ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
2019; 10(3); 235-248

The Influence of Using Plausible Values and Survey Weights on Multiple Regression and Hierarchical Linear Model Parameters*

Osman TAT **   İlhan KOYUNCU ***   Selahattin GELBAL ****

Abstract

In large-scale assessments like Programme for International Students Assessment (PISA) and the Trends in International Mathematics and Science Study (TIMSS), plausible values are often used as students' ability estimations. In those studies, stratified sampling method is employed in order to draw participants, and hence, the data gathered has a hierarchical structure. In the context of large-scale assessments, plausible values refer to randomly drawn values from posterior ability distribution. It is reported that using one of plausible values or mean of those values as independent variable in linear models may lead to some estimation errors. Moreover, it is observed that sampling weights sometimes are not used during analysis of large-scale assessment data. This study aims to investigate the influence of three approaches on the parameters of linear and hierarchical linear regression models: 1) using only one plausible value, 2) using all plausible values, 3) incorporating sampling weights or not. Data used in the present study is obtained from school and student questionnaires in PISA (2015) Turkey database. Results revealed that the use of sampling weights and number of plausible values has significant effects on regression coefficients, standard errors and explained variance for both regression models. Findings of the study were discussed in details and some conclusions were drawn for practice and further research.

Key Words: Hierarchical linear modeling, multiple linear regression, plausible values, survey weights, large-scale assessments, PISA.

INTRODUCTION

When determining the group performance, large-scale assessment data are used in many countries so as to take initiatives and develop educational policies. In addition to the cognitive tests measuring the student performance, several scales are used in those applications in order to collect student-, teacher- and school-level information. Through that data, instead of individual assessment, school- and study-related student skills are taken together, and group-level inferences are made. In this type of large-scale assessments, different booklets are designed and applied to students in pairwise blocks in order to prevent the loss of time resulting from the measurement of performance in a wide range of subjects. In this case, as all students do not answer the same questions, it is incorrect and inaccurate to estimate their performance via classical statistical methods and to make a group-level comparison (Organization for Economic Cooperation and Development-OECD, 2017). Hence, such applications employ multiple values demonstrating the possible distribution of student abilities (Von Davier, Gonzales & Mislevy, 2009). The so-called *plausible values* are based on student responses to subset of tests, as well as affective features and available background information (demographic information) (Mislevy, 1991; OECD, 2009).

Plausible values refer to random values drawn from the posterior distributions of ability scores in the context of large-scale assessments (Von Davier et al., 2009). Maximum Likelihood (ML) (Rasch,

* Preliminary results of this study were presented at 6th International Conference on Education, Zagreb-Croatia, 2017.

** Res. Assist., Hacettepe University, Faculty of Education, Ankara-Turkey, osmntt@gmail.com, ORCID ID: 0000-0003-2950-9647
*** Assist. Prof. PhD., Adıyaman University, Faculty of Education, Adıyaman-Turkey, ilhankync@gmail.com ORCID ID: 0000-0002-0009-5279
**** Prof. PhD., Hacettepe University, Faculty of Education, Ankara-Turkey, sgelbal@gmail.com, ORCID ID: 0000-0001-5181-7262

To cite this article:
Tat, O., Koyuncu, İ., & Gelbal, S. (2019). The influence of using plausible values and survey weights on multiple regression and hierarchical linear model parameters. *Journal of Measurement and Evaluation in Education and Psychology, 10*(3), 235-248. doi: 10.21031/epod.486999

Received: 23.11.2018
Accepted: 17.06.2019

1960), Weighted Maximum Likelihood (WML) (Warm, 1985), Joint Maximum Likelihood (JML) (Wright & Stone, 1979), and Expected A Posteriori (EAP) (Bock & Aitkin, 1981) used in estimations made through the Rasch model within the Item Response Theory are estimation methods that cover up each other's flaws. However, these methods make point estimations and do not give more than one ability estimation different from each other coming from the posterior distribution for individuals as in plausible values (Wu, 2005). The first usage of plausible values was inspired by Rubin's (1987) multiple imputation research when analyzing the US National Assessment of Educational Progress (NAEP) data in 1994. Using plausible values in large-scale tests became more common as they were also used in the next NAEP applications, the Trends in International Mathematics and Science Study (TIMSS) by OECD, as well as the Programme for International Student Assessment (PISA). In general, five plausible values are produced for each student, though there is not a strong basis for this limitation in the literature (Von Davier et al., 2009; Wu, 2005).

Plausible values correspond to the distribution of abilities a student can have depending on his / her responses to items. They are obtained by randomly drawn values out of the posterior probability distribution for θ ability values in the Item Response Theory (IRT) (Wu, 2005). The technical reports of the NAEP applications in 1983-1984 and the PISA in 2000 give detailed information about how those values are calculated and how they are drawn from the probability distribution (Adams & Wu, 2002; Beaton, 1987). Plausible values are not individual scores in the traditional sense, and should therefore not be analyzed as multiple indicators of the same score or latent variable (Mislevy, 1993). When compared to the EAP and WML methods that make point estimations, using plausible values will yield less biased results in group-level assessments, as Von Davier et al. (2009) demonstrated in their research. They point out, however, that using the averages of plausible values (PV-W) leads to more biased estimates than using the average of statistics (PV-R) derived by analyzing each value; therefore, the averages of plausible values should not be used as dependent variable (Von Davier et al., 2009). Furthermore, the simulation research by Wu (2005) shows that using any plausible value alone is enough to make highly correct estimates regarding the population parameters.

