ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2020; 11(1); 43-60

The Importance of Sample Weights and Plausible Values in Large-Scale Assessments

Serkan ARIKAN * Ferah ÖZER ** Vuslat ŞEKER *** Güneş ERTAŞ ****

Abstract
International large-scale assessments such as PISA (The Programme for International Student Assessment), PIAAC (The Programme for the International Assessment of Adult Competencies) and TIMSS (Trends in International Mathematics Science Study), play a key role in determining educational policies besides their primary objectives of measuring, evaluating and monitoring the educational process. Therefore, it is critical to analyze the data gathered from the large scale assessments using scientifically accurate statistical methods as the results have the potential to influence millions of stakeholders through major policy changes. Analysis of these data that consists of hundreds of different genuine variables requires expertise and using specific methods. This study illustrates issues to be considered while analyzing PISA, PIAAC and TIMSS data by presenting relevant syntax and exemplifying the possible incorrect results that might be encountered. In Turkey, there are very limited courses that focus on large scale data analysis. Workshops are also very limited to reach major groups. The aim of this study is to raise awareness related to sample weights and plausible values. Comparative findings of the study showed that without using sample weights and plausible values there is a high probability to get incorrect results. In this study, t-test and multiple regression analyses conducted by IDB Analyzer and multilevel regression analysis by Mplus were exemplified.

Keywords: Sample weights, plausible values, large scale assessment, IDB Analyzer, Mplus

INTRODUCTION

International large-scale assessments such as PISA (The Programme for International Student Assessment), PIAAC (The Programme for the International Assessment of Adult Competencies) and TIMSS (Trends in International Mathematics Science Study), play a key role in determining educational policies besides their primary objectives of measuring, evaluating and monitoring the educational process (Bialecki, Jakubowski, & Wisniewski, 2017; Figazzolo 2009; Novoa & Yariv-Mashal, 2003; Steiner-Khamsi & Waldow, 2018). In the early periods of these assessments, the developers highly emphasized that the aim of the assessment was mainly monitoring the process rather than cross-country comparisons (Landahl, 2018). Yet, in the following periods, cross-country comparisons raised the interest of both local and international media, which led the test results to be used as also for indicators of economic growth and rationales for policy reforms. Moreover, Addey, Sellar, Steiner-Khamsi, Lingard and Verger (2017) explained the reasons for participation of the countries to these tests as follows: to provide data-based information for policies, technical capacity-infrastructure building, to provide financial support and assistance, prominence in international relations, decision making in domestic politics, economic reasons, reforms to curriculum and teaching

* Asst. Prof, Boğaziçi University, Faculty of Education, İstanbul, serkan.arikan1@boun.edu.tr, ORCID ID: 0000-0001-9610-5496.
** Res. Assist., Boğaziçi University, İstanbul, Faculty of Education, ferah.ozer@boun.edu.tr, ORCID ID: 0000-0001-8621-3522.
*** Res. Assist., Boğaziçi University, İstanbul, Faculty of Education, vuslat.seker@boun.edu.tr, ORCID ID: 0000-0002-3279-5544.
**** Res. Assist., Boğaziçi University, İstanbul, Faculty of Education, gunes.ertas@boun.edu.tr, ORCID ID: 0000-0001-8785-7768.

To cite this article:
Arıkan, S., Özer, F., Şeker, V. & Ertaş, G. (2020). The Importance of Sample Weights and Plausible Values in Large-Scale Assessments. *Journal of Measurement and Evaluation in Education and Psychology, 11*(1), 43-60. doi: 10.21031/epod.602765

Received: 06.08.2019
Accepted: 14.02.2020

approaches. In addition to those, international organizations such as OECD (Organisation for Economic Co-operation and Development), UNESCO, World Bank utilize these assessment results to monitor educational policy reforms in countries and to determine for further investments/grants for developing countries (Addey & Sellar, 2018; Aydın, Selvitopu, & Kaya, 2018). In summary, to date, large-scale assessment data provide crucial information for the efficiency of countries' educational system elements and comparable data about the current student, teacher, and administrator profiles.

Regarding the main reason for the participation of the countries to large scale assessments (Adler, 2017), it is known that in recent years the data-driven results gathered from PISA, PIAAC, and TIMSS have been used for some major and minor educational policy reforms in different countries. In some cases, these major reforms include curricular changes, orientation and the integration of disadvantaged groups; whereas minor reforms include changes in textbooks, educational materials, integration of educational hardware-software and local school cultures. Specifically, it is known that France (Carvalho & Costa, 2015; Michel, 2017), Portugal (Carvalho & Costa, 2015), Poland (Bialecki et al., 2017), Hungary (Carvalho & Costa, 2015), Germany (Ertl, 2006), Sweden (Landahl, 2018), Israel (Pizmony-Levy, 2018) and Spain (Tiana Ferrer, 2017) utilized these source of data to legitimize recent radical policy reforms or curricular changes that were carried out by the different governmental institutions (Wiseman, 2013). Similarly, in Turkey major curricular reforms and changes on the national high-stakes exams have been made since the beginning of the 2000s. Especially in the curriculum changes of 2013 and 2018, the importance of providing learning environments and opportunities that promote higher cognitive skill development, such as analyzing, reasoning, and evaluating has been highly emphasized as an influence of PISA and TIMSS. In line with these policy changes, high-stake central exams were also affected by these major structural changes. For instance, High School Entrance Exam (LGS) has started to measure higher-order thinking skills along with subject matter knowledge (MEB, 2018). Indeed, the so-called national version of PISA administration, namely ABİDE, which aims to measure higher-order thinking skills such as critical thinking, problem-solving and interpretation could also be considered as one of the exemplary initiatives for recent reforms regarding PISA & TIMSS alignment.

