Statistical Techniques Utilized in Analyzing TIMSS Databases in Science Education from 1996 to 2012: A Methodological Review

Pey-Yan Liou, Ph.D.

Yi-Chen Hung

National Central University

Please address all correspondence to:

Pey-Yan Liou, Ph.D.

Graduate Institute of Learning and Instruction & Center of Teacher Education,

National Central University,

No, 300, Jhongda Rd., Jhongli City, Taoyuan, 32001 Taiwan (R.O.C.)

Tel: 886-3-4227151 ext. 26505

Fax: 886-3-4273371

E-mail: pyliou@ncu.edu.tw

Paper to be presented at the 5th IEA International Research Conference

Singapore – June 26-28, 2013

Statistical Techniques Utilized in Analyzing TIMSS Databases in Science Education from 1996 to 2012: A Methodological Review

Abstract

A methodological review of science education articles using the TIMSS databases published by journals indexed in the SSCI database from 1996 to 2012 was conducted. The identified thirty-four articles are analyzed in terms of the statistical techniques used to analyze the TIMSS databases. The results indicate that the weights and design effects, essential adjustments for analyzing the TIMSS databases, were used in more than half of the studies. The study also summarized the most commonly used quantitative methods for analyzing TIMSS in these articles. Suggestions regarding the use of statistical techniques and reporting are made for researchers who utilize the TIMSS databases in their research.

Keywords: TIMSS; large-scale data analysis; data analysis reporting; weights; design effects

Introduction

It has been widely recognized that to fulfill the future demands of a nation's growth in the global economy, it is essential to develop human capital in the fields of science, technology, engineering and mathematics (STEM). In the field of science education, increasing attention has been paid to these international large-scale assessments (LSA), such as the Trends in International Mathematics and Science Study (TIMSS). Taking one of the leading journals, *Science Education*, as an example, a new section, Science Education Policy, was announced as a direct result of the implementation and consequences of the international LSA. The international LSA are one of the explicit cases of the interaction between policy and research in science education. From the results of these assessments, science educators and stakeholders have a chance to reexamine educational practices based on evidence (Collins, 2004).

The organizations hosting these international LSA publish summary reports with a broad scope. At the same time, the databases are released for public use. TIMSS for instance is one of the largest ILSA and collects rich contextual information at different levels related to student science performance in both national and

international contexts (Martin, Mullis, & Foy, 2008). Researchers can use the released TIMSS databases for secondary data analysis regarding their own areas of interest.

TIMSS are complex sampling survey data, and certain quantitative methodological skills are needed to analyze such data. Traditional statistical methods and software analysis programs, assuming that the data were collected through the simple random sampling technique, may not work well for obtaining the correct parameter estimates and corresponding standard errors when analyzing TIMSS data. Among several statistical issues related to using the TIMSS databases, sampling weights and design effects are the most essential issues needing to be addressed. Without using correct quantitative methodologies, the estimated results from analyzing the TIMSS databases could be invalid (e.g., Rutkowski, Gonzalez, Joncas, & von Davier, 2010).

Literature reviews of the use of sampling weights and design effects in several fields where LSA has been utilized for research have been conducted recently. For instance, Bell et al. (2012) extensively examined articles between 1995-2010 using the three commonly used adolescent health surveys, the *National Longitudinal Study of Adolescent Health*, *Monitoring the Future*, and the *Youth Risk Behavior Surveillance System*. Sixty-one percent of the articles reported using sampling weights, while 60% reported accounting for design effects. In the field of education research, Ene, Askew, and Bell (2012) conducted a review of peer-reviewed articles using the *Early Childhood Longitudinal Study – Kindergarten* or the *National Assessment of Educational Progress*. Their initial results indicated that 75% of the articles reported using weighting, while 62.5% reported accounting for the design effects. Both of the studies observed that the frequency and clarity of reporting varied across databases. Their results inform researchers of the importance of incorporating weights and design effects to achieve the required quality and generalization of the published research findings when analyzing complex survey data. Thus, the first purpose of this study was to conduct a comprehensive review that examines the reporting of the use of sampling weights and design effects in the identified empirical studies with a focus on science education based on TIMSS databases.