Instead of assigning point estimations of ability for each student, plausible values from the posterior ability distribution are used in large-scale assessments such as Trends in International Mathematics and Science Study (TIMSS), the PISA, and International Computer and Information Literacy Study (ICILS). The data obtained via those large-scale applications is hierarchically structured within multiple levels (student, school, regions, country, etc.). In fact, it is possible to encounter this data structure in several areas of social science research like organizational, intercultural, and developmental studies (Bryk & Raudenbush, 2002). The data in educational sciences may involve two or more levels as well, with students being nested within classes, classes within schools, and schools within cities or regions, in addition to the repeated measures for students or any unit of analysis. Over the Ghana Youth Save data, for instance, Chowa, Masa, Ramos, and Ansong (2015) examined how the properties of students and schools would affect the academic achievement of youth. Students were nested within schools in the mentioned study. By using the longitudinal data from the students participating the National Longitudinal Survey of Youth (NLSY), Stipek and Valentino (2015) investigated how well measures of short-term and working memory and attention in early childhood predicted longitudinal growth trajectories in mathematics and reading comprehension. The measures in due course were nested within the variable of student as a secondary unit. In the Sustaining Effects Study (SES), Bryk and Raudenbush (1988) used a three-level hierarchical linear model to analyze the relationship between the intensity of student and school poverty for the first to third grade students and their reading comprehension and learning mathematics.

It is common to observe two type of data use if the hierarchical data structure is not taken into consideration. Those are aggregation and disaggregation methods. Aggregation is integrating sub-units of data in upper units. Conjoining the test scores of students at the class level and obtaining school-level scores by weighting their average class-level scores can be taken as examples of aggregation. As individual differences are ruled out in this method, relationships between aggregated variables may be much stronger or lead to misinterpretations (Atar, 2010; Bryk & Raudenbush, 2002; Snijders & Bosker, 2003; Woltman, Feldstain, MacKay, & Rocchi, 2012). In disaggregation, upper units are degraded to lower levels. Assigning the data about a school- and class-level variable to students can be an example of disaggregation. In this case, as all of the students within the same school or class have the same properties, independence of observations as a significant assumption of statistical analyses will be violated (Snijders & Bosker, 2003; Woltman et al., 2012). In conclusion, using linear regression models with aggregation and disaggregation methods will lead to related residuals, as well as to biased coefficients and standard errors on regression equations by ignoring between-group differences (Bryk & Raudenbush, 2002).

Being a way to analyze nested data, hierarchical linear models eliminate the mentioned disadvantages of aggregation and disaggregation methods. Hierarchical linear models have removed the obstacles concerning the examination of analysis unit and measurement change that were important problems in the past (Raudenbush & Bryk, 1986). Thus, estimates for variables at each level, interactions between variables at the same and different levels, as well as components of variance-covariance can be investigated through a single analysis (Bryk & Raudenbush, 2002). The advantages of using hierarchical linear models for hierarchical data include formulating within and between level relations correctly; eliminating the biases resulting from aggregation; enabling to propose more diversified and far-reaching research questions and hypotheses in empirical studies; detecting the appropriate error structures including random effects, and allowing for estimates of standard errors stemming from group effects, including the components of variance and covariance (Raudenbush, 1988). According to Goldstein (2011), hierarchical models enable statistically efficient estimates of regression coefficients, provide correct standard errors, confidence intervals, significance tests, and make it possible to examine within and between relations, as well as to compare the whole levels by taking all factors into consideration. The data analysis section of this research touches on the statistical aspects of hierarchical linear models (HLM) analyses and how they are carried out.

Ignoring the hierarchical structure in the data may lead to a considerable differentiation in the outcome. Roberts (2004) found that the relationship between urbanicity and science achievement was .77 when the hierarchical data structure was ignored, whereas the same relationship was -.88 when the students were nested within school. Likewise, a number of studies argue that using traditional linear models instead of hierarchical ones will yield biased results (Bryk & Raudenbush, 2002; Goldstein, 2011; Osborne, 2000; Raudenbush, 1988; Raudenbush & Bryk, 1986; Woltman et al., 2012). In her study which is a comparison of linear regression and hierarchical linear model, Atar (2010) found that the coefficient of *Attitude Towards Science* in linear regression differs among second level units (schools) in a range from -0.2 to 1.09. The findings shown the degree of attitude towards science significantly differs between schools and multilevel nature of the data should be taken into consideration.

According to Gelman (2006), hierarchical linear models are useful in terms of data reduction and casual inference compared to classical regression analysis. However, using hierarchical linear models do not guarantee the unbiasedness of parameter estimation in the hierarchical data, because some errors of estimates may be observed if the selected sample does not represent the number of students in the population, as it might be the case in the other linear models as well. For this reason, large-scale assessments make use of survey weights pertaining to different levels (Meinck, 2015).

The survey weights used in large-scale tests like PISA make it easier to analyze data, to calculate estimates of sampling errors appropriately, as well as to make valid estimates and inferences of the population. In this way, users are enabled to make unbiased estimates of standard errors, conduct significance tests, and create confidence intervals in consideration of the complex sample design for each participating country. The survey weights are not the same for all students in a given country, because they are to provide full representation of every selected school, to balance the participation of school populations at certain rates, to take school non-responses into consideration, to prevent larger weights in relatively small groups, and to balance the influence of additional number of students sampled for surveys in some countries (OECD, 2014). The statistical procedures underlying the survey weights in tests like TIMSS, and PISA can be found in Cochran (1977), Lohr (2010), Särndal, Swensson and Wretman (1992). The survey weights in those tests consist of the school base weight and the within-school base weight, as well as five adjustment factors. Adjustment factors are used to consider non-participation by other schools that are somewhat similar in nature to a particular school, to balance the age and grade levels of students, to consider non-participating students within the school according to their gender, grade, and region, and to reduce the unexpected school-based and other weight factors. The detailed information on how these weight factors were calculated for PISA 2012 can be found in the technical report (OECD, 2014).