Factors such as increased number of large scale assessments-related publications on local and international media (Martens & Niemann, 2010) and elicited media perception related to PISA (Michel, 2017) led the raised awareness on the public **(**Froese-Germain, 2010; Gür, Çelik & Özoğlu, 2012; Steiner-Khamsi & Waldow, 2018). In line with these factors, easy accessibility of the data, serving as a promising field to use the contemporary analysis methods, and providing opportunities for cross-cultural and cross-country comparisons also led the educators and researchers to study on this matter profoundly, which grounded for many national and international publications. In this vein, it is clear that data obtained from large scale assessments have a crucial mission to affect further educational policies. Considering crucial role and mission of large scale assessments, it is critical to analyze these data using accurate statistical methods. Analysis of these data that consists of hundreds of different genuine variables requires expertise. This study illustrates issues to be considered while analyzing PISA, PIAAC and TIMSS data by presenting relevant syntax and exemplifies the possible incorrect results that might be encountered when these issues are not taken into account. In this way, it is aimed to guide researchers studying large scale assessment data to use proper methods.

Large-Scale Tests

There are variety of large-scale assessments and the most widely used ones are PISA, PIAAC, and TIMSS. In the following sections, these assessments are briefly introduced.

Programme for international student assessment (PISA)

PISA is a program organized by the OECD in every three years since 2000 to measure 15-year-old students' performance on mathematics, science, and reading. PISA aims to reveal to what extent students have knowledge and skills needed for modern societies after they complete compulsory education (MEB, 2016a; OECD, 2018). There are three main subject areas in PISA: reading, mathematical literacy, and scientific literacy. PISA measures the degree to which students make use of their learning in these areas in different contexts. While PISA examined reading ability in more detail in 2000, 2009 and 2018, it focused on mathematics literacy in 2003 and 2012, and scientific literacy in 2006 and 2015. In addition, the program collects data from students, teachers, principals, and parents via questionnaires. In the latest PISA carried out in 2018, there were 76 member or nonmember countries. Turkey has been participating in PISA consistently since 2000.

Programme for the international assessment of adult competencies (PIAAC)

PIAAC aims to evaluate the key information processing skills needed for individuals aged 16-65 to participate in social life. The Survey of Adult Skills, as a product of the programme, aims to assess the adults' proficiency by focusing on three key information processing skills: literacy, numeracy, and problem-solving. It is assumed that adults who are proficient in those skills will be able to get benefit from the opportunities generated by technological and structural changes in modern societies (OECD, 2016). In addition to the survey of adult skills, PIAAC includes a comprehensive survey of participants' information related to socio-demographic characteristics. PIAAC was first implemented in 2011-2012 with the participation of 24 countries and on the second round in 2014-2015 with the participation of 9 more countries, the total number of participant countries had reached to 33. Turkey was among those 9 countries that participated on the second round of the study. According to the results of the report *Skills Matter: Further Results from the Survey of Adult Skills* published in 2016, Turkey was significantly below the OECD average (OECD, 2016; TEDMEM, 2016).

Trends in international mathematics science study (TIMSS)

TIMSS is an international study to evaluate the skills and knowledge gained in mathematics and science fields for the 4^{th} and 8^{th} grade students (MEB, 2016b; Mullis & Martin, 2017). TIMSS has been co-developed and administrated by Boston College and International Association for the Evaluation of Educational Achievement (IEA). Since its inauguration in 1995, the test was administrated in 1999, 2003, 2007, 2011, 2015 and 2019 consecutively every 4 years, with the increased number of participating countries in every year. Moreover, the expected number of countries for 2019 administration is likely to be 70 (Mullis & Martin, 2017). Turkey has been included in the TIMSS study in 1999, 2007, 2011, 2015 and 2019 (MEB, 2016b).

TIMSS generally focuses on curricular objective frameworks to evaluate the skills and knowledge gained in mathematics and science fields. Thus, TIMSS curriculum framework is basically three folded as follows: *intended curriculum* in national, social and educational contexts, *implemented curriculum* at home, school, teacher and classroom contexts; *attained curriculum* in student achievement and attitudes contexts. Within these contexts, the TIMSS evaluation framework basically consists of *subject matter dimension*, that focuses on the subject matter knowledge level and *cognitive dimension* that focuses on thinking processes. By providing detailed data among countries' mathematics and science curricula, TIMSS presents the opportunity to make cross-country comparisons as well as local comparisons (MEB, 2016b)

The Important Features of Large Scale Assessment Datasets

There are two important features of large scale assessment (LSA) datasets. The first one is the sample weights which are related to the sampling design of LSA's. The second one is the plausible values related to rotated test design used in the test administration (Rutkowski, Gonzalez, Joncas, & von Davier, 2010). The following section explains these concepts.

Sampling weights

Large scale assessments aim to choose the most representative sample generalizable to the population since it is not possible to use the entire population due to financial inadequacy and time limitations. The sample is useful the extent to which it estimates the characteristics of the population. The most common technique for clarifying the issue of differences between the distribution of characteristics in the sample and in the population is using sampling weights (Rust, 2013). In PISA 2015 technical report, the necessity of using sampling weights was highlighted as to ensure each student in the sample was represented with the correct number of students in the population (OECD, 2017). Sampling weights are used in studies that TÜİK (Turkey Statistics Institution) conducted at the national level and international large scale tests (PISA, TIMSS, & PIAAC, etc.).

In PISA and TIMSS, multistage sampling design is used for sample selection. The use of a multistage design has a significant impact on the precision of resulting estimates (Rust, 2013). In the first stage, schools are selected proportional to their size; and in the second stage classes and/or students are randomly selected from the selected school (LaRoche & Foy, 2016; OECD, 2017). The size of the school is determined by the number of students eligible to participate in the study. For instance, the number of students aged 15 in PISA and the number of students enrolled in 4^{th} or 8^{th} grade in TIMSS are considered to calculate the school size. In PIAAC, all non-institutionalized adults between the ages of 16 and 65 are considered.