Rutkowski et al. (2010) have summarized important issues about weights and the types of analyses to which they should be applied when analyzing the TIMSS databases. There are generally five sets of weights: (1) total student weight

(TOTWGT), (2) student house weight (HOUWGT), (3) student senate weight (SENWGT), (4) overall and subjectwise teacher weight (SCIWGT for analyzing science teacher data and MATWGT for mathematics teacher data), and (5) school weight (SCHWGT). Total student weight is appropriate for single-level student level analyses. Student house weight, also called normalized weight, is used when analyses are sensitive to sample size. Student house weight is essentially a linear transformation of total student weight so that the sum of the weights is equal to the sample size. Student senate weight is used when analyses involve more than one country because it is student total weight scaled in such a way that all students' senate weights sum to 500 in each country. Overall and subjectwise teacher weight should be used when analyzing teacher datasets as attributes of the student. School weight should be used when analyzing school-level data, as it is the inverse of the probability of selection for the selected school.

Another important element for analyzing the TIMSS databases is design effects (Hahs-Vaughn, 2005; Thomas & Heck, 2001; Thomas, Heck, & Bauer, 2005). The definition of design effects is the variance of an estimate under the complex sample design to the variance calculated assuming simple random sampling (SRS). According to the assumption of independence, most of the applied statistical methods cannot be used to analyze the TIMSS databases directly due to dependency among sampled observation units. Inaccurate standard errors may be produced if no adjustment is made when analyzing complex survey data (Hahs-Vaughn, 2005, 2006; Lohr, 1999; Thomas & Heck, 2001). Failure to consider homogeneity of clusters leads to overly small estimations of standard errors and to committing Type I errors in hypothesis testing (Thomas & Heck, 2001). In order to take the issue of dependency of complex survey sampling into consideration, two approaches have been proposed. One is the model-based approach, and the other is the design-based approach (Hahs-Vaughn, 2005; Muthén & Satorra, 1995; Thomas & Heck, 2001). The model-based approach, also called multilevel modeling or hierarchical modeling, directly incorporates the clustered sample design into the analytical models (Raudenbush & Bryk, 2002). The design-based approach, also known as the aggregated approach, estimates a model by focusing on only a single level of analysis (Hahs-Vaughn, 2005; Thomas & Heck, 2001). TIMSS employs the jackknife repeated replication method.

Additionally, since most of the information from the TIMSS databases is

quantitative, investigating which statistical techniques have been utilized in published articles also seems to be a valuable summary for researchers who intend to use the databases for future research. In the field of education, several studies (e.g., Goodwin & Goodwin, 1985; Skidmore & Thompson, 2010) have provided a methodological review of statistical analyses in peer-reviewed journals for certain periods. The suggestions of these studies are made for researchers and readers to interpret and evaluate the research outcomes reported. Therefore, another aim of this study was to summarize the prevalent patterns of statistical analyses used most frequently by authors who have published empirical research based on the TIMSS databases.

Research Questions

Based on the literature review, two research questions are proposed with a focus on science education.

1. How frequently do researchers report the use of sampling weights and account for design effects when publishing findings based on the TIMSS databases in the field of science education?
2. What types of data analysis have been utilized in analyzing the TIMSS databases in the field of science education?

Methods

Selected papers for analysis

This study selected those empirical papers that used the TIMSS databases and were published in journals indexed in the Social Science Citation Index (SSCI) database from 1996 to 2012. SSCI is one of the highly recognized databases which indexes major journals in social sciences including education research. The research time period was set after 1996 because the first assessment of TIMSS is in 1995. Additionally, with an attempt to review articles which are potentially more consistent in terms of quality, this study focuses only on journal articles. Other types of research such as books, book chapters, conference proceedings, unpublished master theses or Ph.D. dissertations were not included.