Purpose of the Study

Plausible values and weights used in large-scale assessments are grounded on conducting more precise and inclusionary measurements. Concordantly, this study aims to compare the analysis results of multiple linear regression and hierarchical linear models in predicting science literacy of students, in terms of plausible values and weights, using the PISA data in 2015. Within the frame of this general aim, we first carried out multiple linear regression and hierarchical linear model analyses which one plausible value regressed on independent variables without weights. Then, the same models repeated in such a way that whole weighted plausible values regressed on independent variables. Through this approach we could observe the impact of usage of plausible values with or without weight in both multiple linear regression and HLM. Accordingly, we investigate four research questions: how do the results of multiple regression and HLM analyses turn out in case of a) one unweighted plausible value, b) all plausible values with weights;

1. In fully unconditional model?
2. Regressed on level-1 explanatory variables (students' epistemological beliefs in science, test anxiety, motivations, and the index of economic, social, and cultural status)?
3. Regressed on level-2 explanatory variables (classroom sizes at schools, educational leadership, and shortage of educational material and staff)?
4. Regressed on both level-1 and level-2 explanatory variables?

METHOD

This study is a correlational research (Fraenkel, Wallen & Hyun, 2012) aiming to demonstrate relationship among plausible values, survey weights and few independent variables in two different analyses, with reference to the hierarchically structured data obtained from the international large-scale education research.

Working Group

The PISA 2015 dataset was used in accordance with the aim of the study. PISA is a triennial international survey conducted by OECD, mainly aiming to measure the mathematics, science, and reading performance of 15-year-old students. The first and the latest PISA surveys were conducted in 1997 and 2015, respectively. Nearly 520 thousand students from 72 countries were assessed. From Turkey, 5895 students from 187 schools in total took the PISA test.

This study incorporates two-level hierarchical data (with students being level-1, and schools being level-2), in line with the nature of hierarchical linear models. The sample of the level-2 consists of 178 schools in Turkey without any missing data. The schools with missing data were excluded from the dataset, since it is impossible to conduct analysis with missing data in level-2 units in HLM software. HLM software works with compete level-2 data. It is an obligation either to impute a value for the missing data or to delete incomplete cases. Ignoring level-2 missing observation will result in listwise deletion of incomplete level-2 units during the creation of system files (Palardy, 2011). The level-1 sample of the research consists of 5703 students receiving education in the afore-mentioned 178 schools. For hierarchical linear models, level-2 sample size of 50 or more with adequate level-1 sample size is expected to provide unbiased estimates (Maas & Hox, 2005). Hence, the sample size of this research is appropriate enough to perform HLM-related analyses.

Data Collection Instruments

In PISA, students take mathematics, science, and reading comprehension tests. Their cognitive skills are assessed in these fields. Besides the cognitive skill tests, one of those three fields are designated as an area of focus in every application, and a student questionnaire is applied to assess affective variables related to the specified area of focus. The data related to students is gathered through cognitive tests and questionnaires in which affective variables are examined. In a similar way, a school questionnaire is applied to school principals in order to gather information in a variety of issues, such as technical infrastructure and status of educational resources at schools. In this study, the level-1 variables from the student questionnaire and the science test were used together with the level-2 variables from the school questionnaire. The variables used in the model selected via Automatic Linear Modeling (Yang, 2013) procedure. This analysis carried out with 14 index or continuous variables. Then, out of 10 most important variables eight variables (two variables exclude for having equal importance levels) decided to be used. We tried to provide a clear representation of the finding as much as possible with parsimonious models based on most important variables. The details about those variables are seen in Table 1.

Table 1. Variables and Their Properties

Level	Variable	Abbreviation	Type	Nature
Student (Level-1)	Science Literacy Scores (10 Plausible Values)	PV1SCIE (1-10)	Dependent	Continuous
	Epistemological Beliefs	EPIST	Independent	Continuous
	Test Anxiety	ANXTEST	Independent	Continuous
	Achievement Motivation	MOTIVAT	Independent	Continuous
	Index of Economic, Social and Cultural Status	ESCS	Independent	Continuous
School (Level-2)	Average Class Size of School	CLSIZE*	Independent	Continuous
	Teachers Participation	LEADTCH*	Independent	Continuous
	Shortage of Educational Material in School	EDUSHORT*	Independent	Continuous
	Shortage of Educational Staff in School	STAFFSHO*	Independent	Continuous
Weights	Final Student Weight	W_FSTUWT		Continuous
	BRR-FAY Replicate Weights (80 in number)	W_FSTURWT1-80		Continuous

*Disaggregated to the student level in multiple regression analysis

Data Analysis

The analysis of this research involves multiple regression and hierarchical linear models with the purpose of investigating the influence of plausible values and survey weights on different statistical analyses. In the multiple regression analysis, PV1SCIE1 was set as the dependent variable, and four different models were analyzed. Those models included no explanatory variables, only student-level variables, only school-level variables, and variables pertaining to both levels. In the multiple regression analysis, school-level variables were disaggregated to students. The mentioned four models were repeated while 10 plausible values (PV1SCIE1-10) were made dependent variables, and student-level weight and replications (W_FSTUWT and W_FSTURWT1-80) were used. In this way, we examined the effects of plausible values and weight use on multiple regression analysis.

Like the multiple regression analysis, the HLM analyses involved eight models in total, in four of which only the first plausible value (PV1SCIE1) was dependent variable, and in four of which all plausible models and weights were employed. For data analysis, the IDB Analyzer (International Association for the Evaluation of Educational Achievement-IEA, 2016) software was used to create the syntax that makes it possible to utilize all plausible values and weights in multiple regression. The main analyses were performed via SPSS 21.0 (International Business Machines-IBM Corp., 2012) and

HLM 7 Hierarchical Linear and Nonlinear Modelling (Bryk, Raudenbush & Congdon, 2010). .05 is significance level for all analyses.

