018). An examination of discrepancies in multiple imputation procedures between SAS and SPSS. *The American Statistician*. doi: 10.1080/00031305.2018.1437078.