Random sampling design is implemented in order to ensure that the sample selection is not biased and that each individual has an equal chance to be selected. Non-random sample designs may cause the bias, whether intentionally or unintentionally. In random sampling also, each individual's chance for selection may not always be equal in the population. In this case, sample weights are used to avoid the bias and to ensure the representativeness of all individuals in the population. A sample unit is determined according to the probability of selection of each individual in the sample. Sample weights are defined as the inverse of the probability of selection for the unit. In other words, if a group has a very low chance to be selected to the sample, the sample unit for the individual representing that group will be higher than the sample unit for the individual coming from the group having high chance to be selected (OECD, 2017). In the analysis, when the sample weights are taken into account for the mean scores of groups, the representation of the population is guaranteed and the estimations are precise. While analyzing the sample data, if the sample weights are used then the contribution of each student to statistical estimations will be proportional to the number of students represented in the population (Gonzales, 2012). Suppose that each individual has an equal chance to be selected among 300. Then, the probability of being selected among 30 individuals will be 1/10 and the weight of each individual will be 10. In this example, since the chance to be selected for each individual is equal, weights for each are also equal. The weights of 30 individuals add up to 300, the total number of individuals in the population. In this case, the weighted mean and the unweighted mean will be equal. For instance, suppose that a sample of 6 students is chosen from a population of 15 girls and 30 boys in a 45-student class. 3 boys and 3 girls are chosen for the sample. While boys are represented more than girls in the population, they are equally represented in the sample. The probability of selection of each 3 girls among 15 girls will be 3/15 = 0.2 and the probability of selection of 3 boys among 30 boys will be

3/30 = 0.1. According to this situation, the weight of each girl in the sample is 5 and the weight of each boy in the sample is 10. Let assume that girls took 8, 7, 7 points from the exam over 10 and boys took 5, 5, 4 points. In this case, while unweighted mean of the sample is [(8 + 7 + 7) + (5 + 5 + 4)]/6 = 6, the weighted mean of the sample which is [(8x5 + 7x5 + 7x5) + (5x10 + 5x10 + 4x10)]/45 = 5.56. Therefore, the weighted mean is 7 % lower than the unweighted mean. In the simplest way, as it is shown in the example, analysis without considering weights would mislead the estimations related to the population.

In multistage sample selection design, in an application that firstly schools are selected and then students are chosen from that school, school weight, within school weight and student weight are determined separately. For example, let the probability of selecting school *j* to be p_j and the probability of selecting students *i* at school *j* (under the condition of school *j* was selected) to be p_{ij}. Then the within school weight is $w_{ij}= 1/p_{ij}$ and the school weight is $w_j= 1/p_j$. In a population of 400 students from 10 different schools having 40 students, firstly 4 schools are randomly selected. Then, 10 students are chosen from each of those schools. The total number of students in the sample is 40. In this case, the probability of selection for each school (4 schools are selected from 10) is p_j = 4/10 = 0.4 and so the school weight is w_j= 2.5. The probability of selection for each student among 4 selected schools (10 students are chosen among 40 in each school) is p_{ij}= 10/40 = 0.25 and within school weight is w_{ij}= 4. Finally, in the case that firstly school is selected and the students are chosen within the school, the probability of selection for a student is $p*_{ij}= p_j$ x p_{ij}= 0.4 x 0.25 = 0.10 and the student weight is $w*_{ij}$= 10.

Since the data gathered from large scale assessments like PISA, PIAAC and TIMSS used multistage sampling, the methods and software that take into account sample weights must be used for all data analysis. The student weights in these data sets are W_FSTUWT (Final trimmed nonresponse adjusted student weight) in PISA, SPFWT0 (Final full sample weight) in PIAAC and TOTWGT (Total student weight) in TIMSS. In multilevel analysis, it is necessary to decompose these weights (Rutkowski et al., 2010). It is important to be aware that the results obtained without considering sample weights will be inaccurate (LaRoche & Foy, 2016; OECD, 2017; Rutkowski et al., 2010). Rutkowski et al. (2010) calculated that the mathematics mean score of Bulgaria as 463.63 when the sample weights were accurately used and 481.38 when sample weights were not used.

Plausible values

The large scale assessments like PISA and TIMSS aim to estimate the performance of population or subgroups in the populations instead of assessing the scores of individuals (Monseur & Adams, 2009; Von Davier, Gonzalez, & Mislevy, 2009). Calculating consistent and valid scores for individuals is not the purpose of large scale assessments. Therefore, the aim is to minimize the errors in population-level estimations (OECD, 2017). Furthermore, the rotated booklet design is used in order to minimize the test burden on individuals (Rutkowski et al., 2010). Students answer only certain parts of the test. However, as a group, student groups answer all of the questions. Therefore, student performance on large scale assessments is reported as plausible values (PVs).

Plausible value method accepts student ability as missing values (Rutkowski et al., 2010). The student ability distributions are estimated through Rubin's (1987) multiple imputation method. Within the distributions, random selections are made and these multiple assigned values are called plausible values (Rutkowski et al., 2010). Plausible values are precedent values for unobservable latent values

(Wu, 2005). Each student has an unobservable latent ability variable and multiple values are assigned to the variable (Laukaityte & Wiberg, 2017; Wu, 2005). OECD (2017) defines plausible values as randomly assigned numbers for individuals from the distribution of scores. The distribution is called marginal posterior distribution. Plausible values including random error variance components should not be considered as test scores, they should be used as defining population performance (OECD, 2017). In short, multiple values are assigned to each individual in order to minimize measurement error (Laukaityte & Wiberg, 2017). If the measurement error is small, multiple values assigned to an individual would be close to each other. On the contrary, if measurement error is large, multiple values assigned to an individual would be far from each other (Wu, 2005). Inferences from large scale assessments become more valid thanks to assigned plausible values and the results of the assessments contribute to the practice more productively (Laukaityte & Wiberg, 2017).

Five plausible values are used in many large scale assessment databases like PISA and TIMSS (OECD, 2017; Laukaityte & Wiberg, 2017). PISA started to report 10 plausible values since 2015. In PIAAC, 10 plausible values are reported. In the National Assessment of Educational Assessment (NAEP) database, 20 plausible values are used. The simulation studies conducted by Laukaityte and Wiberg (2017) showed that using multiple plausible values increases the accuracy of the estimation and decreases measurement error.

It is necessary to use methods and software that take into account plausible values in large scale assessments like PISA, PIAAC, and TIMSS. The researchers should be aware that the outcomes ignoring plausible values would be erroneous (LaRoche & Foy, 2015; OECD, 2017, Rutkowski et al., 2010).