Before the carrying out the multiple regression and HLM analyses we tested assumptions of both analyses. Firstly, we checked multiple regression assumptions in terms of linear relationship between dependent variable and independent variables, multicollinearity, independence of residuals (uncorrelated residuals), constant residual variance (homoscedasticity), normal distribution of residuals and outliers for all models (except intercept only models). By scatter plots drawn with dependent variable against independent variables for all models, we could observe linear relationship among outcome and explanatory variables. For all models, multicollinearity tested with tolerance and VIF statistics. Accordingly, it is found that none of tolerance value is smaller than 0.2 (0.7-0.9) and none of VIF is greater than 10 (1.3-1.4). The independence of residuals tested via Durbin-Watson statistics. According to the test, it is indicated that for all models mentioned statistics is in a range from one to three. For the constant residuals (homoscedasticity) we benefited from a graph of predicted standard points against standard residuals. Through the P-P graph we could decide residuals are on the diagonal line and are normally distributed. Finally, outliners tested with Cook's distance method. We did not meet any distance greater than one.

Since the first HLM model is fully unconditional model we did not check the assumptions. For all other models, homogeneity of variances and normality of residuals for each level are strictly recommended (Snijders & Bosker, 1999). For all models we created scatter plots for level-1 among level-2 units and we observed that residuals are randomly distributed among level-2 units. Finally, we drawn P-P plots of predicted standard points against standard residuals and determined that the residuals are normally distributed.

RESULTS

Findings on the First Sub-Problem

Table 2 demonstrates the details about four different models that are constructed by considering the absence of any explanatory variable. As seen in the table, the multiple regression model in which all plausible values and weights are used is the highest predictor of Turkish general science literacy score, whereas the HLM analysis in which all plausible values and weights are used is the lowest predictor. The smallest standard error estimation is obtained via multiple regression model (1.02), while the highest standard error is obtained through the HLM analysis where all plausible values and weights are used.

Table 2. Fixed Effects Pertaining to The First Model

Analysis	Fixed Effect	Coefficients	Se	t
Multiple Regression (PV1SCIE1)	Grand Mean of Science Literacy	423.19*	1.02	414.89
Multiple Regression (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy	426.22*	4.06	104.98
HLM (PV1SCIE1)	Grand Mean of Science Literacy, γ_{00}	418.48*	4.35	96.13
HLM (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy, γ_{00}	417.71*	4.90	85.29

*p < .05

The random effects from two different random effects ANOVA models are presented in Table 3. The results of both analyses indicate that the mean of student science achievement differs from a school to another. While the level-1 error term estimated in both analyses is too close, the level-2 error term estimated with the HLM analysis using all plausible values and weights is higher as compared to the first analysis.

Table 3. Random Effects Pertaining to The First Model

Analysis	**Random Effect**	**Sd**	**Variance**	χ^2
HLM (PV1SCIE1)	Level-2 Error Term, u_{0j}	55.43	3073.53	5499.68*
	Level-1 Error Term, r_{ij}	53.61	2873.90	
HLM (PV1SCIE1-10) Weighted	Level-2 Error Term, u_{0j}	59.46	3536.05	6332.42*
	Level-1 Error Term, r_{ij}	53.37	2848.21	

*p < .05

The intra-class correlation coefficient was used to determine the percentage of variance in science literacy explained at school level. Accordingly, the proportions obtained from both analyses are as follows:

$$\rho_1 = {}^{\tau_{00}}/_{(\tau_{00}+\sigma^2)} = 3073.53/\ (3073.53+2873.90) = 0.517 \quad (1)$$

$$\rho_2 = {}^{\tau_{00}}/_{(\tau_{00}+\sigma^2)} = 3536.05/\ (3536.05+2848.21) = 0.554 \quad (2)$$

In the first analysis, it was determined that approximately 52% (ρ_1 = 0.517) of the variance in dependent variable can be explained at school level. On the other hand, when all plausible values were used, approximately 55% (ρ_2 = 0.554) of the variance in dependent variable could be explained at level-2.

Findings on the Second Sub-Problem

Table 4 shows the coefficients pertaining to two different multiple regression analyses, in which four student-level variables were included in the model, as well as the fixed effects from two different random coefficients models.

Table 4. Fixed Effects Pertaining to The Second Model

Analysis	**Fixed Effect**	**Coefficients**	**Se**	**t**
Multiple Regression (PV1SCIE1)	Grand Mean of Science Literacy	452.97*	1.65	274.11
	EPIST	14.74*	0.83	17.80
	ANXTEST	-6.81*	0.94	-7.29
	MOTIVAT	6.05*	0.98	6.16
	ESCS	17.84*	0.82	21.66
Multiple Regression (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy	456.05*	4.56	100.08
	EPIST	15.23*	1.3	11.75
	ANXTEST	-6.27*	1.4	-4.47
	MOTIVAT	6.53*	1.38	4.74
	ESCS	18.69*	2.05	9.12
HLM (PV1SCIE1)	Grand Mean of Science Literacy, γ_{00}	423.31*	4.61	91.74
	EPIST,γ_{10}	7.37*	0.68	10.81
	ANXTEST,γ_{20}	-6.35*	0.82	-7.72
	MOTIVAT, γ_{30}	3.26*	0.90	3.63
	ESCS, γ_{40}	1.91*	0.78	2.46
HLM (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy, γ_{00}	422.10*	5.18	81.49
	EPIST, γ_{10}	7.41*	0.85	8.71
	ANXTEST, γ_{20}	-6.15*	0.93	-6.61
	MOTIVAT, γ_{30}	3.89*	1.05	3.70
	ESCS, γ_{40}	1.87*	0.91	2.03

*p < .05

In line with Table 4, it is possible to say that overall science literacy is over estimated in each analysis. The coefficients in these analyses reflect the mean science literacy when independent variables are controlled. In every analysis, all the independent variables significantly predict the dependent variable. Whereas the variable with the largest coefficient is the index of Economic, Social, and Cultural Status (ESCS) in the multiple regression analysis, this variable was estimated very lower in the HLM analyses. Students' levels of epistemological belief (EPIST) is the variable with the highest coefficient in the HLM analyses. Furthermore, it can be said that all coefficients in the multiple regression analyses were estimated higher when compared to the HLM analyses. It was seen that the standard errors pertaining to the coefficients in the unweighted multiple regression and HLM analyses were estimated low when compared to the weighted multiple linear regression. The random effects from the random coefficient models are presented in Table 5.