This study is derived from a larger literature review project about TIMSS and PISA research. Due to the scope of this study, the analyzed papers are constricted to those utilizing the TIMSS, and not the PISA, databases. The data-gathering procedure

involved a series of steps. First, three keywords were entered in the SSCI database, including TIMSS or PISA and Science, to identify possible articles. These three keywords were entered in each of the term boxes, and the field tags were all set to "Topics." Boolean search operators were used to combine the three keywords as the search query: "TIMSS" OR "PISA" AND "science." At the same time, language as *English, d*ocument type as *Article*, category as *Education Educational Research* or *Psychology Educational* or *Psychology Multidisciplinary* or *Education Scientific Disciplines* were defined. Two hundred and twenty-eight articles were identified. Furthermore, the two authors read through the abstract and the whole content of these 228 articles together to determine whether they were empirical studies. Only studies in which the TIMSS databases were directly utilized were counted as empirical studies in this article. Articles which only used numerical results (e.g., student science achievement) found in the TIMSS reports without further conducting statistical analyses were not considered as empirical studies. In the end, 34 articles in which the TIMSS databases were analyzed were identified as the sample for this study.

Coding for statistical techniques used in studies derived from the TIMSS databases

The categories for the statistical methods used in each search article were developed based on the categorization scheme utilized by Goodwin and Goodwin (1985), Bangert and Baumberger (2005), and Baumberger and Bangert (1996), who tabulated and summarized the statistical techniques used in the research articles published by the three educational journals, respectively. However, some modifications of the categorization were made. A given technique used more than once in the same article was coded only once.

Coding procedures

Each of the selected research articles was coded independently by a single researcher. The papers were categorized respectively by two researchers (one has a doctoral degree in educational psychology with a focus on quantitative method, while both have a record of publishing studies using the TIMSS databases). When reviewing the articles, the two researchers were aware that the weights and design effects may be not included in every statistical procedure in some of the articles. A decision about coding the existence of the technique was made if the weights and design effects were

incorporated in the major statistical analysis. Over the course of the coding stage for the study, the reviewers met weekly to discuss the articles included in that week's reliability check. Differing opinions were further discussed and solved to achieve consensus. Thus, there was 100 percent agreement on the coding by the final round of reliability checks.

Results

Sampling weights and design effects

Of the 34 articles included in the review, eighteen (53%) used sampling weights to adjust the unequal probability of subjects representing the characteristics of the population. Meanwhile, twenty-one of the articles (63.6%[1]) used design effects to compute the variance estimation and associated standard errors. Fourteen out of the 34 articles utilized weights and design effects together in their analyses.

TIMSS has different weights for taking the unequal probability of subjects into account. Eighteen of the 34 articles stated that sampling weights were applied in the analyses. Two of the articles utilized both the school and student weights. Four of the 34 studies with a focus on the school level used the school weight. The student house weight is utilized in three articles and the student senate weight in two articles. It seems that the remaining studies used the total student weight because none of them specify which student weight was utilized. Most of the articles only mention "student sample weight" or "sampling weight" or "weight variable" as being utilized in the analyses. Five of the articles conducted analyses in the IEA International data analyzer (IDB analyzer). Thus, the total student weight is assumed to be utilized in these studies.

Design effects, for the stratification and clustering effect of complex sampling data, were utilized in 21 studies. The incorporation of design effects was to estimate correct standard errors. HLM, IDB analyzer and AM were the most commonly used software packages in these 21 articles for taking design effects into account. Although some studies did not explicitly mention the reason for taking design effects into consideration, the use of the particular software automatically solved the issue of design effects in their analyses.

[1] The denominator for calculating the percentage of the design effects used is 33, not 34. In Liu and Ruiz's study (2008), the focus is not individual students' total test score and corresponding variances, so the design effects are not required in the analysis.