Incorrect Data Analysis Approaches related to Large Scale Assessment Analysis

Rutkowski et al. (2010) listed two common incorrect data analysis approaches when LSA data is used. The first incorrect approach is to use only one of the plausible values. The second one is to take the average of all plausible values. For example, for TIMSS dataset, using only PV1 or averaging PV1 to PV5 are among these common incorrect data analysis approaches. Rutkowski et al. (2010) also added that taking the averages of plausible values creates more severe problems than taking only one plausible value. Therefore, they warned researchers not to use averages of plausible values. In the use of both incorrect approaches, standard errors will be estimated erroneously and p values will be affected. In addition to these aforementioned incorrect approaches, using plausible values as an indicator of a latent variable (such as math performance) in a structural equation model is another incorrect approach. In Turkey, there are studies that used correct approaches as well as incorrect approaches.

Present Study

The main purpose of this study is to raise awareness about LSA data analysis by explaining the structure and showing exemplary analysis. To fulfil this purpose 3 main research questions including group comparison with t-test, multiple linear regression, and multilevel regression were selected. The syntaxes of each analysis related to research questions were also provided in the appendices A-D. The research questions (RQs) of the study are as follows:

1) What are the effects of not taking into account the sample weights and plausible values in group comparison?
2) What are the effects of not taking into account the sample weights and plausible values in multiple regression?

3) Which procedures are used to take into account the sample weights and plausible values in multilevel regression?

To answer these research questions, the following sub-research questions were generated. For the RQ1, "Is there a statistically significant difference between mean TIMSS 2015 mathematics scores of boys and girls in Turkey?" and "Is there a statistically significant difference between mean PIAAC 2015 reading scores of adults who looked for a job last month and who did not look for a job last month in Turkey?"; for the RQ2, "Do disciplinary climate in science classes, epistemological beliefs, index of economic, social and cultural status, inquiry-based science teaching and learning practices, instrumental motivation, enjoyment of science, science self-efficacy, teacher-directed science instruction, teacher support in science classes predict PISA 2015 science performance of students in Turkey?"; for the RQ3 "Do student-level variables, parents make sure that time is allocated for the homework, parents check if the homework is completed, time spent on the homework; and teacher level variables, correcting assignments and giving feedback, letting students to correct their own homework, discussing homework in the classroom, monitoring completeness of the homework, using homework for grading predict TIMSS 2011 reasoning score of students in Turkey?" were used.

METHOD

Sample

In this study, PISA, PIAAC, and TIMSS datasets were used to introduce different LSA data. The sample used in the study is described in this section. In PISA 2015 dataset, there were 5895 students located in 187 schools from Turkey. The majority of students were 10th graders (MEB, 2016a). In PIAAC 2015 Turkey dataset, there were 5227 adults ranging from 16 to 65 years old (OECD, 2016). In TIMSS 2015 dataset, there were 6928 8th grade students located in 239 schools from Turkey (MEB, 2014). In TIMSS 2015 dataset, there were 6079 8th grade students located in 238 schools from Turkey (MEB, 2016b).

Instrument

PISA, PIAAC, and TIMSS have both tests to measure achievement or performance level and questionnaires to collect demographic and attitudinal data of participants. The first research question had two sub-research questions. The first sub-research question was related to the TIMSS 2015 dataset. Mathematics achievement in TIMSS was reported with 5 plausible values (BSMMAT01-BSMMAT05). Mathematics achievement was estimated using item response theory (IRT). The other variable of the research question, gender was taken from the questionnaire data (BSBG01). In the second sub-research question, PIAAC 2015 reading scores of the adults and whether they looked for a paid job was used as variables. Reading scores of adults were reported with 10 plausible values (PVLIT1- PVLIT10). The reading ability of the adults was estimated using IRT. The status of looking for paid job information (yes or no) was gathered from the adult questionnaire (C_Q02b).

In the second research question, the independent variables used in the model were disciplinary climate in science classes, epistemological beliefs, index of economic, social and cultural status, inquiry-based science teaching and learning practices, instrumental motivation, enjoyment of science, science self-efficacy, teacher-directed science instruction, teacher support in a science classes of PISA 2015 (DISCLISCI, EPIST, ESCS, IBTEACH, INSTSCIE, JOYSCIE, SCIEEFF, TDTEACH, TEACHSUP). These student-level independent variables are index scores of related questionnaire items. The science performance score was reported as 10 plausible values estimated by IRT (PV1SCIE-PV10SCIE).

In the last research question, TIMSS 2011 variables that were in the hierarchical structure, students nested in the classrooms, were used. Student level variables were parents make sure that time is allocated for the homework, parents check if the homework is completed, time spent on the homework (BSBG11C, BSBG11D, BSBM20B); and teacher level variables were correcting assignments and giving feedback, letting students to correct their own homework, discussing homework in the classroom, monitoring completeness of the homework and using homework for grading (BTBM25CA, BTBM25CB, BTBM25CC, BTBM25CD, BTBM25CE). The dependent variable, reasoning ability of the students, were estimated using IRT with 5 plausible values (BSMREA01-BSMREA05).

Data Analysis

In this section, how the analyses were performed and important concepts related to LSA data analysis were explained. The first research question was group comparison analysis. In both sub-research questions t-test was conducted as the grouping variables contained two categories. As explained in the introduction, LSA data analysis requires taking into account sample weights and plausible values. IEA's IDB Analyzer can conduct t-test by taking into account sample weights and plausible values (IEA, 2019). IDB Analyzer is an interphase program that can read SPSS files. In the first step, necessary variables including plausible values are selected. In the next step, the sample weight is selected. After these steps, IDB Analyzer produces an SPSS syntax and running the syntax produces the output. IDB Analyzer output does not give significance value (p-value), however, it reports t values. Using t value and the degrees of freedom, statistical significance can be decided. All of these values are reported in "*_sig.sav" output files. In the research question related to TIMSS, Total Student Weight (TOTWGT) was used. In the research question related to PIAAC, Final Full Sample Weight (SPFWT0) was used.

In the second research question, multiple linear regression was used as there were more than one independent variable to predict one dependent variable. IDB Analyzer also can conduct multiple regression by taking into account sample weights and plausible values. In PISA 2015, FINAL TRIMMED NONRESPONSE ADJUSTED STUDENT WEIGHT (W_FSTUWT) was used as a sample weight.