Table 5. Random Effects Pertaining to The Second Model

Analysis	Random Effect	Sd	Variance	χ^2
HLM (PV1SCIE1)	Level-2 Error Term, u_{0j}	56.50	3192.56	1982.77*
	EPIST Effect, u_{1j}	1.51	2.28	143.22
	ANXTEST Effect, u_{2j}	3.94	15.51	159.71
	MOTIVAT Effect, u_{3j}	5.26	27.69	176.79
	ESCS Effect,u_{4j}	1.68	2.82	146.86
	Level-1 Error Term, r_{ij}	52.15	2719.55	
HLM (PV1SCIE1-10) Weighted	Level-2 Error Term, u_{0j}	60.19	3622.63	2220.32*
	EPIST Effect, u_{1j}	2.47	6.09	144.92
	ANXTEST Effect, u_{2j}	2.72	7.38	146.64
	MOTIVAT Effect, u_{3j}	5.16	26.59	163.78
	ESCS Effect, u_{4j}	3.02	9.11	149.17
	Level-1 Error Term, r_{ij}	51.89	2693.07	

*p < .05

In random effects, level-1 error variances are expected to become smaller when level-1 independent variables are included in the model. Equation 3 and 4 were used to determine to what extent the level-1 variance is explained by the level-1 variables included in the model.

$$\text{Unweighted HLM } \rho_1=(\sigma^2_{ANOVA}-\sigma^2_{RIM})/\sigma^2_{ANOVA}: = (2873.90 - 2719.55) / 2873.90 = 0.05 \quad (3)$$

$$\text{Weighted HLM } \rho_2=(\sigma^2_{ANOVA}-\sigma^2_{RIM})/\sigma^2_{ANOVA} = (2848.21 - 2693.07) / 2848.21 = 0.05 \quad (4)$$

Both HLM models explained the level-1 variance to the equal extent, although the variance of level-1 error was smaller in the HLM analysis in which all plausible values and weights were used. The level-2 error variance was estimated higher when weights were used.

Findings on the Third Sub-Problem

In Table 6, the coefficients pertaining to two different multiple regression analyses, in which four school-level variables were disaggregated, as well as the fixed effects from two different HLM analyses.

Table 6. Fixed Effects Pertaining to The Third Model

Analysis	Fixed Effect	Coefficients	Se	t
Multiple Regression (PV1SCIE1)	Grand Mean of Science Literacy	411.89*	4.621	89.14
	CLSIZE	0.290*	0.09	3.07
	LEADTCH	5.00*	0.90	5.56
	EDUSHORT	-9.58*	0.94	-10.25
	STAFFSHO	-8.43*	1.01	-8.35
Multiple Regression (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy	416..84*	25.41	16.4
	CLSIZE	0.27*	0.53	0.51
	LEADTCH	4.38*	5.2	0.84
	EDUSHORT	-10.43*	3.75	-2.78
	STAFFSHO	-11.05*	4.3	-2.57
HLM (PV1SCIE1)	Grand Mean of Science Literacy, γ_{00}	399.33*	16.37	24.40
	CLSIZE,γ_{10}	0.52	0.35	1.50
	LEADTCH, γ_{20}	4.05	3.65	1.11
	EDUSHORT, γ_{30}	-9.09*	3.02	-3.01
	STAFFSHO, γ_{40}	-9.85*	4.01	-2.45
HLM (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy, γ_{00}	399.89*	20.49	19.51
	CLSIZE,γ_{10}	0.57	0.42	1.35
	LEADTCH, γ_{20}	3.38	4.42	0.77
	EDUSHORT, γ_{30}	-8.44*	3.70	-2.29
	STAFFSHO, γ_{40}	-15.39*	4.63	-3.32

*p < .05

As seen in Table 6, all the variables in both multiple regression analyses significantly predict the dependent variable, while only the shortage of educational materials (EDUSHORT) and the shortage of educational staff (STAFFSHO) remain significant in the HLM analyses. Standard errors increase with the use of weighted plausible values in both regression and HLM analyses. It is seen that some of the weighted multiple regression coefficients are slightly greater than those of the unweighted multiple regression analysis coefficients. The effect of weighting on the coefficients was not found considerable in the HLM analyses. The random effects related to the HLM analyses are presented in Table 7.

Table 7. Random Effects Pertaining to The Third Model

Analysis	Random Effect	Sd	Variance	χ^2
HLM (PV1SCIE1)	Level-2 Error Term, u_{0j}	51.64	2667.09	4700.68*
	Level-1 Error Term, r_{ij}	53.61	2873.60	
HLM (PV1SCIE1-10) Weighted	Level-2 Error Term, u_{0j}	53.46	2857.64	5243.57*
	Level-1 Error Term, r_{ij}	53.36	2847.75	

*p < .05

The variance of level-2 error term was estimated higher in the weighted HLM analysis, as demonstrated in Table 7. The level-2 variance is expected to decrease with the inclusion of level-2 variables into the completely unconditional model. Equation 5 and Equation6 were utilized to determine to what extent the level-2 variance is explained by the level-2 variables included in the model.