Summary of statistical techniques

Table 1 shows the frequency and percentages of the occurrence of the use of statistical techniques in the 34 articles. A total of 100 statistical methods were coded for all articles reviewed for this study. It should be noted that the number of statistical procedures identified was far greater than the number of articles reviewed because most studies incorporated more than one technique. These techniques were coded into three categories based on their level of difficulty – basic, intermediate, and advanced. More than half (56%) of the techniques were at the basic level, while 26% were at the advanced level. Fewer techniques were at the intermediate level (9%).

Table 1.

Major Statistical Techniques by Frequency and Percentage

Techniques	Frequency	Percentage
Basic		
Descriptive statistics	29	29.%
Correlation	12	12%
One-way ANOVA	4	4%
Chi-square	2	2%
Independent *t*-test	9	9%
Intermediate		
Multiple regression	8	8%
Post hoc multiple comparison	1	1%
Advanced		
Canonical correlation	1	1%
Factor analysis	4	4%
Discriminant analysis	1	1%
Structural equation modeling	1	1%
Hierarchical linear modeling	15	15%
Item response theory	4	4%
Others	9	9%
Total	100	

Note. Percentage is based on all 100 statistical techniques coded in the 34 articles.

Conclusions and Academic Significance of Research

The importance of the TIMSS has been recognized around the world. TIMSS have a strong policy orientation, so the data can be utilized, for example, for examining the current educational practice and student learning environment, and for indicating how successive education policies in a country have been implemented. It can be anticipated that the TIMSS databases can be more fully exploited to benefit education both domestically and internationally. To utilize the TIMSS databases soundly, researchers should be equipped with knowledge about the nature of complex sampling data. Sampling weights and design effects are the "basic ingredients" needed for a complete analysis of survey data (Bell et al., 2012). Without taking the basic ingredients into consideration in their analyses, the statistical results may be invalid, and it may therefore be impossible for others to assess the quality and generalizability of the published research findings.

For the summary of sampling weights, the results show that 18 out of the 34 articles incorporated sampling weights when conducting secondary data analysis of the TIMSS databases. Compared with the results from Bell et al.'s (2012) and Ene, Askew, and Bell's (2012) studies, 53% of articles reporting using weights seems to be lower. In the other 16 articles, weights may be have been used in the analyses; however, since no information was provided in these studies, readers have no way to validate the use of weights. It is imperative to report the necessary information about weights to allow readers to evaluate the appropriateness of the particular weight. Furthermore, since researchers are using the TIMSS databases, there should be full transparency in the weights used (Hahs-Vaughn, 2006). Taking the three types of weights in TIMSS for instance, each weight should be used in a given situation. If the authors only state that "weight" was used, it causes ambiguity regarding which weight was used and whether its use was appropriate. Applying an appropriate sampling weight in every analysis is essential for both point estimates and standard error estimates (Rutkowski et al., 2010).

Meanwhile, for the summary of design effects, the results also show that 21 out of the 34 articles accounted appropriately for the sampling design of the survey. Since around 60% of the articles incorporated design effects in Bell et al.'s (2012) and Ene, Askew, and Bell's (2012) studies, 63.6% of the identified articles in this study seems

to be comparable. If the impact of variance estimation without using design effects fails to be recognized, there is more likelihood that any variance estimates will be underestimated. When such variance estimation techniques cannot be executed, it should be noted to show the consequence of the standard errors (Rutkowski et al., 2010). Therefore, improvement is needed in reporting the use of sample weights and accounting for design effects when disseminating science education research based on the TIMSS databases.

Most of the data of the TIMSS databases are quantitative, so applying appropriate statistical techniques is essential to answering the proposed research questions. The findings indicate that more than half (56%) of the techniques adopted were at the basic level, 26% at the advanced level, and 9% at the intermediate level. This is inconsistent with previous studies in which the basic and intermediate levels have the highest percentages of usage (e.g., Bangert & Baumberger, 2005; Baumberger & Bangert, 1996; Goodwin & Goodwin, 1985). One plausible reason is that advanced statistics have become more developed and accessible for researchers to use in recent years. Another explanation may be that for publishing papers using publicly accessible data in academic journals, using advanced statistical methods could uncover some valuable phenomena.