In the last research question, multilevel regression analysis was conducted as the research question contained student-level variables, as well as teacher-level variables. Mplus program was used as Mplus not only can take into account sample weights and plausible values but also multilevel structure of the variables (Muthen & Muthen, 2015). In order to take into account the sample weights, sample weights should be defined in the Mplus syntax. As Rutkowski et al. (2010) advised for multilevel analysis, sample weights were decomposed manually. For level 1 sample weights, the product of WGTADJ2*WGTFAC2*WGTADJ3*WGTFAC3 was used (CLASS WEIGHT ADJUSTMENT* CLASS WEIGHT FACTOR* STUDENT WEIGHT ADJUSTMENT* STUDENT WEIGHT FACTOR). For level 2 sample weights, the product of WGTADJ1* WGTFAC1 (SCHOOL WEIGHT ADJUSTMENT* SCHOOL WEIGHT FACTOR) was used. The product of level1 and level2 sample weights is equal to TOTAL STUDENT WEIGHT. Mplus requires creating separate text files that include one of the plausible values and the rest of the variables. For instance, if there are 5 plausible values, 5 text files that include one of the plausible values as one column and the rest of the variables in the other columns need to be created. Then the names of these text files are listed in a different text file which is the main input file and it is defined in MPLUS syntax (FILE = dataimputedlist.dat;). Also, the data structure should be stated in the syntax (TYPE = IMPUTATION;). Then, the relationships among variables should be defined.

RESULTS

This study aims to compare LSA data analysis with and without taking into account sample weights and plausible values. Also, it is aimed to guide researchers by showing LSA data analysis by providing syntaxes. The results of four main research questions were reported in the following sections comparatively.

Group Comparison Studies

In this section, two sub-research questions were analyzed. The first one is "Is there a statistically significant difference between mean TIMSS 2015 mathematics scores of boys and girls in Turkey?". t-test was conducted as the grouping variable, gender, contained two categories, boys and girls. With and without taking into account sample weights and plausible values were reported in Table 1.

When sample weights and plausible values were used, it was concluded that there was no statistically significant difference between mean TIMSS 2015 mathematics scores of boys and girls in Turkey (t=1.79, p>.05). This result is also the same as the TIMSS 2015 National Mathematics and Science Pre-Report (MEB, 2016b).

Table 1 also includes the results when each plausible value or the average of the plausible values were used. In all cases, there were statistically significant differences between mean TIMSS 2015 mathematics scores of boys and girls in Turkey. These findings totally contradict with the previous finding. Therefore, when sample weights and plausible values are not used, it is highly probable to obtain incorrect results.

Table 1. Comparison of Mathematics Scores of Girls and Boys

Method	Girls (SE)	Boys (SE)	Mean Difference (SE)	*t*
IDB Analyzer PV1-PV5	461.14 (4.80)	454.73 (5.31)	6.40 (3.57)	1.79
SPSS PV1	459.23 (1.90)	452.77 (1.86)	6.46 (2.66)	2.43*
SPSS PV2	460.50 (1.91)	452.87 (1.87)	7.63 (2.67)	2.85**
SPSS PV3	460.26 (1.91)	451.33 (1.91)	8.93 (2.70)	3.31**
SPSS PV4	458.04 (1.97)	449.01 (1.94)	9.03 (2.77)	3.26**
SPSS PV5	459.37 (1.94)	453.84 (1.90)	5.53 (2.72)	2.04*
SPSS PVmean	459.48 (1.87)	451.97 (1.83)	7.51 (2.62)	2.87**

*p < .05. **p < .01. ***p < .001. SE: Standard Error

In the second sub-research question, PIAAC dataset was used. The research question is "Is there a statistically significant difference between mean PIAAC 2015 reading scores of adults who looked for a job last month and who did not look for a job last month in Turkey?"

When sample weights and plausible values were used, it was concluded that there was no statistically significant difference between mean PIAAC 2015 reading scores of adults who looked for a job last month and who did not look for a job last month in Turkey (t=1.16, p>.05).

Table 2 also includes the results when each plausible value or the average of the plausible values was used. Among 11 cases, there were contradictory results. In 3 of these results, significant differences were found and in 8 of them, no difference was found. As similar to the first sub-research question, when sample weights and plausible values are not used, it is probable to obtain incorrect results.

In both sub-research questions, the difference in findings stems from standard errors. The standard errors were higher when sample weights and plausible values were taken into consideration than when they were not used. The change in the standard error directly affects the t value and the ultimate decision.

Tablo 2. Comparison of Reading Scores of Adults Who Looked For a Job and Not

Method	Looked for a job (SE)	Did not look for a job (SE)	Mean difference (SE)	*t*
IDB Analyzer PV1-PV10	226.11 (4.16)	221.05 (1.45)	5.06 (4.36)	1.16
SPSS PV1	229.06 (2.51)	223.90 (.83)	5.16 (2.73)	1.89
SPSS PV2	229.40 (2.59)	223.23 (.83)	6.17 (2.75)	2.25*
SPSS PV3	227.01 (2.57)	224.33 (.83)	2.67 (2.73)	.98
SPSS PV4	226.87 (2.45)	224.12 (.84)	2.76 (2.74)	1.01
SPSS PV5	226.52 (2.55)	222.94 (.83)	3.58 (2.71)	1.32
SPSS PV6	231.42 (2.58)	224.81 (.84)	6.61 (2.75)	2.40*
SPSS PV7	226.62 (2.55)	223.93 (.82)	2.70 (2.71)	1.00
SPSS PV8	226.73 (2.56)	223.70 (.83)	3.03 (2.72)	1.11
SPSS PV9	227.88 (2.51)	222.49 (.84)	5.39 (2.76)	1.95
SPSS PV10	231.14 (2.63)	222.82 (.84)	8.32 (2.75)	3.02**
SPSS PVmean	228.27 (2.34)	223.63 (.76)	4.64 (2.51)	1.85

*p < .05. **p < .01. ***p < .001. SE: Standard Error.

Single-Level Regression Study

In this section "Do disciplinary climate in science classes, epistemological beliefs, index of economic, social and cultural status, inquiry-based science teaching and learning practices, instrumental motivation, enjoyment of science, science self-efficacy, teacher-directed science instruction, teacher support in a science classes predict PISA 2015 science performance in Turkey?" sub-research question was investigated. The results were given in Table 3.