Unweighted HLM: $\rho_1=(\sigma^2_{ANOVA}-\sigma^2_{MAOR})/\sigma^2_{ANOVA}$ = (3073.53 - 2667.09) / 3073.53 = 0.13 (5)

Weighted HLM: $\rho_2=(\sigma^2_{ANOVA}-\sigma^2_{MAOR})/\sigma^2_{ANOVA}$ = (3536.05 - 2847.75) / 3536.05 = 0.20 (6)

Whereas 13% of the level-2 variance is explained in the unweighted HLM analysis after four level-2 independent variables are included into the model, this percentage rises to 20% in the weighted HLM analysis. Hence, it is possible to say that weighting had a certain effect on the variance explained in the HLM analysis.

Findings on the Fourth Sub-Problem

Table 8 shows the coefficients from two different regression models, in which the level-1 variables were modelled together with the level-2 variables that were found to be significant. The table also shows the fixed effects pertaining to the model of intercepts and slopes as two different dependent variables.

Table 8. Random Effects Pertaining to The Fourth Model

Analysis	**Fixed Effect**	**Coefficients**	**Se**	**t**
Multiple Regression (PV1SCIE1)	Grand Mean of Science Literacy	453.91*	1.65	275.59
	EDUSHORT	-7.11*	0.884	-8.04
	STAFFSHO	-7.00*	0.96	-7.26
	EPIST	14.00*	0.81	17.23
	ANXTEST	-6.79*	0.92	-7.42
	MOTIVAT	5.76*	0.96	5.98
	ESCS	15.04*	0.83	18.17
Multiple Regression (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy	457.20*	4.79	95.35
	EDUSHORT	-7.8*	3.07	-2.54
	STAFFSHO	-9.07*	3.83	-2.37
	EPIST	14.16*	1.22	11.58
	ANXTEST	-6.32*	1.3	-4.86
	MOTIVAT	6.21*	1.35	4.61
	ESCS	15.63*	1.94	8.07
HLM (PV1SCIE1)	Grand Mean of Science Literacy, γ_{00}	429.33*	4.71	91.20
	EDUSHORT, γ_{01}	-8.85*	3.25	-2.72
	STAFFSHO, γ_{02}	-6.97*	3.55	-1.96
	EPIST, γ_{10}	7.43*	0.68	10.85
	ANXTEST, γ_{20}	-6.32*	0.82	-7.69
	MOTIVAT, γ_{30}	3.22*	0.89	3.61
	ESCS, γ_{40}	1.85*	0.78	2.39
HLM (PV1SCIE1-10) Weighted	Grand Mean of Science Literacy, γ_{00}	431.37*	5.26	81.97
	EDUSHORT, γ_{01}	-8.24*	3.54	-2.33
	STAFFSHO, γ_{02}	-12.35*	3.95	-3.13
	EPIST, γ_{10}	7.43*	0.83	8.92
	ANXTEST, γ_{20}	-6.09*	0.93	-6.52
	MOTIVAT, γ_{30}	3.78*	1.05	3.59
	ESCS, γ_{40}	1.86*	0.93	2.00

*p < .05

According to Table 8, the significant variables in the multiple regression analyses are the epistemological beliefs (EPIST) of students and the index of economic, social, and cultural status (ESCS), both being the student-level variables, while the predictors with highest coefficients in the HLM analyses are the level of epistemological beliefs and test anxiety (ANXTEST), both being the student-level variables again. Besides, it is seen that the coefficients of regression analyses are estimated higher than those of the HLM analyses, whereas the standard errors pertaining to the coefficients are estimated lower, as it was the case in the previous models. In case of weighting, a remarkable increase is observed in standard errors of level-2 variables in the multiple regression analyses. As for the HLM analyses, weighting does not create any considerable change on the coefficients and standard errors thereof. Table 9 demonstrates the random effects pertaining to the HLM analyses.

Table 9. Random Effects Pertaining to The Fourth Model

Analysis	**Random Effect**	**Sd**	**Variance**	χ^2
HLM (PV1SCIE1)	Level-2 Error Term, u_{0j}	52.97	2805.70	1666.33*
	EPIST Effect,u_{1j}	1.63	2.66	143.29
	ANXTEST Effect,u_{2j}	3.95	15.59	159.68
	MOTIVAT Effect, u_{3j}	5.13	26.35	176.69
	ESCS Effect, u_{4j}	1.65	2.72	146.78
	Level-1 Error Term, r_{ij}	52.15	2719.60	
HLM (PV1SCIE1-10) Weighted	Level-2 Error Term, u_{0j}	55.15	3041.55	1786.47*
	EPIST Effect, u_{1j}	2.50	6.24	144.95
	ANXTEST Effect, u_{2j}	2.70	7.27	146.54
	MOTIVAT Effect, u_{3j}	5.10	26.99	163.38
	ESCS Effect, u_{4j}	2.96	8.79	149.15
	Level-1 Error Term, r_{ij}	51.89	2692.70	

*p < .05

In order to determine the percentages of variance explained for the models of intercepts and slopes as dependent variables, the variances obtained from these models were compared with those obtained from the random effects ANOVA model.

Level-1 variance explained:

Unweighted HLM: $\rho_1=(\sigma^2_{ANOVA}-\sigma^2_{MAOANCOVA})/\sigma^2_{ANOVA}$ = (2873.90 - 2719.60) / 2873.90 = 0.05 (7)

Weighted HLM: $\rho_2=(\sigma^2_{ANOVA}-\sigma^2_{MAOANCOVA})/\sigma^2_{ANOVA}$ = (2848.21 - 2692.70) / 2848.21 = 0.05 (8)

Level-2 variance explained:

Unweighted HLM: $\rho_1=(\sigma^2_{ANOVA}-\sigma^2_{MAOANCOVA})/\sigma^2_{ANOVA}$ = (3073.53 - 2805.70) / 3073.53 = 0.09 (9)

Weighted HLM: $\rho_2=(\sigma^2_{ANOVA}-\sigma^2_{MAOANCOVA})/\sigma^2_{ANOVA}$ = (3536.05 - 3041.55) / 3536.05 = 0.14 (10)

Accordingly, the level-1 variance explained remained the same when one plausible value was used and weighting was not applied in the analyses performed via the model of intercepts and slopes as dependent variables. Per contra, the level-2 variance explained was found higher (14%) when all plausible values and weights were used together.