In sum, the significance of this study lies in systematically reviewing and summarizing the basic ingredients for analyzing international LSA. Additionally, this study is the first attempt to review the statistical techniques used in the articles in the field of science education by analyzing articles derived from the TIMSS databases from journals indexed in the SSCI database. The analysis of research articles can provide empirical insights into the data analysis methods, appropriate techniques and reporting that have been used in the field. This study also provides some guidelines regarding these issues for science education researchers interested in analyzing these international LSA databases to be aware of with respect to their own practice.

The limitation of this present study as well as suggestions for future research is to search more databases or to adopt different criteria for selecting analyzed papers. In this study, 34 studies utilizing the TIMSS databases were selected from the SSCI database. Some published articles from journals not in the SSCI list were not included in this analysis. Future studies may increase the number of analyzed articles to capture a more thorough picture of current published studies.

References

Bangert, A. W., & Baumberger, J. P. (2005). Research and statistical techniques used in the Journal of Counseling & Development: 1990 - 2001. *Journal of Counseling & Development, 83*, 480-488.

Baumberger, J. P., & Bangert, A. W. (1996). Research designs and statistical techniques used in the Journal of Learning Disabilities: 1989 - 1993. *Journal of Learning Disabilities, 29*(3), 313-316.

Bell, B. A., Onwuegbuzie, A. J., Ferron, J. M., Jiao, Q. G., Hibbard, S. T., & Kromrey, J. D. (2012). Use of design effects of sample weights in complex health survey data: A review of published articles using data from 3 commonly used adolescent health surveys. *American Journal of Publish Health, 102*(7), 1399-1405.

Collins, A. (2004). Guest editorial. *Science Education, 88*(1), 1-3.

Ene, M., Askew, K., & Bell, B. A. (April, 2012). *Reporting of design effects and sample weights: A review of published early childhood longitudinal study, Kindergarten Cohort and NAEP articles.* Paper presented at the annual meeting of American Education Research Association, Vancouver, Canada.

Goodwin, L. W., & Goodwin, W. L. (1985). Statistical techniques in *AERJ* articles, 1979-1983: The preparation of graduate students to read the educational literature, *Educational Researcher, 14*, 5-11.

Hahs-Vaughn, D. L. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education, 73*(3), 221-248.

Hahs-Vaughn, D. L. (2006). Analysis of data from complex samples. *International Journal of Research and Method in Education, 29*(2), 165-183.

Liu, X. & Ruiz, M. E. (2008). Using data mining to predict K-12 students' performance on large-scale assessment items related to energy. *Journal of Research in Science Teaching, 45*(5), 554-573.

Lohr, S. L. (1999). *Sampling: Design and analysis*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2008). *TIMSS 2007 International Science Report: Findings From IEA's Trends in International Mathematics and Science Study at the Eighth and Fourth Grades*. Chestnut Hill, MA: Boston College.

Muthén, B. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267-316.

Ng, K. T, Lay, Y. F., Areepattamannil, S., & Treagust, D. F., & Chandrasegaran, A. L. (2012). Relationship between affect and achievement in science and mathematics in Malaysia and Singapore. *Research in Science & Technological Education, 30*(3), 225-237.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage. 2nd.

Rutkowski, D., Rutkowski, L., & Plucker, J. A. (2012). Trends in education excellence gaps: a 12-year international perspective via the multilevel model for change. *High Ability Studies, 23*(2), 143-166.

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.

Skidmore, S. T., & Thompson, B. (2010). Statistical techniques used in published articles: A historical review of reviews. *Educational and Psychological Measurement, 70*(5), 777-795.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42*(5), 517-540.

Thomas, S. L., Heck, R. H., & Bauer, K. W. (2005). Weighting and adjusting for design effects in secondary data analysis. *New Directions for Institutional Research, 2005*(127), 51-72.