When sample weights and plausible values were taken into account instrumental motivation and teacher support in science classes could not predict the science performance of students. The disciplinary climate in science classes, epistemological beliefs, index of economic, social and cultural status, inquiry-based science teaching and learning practices, enjoyment of science, science self-efficacy, teacher-directed science instruction could predict science performance.

When sample weights and plausible values were not used, among 11 cases, 8 of them produced incorrect results. The main problem was that more variables were found to be significantly related to

the dependent variable which was also related to incorrect standard error estimation. Both using only PV1 or PVmean produced incorrect results. On general R square values were not changed dramatically however, R^2 of PVmean was higher. This example also illustrates that plausible values and sample weights should be used.

Tablo 3. Factors Predicting Science Performance

Method	discipline	beliefs	SES	Inquiry b. science	motivation	enjoy	Self-efficacy	Teacher-directed	support	R^2
IDB Analyzer PV1-PV10	.09***	.19***	.27***	-.19***	.03	.09***	.08***	.04*	.03	.20
SPSS PV1	.08***	.19***	.26***	-.18***	**.03***	.09***	.08***	.05***	**.03***	.19
SPSS PV2	.07***	.20***	.26***	-.18***	**.03***	.11***	.08***	.04**	.02	.20
SPSS PV3	.09***	.20***	.27***	-.19***	**.03***	.09***	.08***	.05***	.02	.20
SPSS PV4	.09***	.19***	.26***	-.19***	.02	.11***	.07***	.05***	.02	.20
SPSS PV5	.09***	.19***	.27***	-.18***	.03	.09***	.07***	.04**	.02	.20
SPSS PV6	.10***	.19***	.26***	-.18***	.03	.09***	.07***	.05***	.02	.19
SPSS PV7	.09***	.19***	.26***	-.20***	**.03***	.09***	.08***	.05***	.03	.20
SPSS PV8	.08***	.19***	.26***	-.18***	**.03***	.10***	.07***	.05***	.03	.19
SPSS PV9	.10***	.19***	.25***	-.20***	.02	.11***	.08***	.04**	**.03***	.20
SPSS PV10	.09***	.19***	.27***	-.19***	**.03***	.09***	.08***	.04**	.01	.20
SPSS PVort	.09***	.20***	.28***	-.20***	**.03***	.10***	.08***	.05***	.02	.22

*$p < .05$. **$p < .01$. ***$p < .001$.

Multilevel Prediction Study

The last sub-research question is "Do student-level variables, parents make sure that time is allocated for the homework, parents check if the homework is completed, time spent on the homework; and teacher level variables, correcting assignments and giving feedback, letting students correct their own homework, discussing homework in the classroom, monitoring completeness of the homework, using homework for grading predict TIMSS 2015 reasoning score in Turkey?". As both student level and teacher level variables were included in the model, multilevel regression was used. The results were given in Table 4.

The intraclass correlation was calculated as 0.32. This value represented that student scores were not independent and scores of the students in the same classrooms were related. Therefore, a multilevel regression analysis was necessary. Also, 32% of the total variance came from between classroom variance and 68% of the total variance came from within classroom variance. The variables of this research question could explain 4% of the variance in student level and 7% of the variance in teacher level. These explained variances were small which implied that the model was not a good one.

The results showed that among student-level variables, parents make sure that time is allocated for the homework and parents check if the homework is completed could predict reasoning scores of students. There was a positive relationship between parents make sure that time is allocated for the homework and reasoning scores. However, there was a negative relationship between parents check if the homework is completed and reasoning scores. Among teacher-level variables, there was a positive relationship between monitoring completeness of the homework and reasoning scores.

Table 4. Standard Coefficients of Multilevel Regression

Variables	Coefficient
Level-1	
time is allocated for the homework	.17***
parents check if the homework is completed	-.19***
time spent on the homework	-.03
Level-2	
correcting assignments	-.04
letting students correct homework	-.05
discussing homework	.10
monitoring completeness of the homework	.16*
grading	.08
Between-class explained variance	%7
Within-class explained variance	%4

*$p < .05$. **$p < .01$. ***$p < .001$.

DISCUSSION & CONCLUSION

It is known that large-scale assessment results are critical in determining educational policies, curriculum reforms and decision-making processes in the use of contemporary innovative practices in education (Hamilton, 2003). The large-scale assessment results also allow cross-country comparisons of various sizes and provide detailed information about the various elements included in the countries' own education system. As a result of its' crucial role in policymaking and the possible influence involving millions of stakeholders, it is required to analyze the data obtained from these tests properly. As it was seen in the cases of examples known as *PISA shock phenomenon* (Wiseman, 2013), misinterpretation of large-scale data sets through primitive and descriptive inferences led irrelevant and radical policy changes in some countries in the past. For instance, Germany's radical policy changes right after their inauguration of PISA 2000 results that were below the OECD average (Waldow, 2009) or Japan's sharp policy changes following the decreased performances in PISA 2000-2003 literacy and maths performance on PISA 2003-2006 could be examples for those misinterpretations (Wiseman, 2013). These instances support the argument that the analysis of the large-scale data sets requires the use of relevant techniques to be embraced (Wiseman, 2013).

As seen in the research questions, in the case of not using sample weights and plausible values appropriately may lead to incorrect results. For instance, as shown in research question 1, in the case of using proper methods of analysis with TIMSS 2015 data led no statistically significant differences between boys' and girls' math performance of Turkey sample. However, statistically significant difference between the groups could be found when the appropriate analysis was not conducted. Similarly, in the second research question, it was shown that multiple regression analysis results could be wrong in the case of not using sample weights and plausible values appropriately. Without taking into consideration of sample weights and plausible values led to 8 incorrect results out of 11 datasets. As Von Davier et al. (2009) and Rutkowski et al. (2010) emphasized within the context of Bulgaria's TIMSS 2007 performance instances, it is vital to use sample weights and plausible values to perform large-scale data set analysis.

Yet, it is seen from the relevant literature regarding the large scale assessment analysis, the awareness regarding embrace these accurate techniques is not as intended. Moreover, the undergraduate or graduate courses offered as well as workshops organized by either researchers or institutions that emphasize how to analyze these LSA data are rare in the national context. As a result of these, even

though there are some studies considering these features of LSA, there are also some studies that use inaccurately only one plausible value or the average of plausible values without using sample weights. In order to overcome these obstacles, this study exemplifies the importance of using sample weights and plausible values by providing the syntaxes. It is recommended for readers of large scale assessments to critically assess whether appropriate techniques are used or not before relying on the research findings. Also, researchers are required to carefully investigate the features of the software embraced in the analysis and to examine the technical reports in the literature for appropriate sample weight use as various sample weights are used in different data sets.