DISCUSSION and CONCLUSION

This study aimed to compare the results of multiple linear regression and HLM in cases of using a plausible value and all plausible values together with survey weights as an indicator of students' science literacy. Within the scope of this aim, the estimates of those methods were compared regarding the four cases, i.e., the absence of any explanatory variable, the existence of student-level variables, school-level variables, and variables from both levels.

In the models without any explanatory variables, the highest average of science literacy was estimated through the multiple linear regression model using all plausible values and weights. In general, both multiple linear regression analyses can be said to have estimated science literacy higher than the HLM analysis did. Weighting was effective in estimating the coefficient-related standard errors in both analyses of regression and HLM. It was observed that standard errors were greater when weighting was applied in both analyses. Hence, it can be asserted that weighting has a considerable role in relation to the significance of coefficients. Students' science literacy varied from a school to another, according to the random effects of both random-effects ANOVA models. It was observed that the percentage of variance explained by schools as level-2 units was higher when weighting was applied. This means that the difference among schools further increased as a result of weighting. In this study, it was seen that approximately 55% of the variance in the dependent variable was explained by level-2 units. This result manifests the importance of using HLM analyses as emphasized in several studies (Bryk & Raudenbush, 2002; Goldstein, 2011; Osborne, 2000; Raudenbush, 1988; Raudenbush & Bryk, 1986; Woltman et al., 2012).

For all the models, science literacy was predicted significantly by students' epistemological beliefs in science, test anxiety, motivation, and the index of economic, social, and cultural status. The literature about large scale studies such as PISA and TIMSS contain many researches that investigate economic, social and cultural development index (Acar & Öğretmen, 2012; Atar & Atar, 2012). The findings of current study related this variable is parallel with former ones. Epistemological beliefs and the index of economic, social, and cultural status were the variables with the biggest coefficients in three out of four models over which level-1 variables were examined, the coefficients related to these variables were estimated quite higher in multiple regression analyses as compared to the HLM. The standard errors estimated in those four models were quite close to each other. However, in case of weighting, the standard errors estimated were observed to be slightly higher compared to the other models. Even though all the explanatory variables were significant and the standard errors were close to each other (except for the unweighted multiple regression), it was concluded that the coefficients obtained from the HLM and multiple linear regression analyses showed remarkable differences. This result is in parallel with Roberts's (2004) observation that research findings differ significantly when the hierarchical data structure is not taken into consideration. The HLM analysis showed equal percentage of level-1 variance explained by the model in which only a plausible value was used, as well as the model in which all plausible values and weights were used together. This result may be in relation to the student-level explanatory variables versus the school-level weights. Besides, this situation is in compliance with Wu's (2005) conclusion in a simulation study that using any of the plausible values alone is enough to estimate the population parameters highly correctly.

The level-2 explanatory variables of class sizes, educational leadership, shortage of educational material and staff proved to be significant on both multiple linear regression models. However, for both HLMs, only the shortage of educational material and the shortage of educational staff were significant. This result stems from the fact that t values turn out to be higher than they must be, because the difference of level is ignored in the multiple linear regression analysis, and the level-2 variables in the nested data tend to be significant. Several other studies have also set forth that HLM is more effective in prediction and able to estimate the coefficients and related standard errors more accurately than the traditional analyses are (Gelman, 2006; Goldstein, 2011; Raudenbush, 1988). On both multiple linear regression and HLM method, using all plausible values in company with weights augmented the coefficient-related standard errors. In this case, it is possible to assert that the usage of weighting reduces the risk of type-2 errors for both analysis methods. In the HLM analysis, the use of all plausible values along with weights increased the percentage of variance explained, though they did not influence the coefficients much. Accordingly, using multiple plausible values and weights appears to enhance the performance of HLM analysis.

It was seen that all student- and school-level variables included in the model were significant factors affecting the students' overall science performance in each model. On the other hand, the variables with highest coefficients were the epistemological beliefs of students and the index of economic, social, and cultural status in the multiple regression analyses, while the level of epistemological beliefs and the shortage of educational material and staff were in the HLM analyses. The coefficients were estimated higher and the related standard errors were estimated lower in the multiple linear regression analyses than they were in the HLM analyses, even when all of the variables were included in the model. Regarding such hierarchical data, several other studies confirm that the results of HLM and those of the traditional linear models differ from each other (Bryk & Raudenbush, 2002; Gelman, 2006; Goldstein, 2011; Osborne, 2000; Raudenbush, 1988; Raudenbush & Bryk, 1986; Woltman et al., 2012). Using all coefficients in company with weights had a considerable effect on the standard errors of coefficients pertaining to the school-level variables in the multiple linear regression analyses, whereas it did not generate any remarkable effect on the HLM analyses. The use of all plausible values together with weights in the HLM analyses produced an effect, similar to that of previous models, on the percentage of variance explained at student and school levels. The percentage of student-level variance explained did not change, while that of school-level variance increased. The inclusion of school-level variables into the model has a different impact on the results, therefore. These results support the necessity of considering the differences of level during analyses.