REFERENCES

Addey, C., & Sellar, S. (2018). Why do countries participate in PISA? Understanding the role of international large-scale assessments in global education policy. In A. Verger, M. Novelli & H. Kosar-Altınyeken (Eds.), *Global education policy and international development: New agendas, issues and policies*, (pp. 97-118). New York, NY: Bloomsbury Publishing.

Addey, C., Sellar, S., Steiner-Khamsi, G., Lingard, B., & Verger A. (2017). Forum discussion: The rise of international large-scale assessments and rationales for participation. *Compare, 47*(3), 434-452. doi:10.1080/03057925.2017.1301399

Aydın, A., Selvitopu, A., & Kaya, M. (2018). Eğitime yapılan yatırımlar ve PISA 2015 sonuçları karşılaştırmalı bir inceleme. *İlköğretim Online, 17*(3), 1283-1301.

Bialecki, I., Jakubowski, M., & Wiśniewski, J. (2017). Education policy in Poland: The impact of PISA (and other international studies). *European Journal of Education, 52*(2), 167-174.

Carvalho, L. M. & Estela Costa, E. (2015) Seeing education with one's own eyes and through PISA lenses: considerations of the reception of PISA in European countries, *Discourse: Studies in the Cultural Politics of Education, 36*(5), 638-646. doi:10.1080/01596306.2013.871449

Ertl, H. (2006). Educational standards and the changing discourse on education: The reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*(5), 619-634.

Figazzolo, L. (2009). *Testing, ranking, reforming: Impact of PISA 2006 on the education policy debate*. Brussels: Education International.

Froese-Germain, B. (2010). The OECD, PISA and the impacts on educational policy. *Canadian Teachers' Federation (NJ1).* Retrieved from http://eric.ed.gov/?id=ED532562

Gonzalez, E. J. (2012). Rescaling sampling weights and selecting mini-samples from large-scale assessment databases. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 5*, 117–134.

Gür, B. S., Celik, Z., & Özoğlu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy*, *27*(1), 1-21.

Hamilton, L. (2003). Assessment as a policy tool. *Review of research in education*, *27*(1), 25-68.

International Association for the Evaluation of Educational Achievement (IEA) (2019). IDB Analyzer (version 4.0). Hamburg, Germany: IEA Hamburg.

Landahl, J. (2018): De-scandalisation and international assessments: the reception of IEA surveys in Sweden during the 1970s. *Globalisation, Societies and Education, 16*(5), 566-576. doi:10.1080/14767724.2018.1531235

LaRoche, S., & Foy, P. (2016). Sample design in TIMSS Advanced 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS Advanced 2015* (pp. 3.1–3.27). Erişim adresi http://timssandpirls.bc.edu/publications/timss/2015- a-methods/chapter-3.html

LaRoche, S., & Foy, P. (2016). Sample implementation in TIMSS 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in TIMSS 2015* (pp. 5.1-5.175). Retrieved from http://timss.bc.edu/publications/timss/2015-methods/chapter-5.html

Laukaityte, I., & Wiberg, M. (2017). Using plausible values in secondary analysis in large-scale assessments. *Communications in Statistics-Theory and Methods*, *46*(22), 11341-11357.

Martens, K., & Niemann, D. (2010). Governance by comparison: How ratings & rankings impact national policy-making in education (No. 139). *TranState Working Paper*. Bremen: University of Bremen Collaborative Research Centre

Milli Eğitim Bakanlığı (MEB). (2014). *TIMSS 2011 ulusal matematik ve fen raporu 8. sınıflar*. Ankara: T.C. Milli Eğitim Bakanlığı Yenilik ve Eğitim Teknolojileri Genel Müdürlüğü.

Milli Eğitim Bakanlığı (MEB). (2016a). *PISA 2015 ulusal raporu*. Ankara: T.C. Milli Eğitim Bakanlığı Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Milli Eğitim Bakanlığı (MEB). (2016b). *TIMSS 2015 ulusal matematik ve fen bilimleri ön raporu 4. ve 8. sınıflar*. Ankara: MEB: Ölçme, Değerlendirme ve Sınav Hizmetleri Genel Müdürlüğü.

Milli Eğitim Bakanlığı (MEB). (2018). *2*018 Liselere Geçiş Sistemi (LGS): Merkezi sınavla yerleşen öğrencilerin performansı. *Eğitim Analiz ve Değerlendirme Raporları Serisi (No. 3)*. Ankara: T.C. Milli Eğitim Bakanlığı.

Michel, A. (2017). The contribution of PISA to the convergence of education policies in Europe. *European Journal of Education, 52*(2), 206-216.

Monseur, C., & Adams, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement, 10*(3), 1-15.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA). Retrieved from http://timssandpirls.bc.edu/timss2019/frameworks/

Muthén, L. K., & Muthén, B. O. (2015). *Mplus user's guide*. (7th ed.). Los Angeles, CA: Muthén and Muthén.

Novoa, A. & Yariv-Mashal, T. (2003). Comparative research in education: A mode of governance or a historical journey? *Comparative Education, 39*(4), 423–438.

The Organisation for Economic Co-operation and Development (OECD). (2016). *Skills matter: Further results from the survey of Adult Skills*. OECD Skills Studies. Paris: OECD Publishing. doi:10.1787/9789264258051-en.

The Organisation for Economic Co-operation and Development (OECD). (2017). *PISA 2015 Technical Report*. Paris: OECD Publishing. Retrieved from https://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf

The Organisation for Economic Co-operation and Development (OECD). (2019). *PISA 2018 Assessment and Analytical Framework*. PISA. Paris: OECD Publishing. doi:10.1787/b25efab8-en.

Pizmony-Levy, O. (2018). Compare globally, interpret locally: international assessments and news media in Israel. *Globalisation, Societies and Education, 16*(5), 577-595. doi:10.1080/14767724.20.18.1531236

Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. New York: John Wiley.

Rust, K. (2013). Sampling, weighting, and variance estimation in international large-scale assessments. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (1st ed., pp. 117–154). New York, NY: Chapman and Hall/CRC Press.

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.

Steiner-Khamsi, G. & Waldow, F. (2018). PISA for scandalisation, PISA for projection: the use of international large-scale assessments in education policy making – an introduction. *Globalisation, Societies and Education, 16*(5), 557-565. doi:10.1080/14767724.2018.1531234

Tiana Ferrer, A. (2017). PISA in Spain: Expectations, impact and debate. *European Journal of Education, 52*, 184-191.

TEDMEM. (2016). *OECD yetişkin becerileri araştırması: Türkiye ile ilgili sonuçlar*. Ankara: Türk Eğitim Derneği Yayınları.

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

Waldow, F. (2009). What PISA did and did not do: Germany after the 'PISA-shock'. *European Educational Research Journal, 8*(3), 476-483.

Wiseman, A. (2013). Policy responses to PISA in comparative perspective. In H.D. Meye, & A. Benavot (Eds.) *PISA, power, and policy: The emergence of global educational governance*. (pp.303–322). Oxford: Symposium Books

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2-3), 114-128.

Appendix A. Syntax of The First Research Question-A

```
Include file =
"C:\Users\exper\AppData\Roaming\IEA\IDBAnalyzerV4\bin\Data\Templates\SPSS_Macros\JB_PV.i
easps".
JB_PV  infile="D:\idb\TIMSS_2015.sav"/
       cvar=IDCNTRY BSBG01 /
       almvars=/
       rootpv=BSMMAT0 /
       tailpv=/
       npv=5/
       wgt=TOTWGT/
       nrwgt=150 /
       rwgt=/
       jkz=JKZONE/
       jkr=JKREP/
       jk2type=FULL/
       nomiss=Y/
       method=JRR/
       kfac=0/
       shrtcut=N/
       viewcod=N/
       ndec=2/
       clean = Y/
       strctry = N/
       intavg = Y/
       graphs=Y/
       selcrit = /
       selvar = /
       outdir="D:\idb"/
       outfile="PVMath_gender".
```

Appendix B. Syntax of The First Research Question-B

```
Include file =
"C:\Users\exper\AppData\Roaming\IEA\IDBAnalyzerV4\bin\Data\Templates\SPSS_Macros\JB_PV.ieasps".
JB_PV   infile=" D:\idb\prgturp1.sav"/
        cvar=CNTRYID C_Q02A /
        almvars=/
        rootpv=PVLIT /
        tailpv=/
        npv=10/
        wgt=SPFWT0/
        nrwgt=80 /
        rwgt=SPFWT/
        jkz=/
        jkr=/
        jk2type=HALF/
        nomiss=Y/
        method=PIAAC/
        kfac=0/
        shrtcut=N/
        viewcod=N/
        ndec=2/
        clean = Y/
        strctry = N/
        intavg = Y/
        graphs=Y/
        selcrit = /
        selvar = /
        outdir=" D:\idb"/
        outfile="paidjoblook".
```

Appendix C. Syntax of The Second Research Question

```
include file =
"C:\Users\Toshibanb\AppData\Roaming\IEA\IDBAnalyzerV4\bin\Data\Templates\SPSS_Macros\JB_
RegGP.ieasps".
JB_RegGP infile="C:\idb\PISA_TUR2015.sav"/
        cvar=CNTRYID /
        convar=DISCLISCI EPIST ESCS IBTEACH INSTSCIE JOYSCIE SCIEEFF TDTEACH
TEACHSUP /
        catvar=/
        codings=/
        refcats=/
        ncats=/
        PVRoots=/
        PVTails=/
        dvar0=/
        rootpv=PV /
        tailpv=SCIE /
        npv=10/
        wgt=W_FSTUWT/
        nrwgt=80 /
        rwgt=W_FSTURWT/
        jkz=/
        jkr=/
        jk2type=/
        nomiss=Y/
        method=BRR/
        missing=listwise/
        kfac=0.5/
        shrtcut=N/
        viewcod=N/
        ndec=2/
        clean = Y/
        strctry = N/
        viewprgs=Y/
        viewlbl=Y/
        qcstats=Y/
        newout=Y/
        intavg = Y/
        selcrit = /
        selvar = /
        outdir="C:\idb"/
        outfile="regression".
```

Appendix D. Syntax of The Third Research Question

```
TITLE: this is an example of a two-level
        regression analysis
```

```
DATA: FILE = dataimputedlist.dat;
        !Create a file list;

      TYPE = IMPUTATION;
      !Define that your data has multiple imputation;

 VARIABLE:

  NAMES = IDSCHOOL IDSTUD BSBG11C BSBG11D BSBM20B
  BTBM25CA BTBM25CB BTBM25CC BTBM25CD BTBM25CE
  REAPV WGTADJ1WGTFAC1 WGTADJ2WGTFAC2WGTADJ3WGTFAC3;

  USEVARIABLES ARE IDSCHOOL BSBG11C BSBG11D BSBM20B
  BTBM25CA BTBM25CB BTBM25CC BTBM25CD BTBM25CE
  REAPV WGTADJ1WGTFAC1 WGTADJ2WGTFAC2WGTADJ3WGTFAC3;

  CLUSTER = IDSCHOOL;
  !Define Cluster Variable here;

  MISSING = ALL (9999);

  WEIGHT = WGTADJ1WGTFAC1;
  BWEIGHT = WGTADJ2WGTFAC2WGTADJ3WGTFAC3;
  !Define Sample Weights Here;

  WITHIN = BSBG11C BSBG11D BSBM20B;
   !Define Level1 variables here;

  BETWEEN = BTBM25CA BTBM25CB BTBM25CC BTBM25CD BTBM25CE;
  !Define Level2 variables here;

ANALYSIS:     TYPE = TWOLEVEL;
!Define number of level here;

MODEL:
        %WITHIN%
        REAPV on BSBG11C BSBG11D BSBM20B;
!Define Level1 relationships here;

        %BETWEEN%
        REAPV on BTBM25CA BTBM25CB BTBM25CC BTBM25CD BTBM25CE;
!Define Level2 relationships here;

OUTPUT: STANDARDIZED;
!For standardized coefficients;
```