When all plausible values are used in concurrence with weights, model coefficients do not increase to a considerable extent, though an increase is observed in the related standard errors, in cases that student- and school-level variables are included into the models separately or together in the multiple linear regression analysis. Any increase in standard errors has a bearing on t values, from which the significance of predictor variables is affected in turn. In this study, the variables included into the model were significant despite the decrease in t values. Thus, it is possible to argue that the usage of all plausible values in company with weights does not create a remarkable change on the parameters of multiple linear regression. Although this result is in parallel with the results of a study by Wu (2005) about the use of plausible values, it shows that the way of using survey weights as proposed by OECD (2017) does not generate any change on the outcome. This finding supported by the finding of Carle's (2009) study. Carle asserts that coefficients of weighted and unweighted models are slightly different from each other. However, standard errors diverge comparably. The coefficients and the related standard errors demonstrated a similar tendency in the HLM analysis. Notwithstanding that, the models in which all plausible values and weights were used in company proved to be more conservative in terms of significance and increased the percentage of variance explained in the HLM analysis, which makes it essentially usable in precise studies.

These research results indicate that the outcomes of using HLM for hierarchically structured data are different from those of the multiple linear regression analysis. Since multiple linear regression is not appropriate and adequate for nested data, HLM analysis should be preferred for that purpose. In this way, the separate and collective effects of explanatory variables at different levels will be observed, and the explanatory variables that predict the dependent variable will be determined accurately and reliably. In this study we used just student-level weights. Under similar conditions new studies can be conducted with scholl or higher level weights.

REFERENCES

Acar, T., & Öğretmen, T. (2012). Çok düzeyli istatistiksel yöntemler ile 2006 PISA fen bilimleri performansının incelenmesi. *Eğitim ve Bilim*, *37*(163). Retrieved from http://egitimvebilim.ted.org.tr/index.php/EB/article/download/1040/346

Adams, R. J., & Wu, M. L. (Eds.) (2002) *PISA 2000 technical report.* Paris: OECD Publications.

Atar, B. (2010). Basit doğrusal regresyon analizi ile hiyerarşik doğrusal modeller analizinin karşılaştırılması. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *1*(2), 78-84.

Atar, H. Y., & Atar, B. (2012). Examining the effects of Turkish education reform on students' TIMSS 2007 science achievements. *Educational Sciences: Theory and Practice*, *12*(4), 2632–2636.

Beaton, A.E. (1987). *Implementing the new design.* (The NAEP 1983-84 technical report, Report No. 15-TR-20). Princeton, NJ: Educational Testing Service.

Bock, R. D. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of the EM algorithm. *Psychometrika 46*, 443-459.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education 97*(1), 65-108.

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

Bryk, A. S., Raudenbush, S. W., & Congdon, R. (2010). HLM7 for Windows [Computer software]. Chicago, IL: Scientific Software International, Inc.

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, *9*(1), 1-13. doi: 10.1186/1471-2288-9-49

Chowa, G. A., Masa, R. D., Ramos, Y., & Ansong, D. (2015). How do student and school characteristics influence youth academic achievement in Ghana? A hierarchical linear modelling of Ghana Youth Save baseline data. *International Journal of Educational Development*, *45*, 129-140.

Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). New York, NY: John Wiley and Sons.

Fraenkel, J. R.; Wallen, N. E.; Hyun, H. H. (2012): *How to design and evaluate research in education* (8th Ed.). New York, NY: McGraw-Hill Humanities / Social Sciences/Languages.

Gelman, A. (2006). Multilevel (hierarchical) modelling: What it can and cannot do. *Technometrics 48*(3), 432-435.

Goldstein, H. (2011). *Multilevel statistical models* (Vol. 922). Oxford: John Wiley & Sons.

International Business Machines Corp. (2015). IBM SPSS Statistics for Windows (Version 23.0) [Computer software]. Armonk, NY: IBM Corp.

International Association for the Evaluation of Educational Achievement, (2016), Help Manual for the IDB Analyzer. Hamburg, Germany. Retrieved from www.iea.nl/data)

Lohr, S. (2010). *Sampling: Design and analysis* (2nd edition). Boston, MA: Brooks / Cole.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modelling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1*(3), 86-92. doi:10.1027/1614-2241.1.3.86

Meinck, S. (2015). Computing sampling weights in large-scale assessments in education [Special issue]. *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to' Approach*. Retrieved from https://surveyinsights.org/?p=5353

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56*(2), 177–196.

Mislevy, R. J. (1993). Should "multiple imputations" be treated as "multiple indicators"? *Psychometrika, 58*(1), 79–85.

Organization for Economic Cooperation and Development (2009). Analyses with plausible values. In *PISA Data Analysis Manual: SPSS*, (Second Edition), OECD Publishing. Retrieved from http://dx.doi.org/10.1787/9789264056275-9-en

Organization for Economic Cooperation and Development (2014). *PISA 2012 technical report*. Paris: OECD.

Organization for Economic Cooperation and Development (2017). *PISA 2015 Technical report*. Paris: OECD.

Osborne, J. W. (2000). Advantages of hierarchical linear modeling. *Practical Assessment, Research & Evaluation*, *7*(1), 1-3.

Palardy, G. J. (2011). Review of HLM 7. *Social Science Computer Review, 29*(4), 515–520. doi: 10.1177/0894439311413437

Rasch, G. (1960). *Studies in mathematical psychology: I. probabilistic models for some intelligence and attainment tests*. Oxford, England: Nielsen & Lydiche.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, *13*(2), 85-116.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*(1), 1-17.

Roberts, J. K. (2004). An introductory primer on multilevel and hierarchical linear modelling. *Learning Disabilities: A Contemporary Journal 2*, 30-38.

Rubin, D. B. (1987). *Multiple imputations for non-response in surveys.* New York, NY: Wiley.

Särndal, C., Swensson, B. & Wretman, J. (1992). *Model assisted survey sampling*. New York, NY: Springer-Verlag.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis. An introduction to basic and advanced multilevel modeling.* London: Sage.

Snijders, T., & Bosker, R. (2003). *Multilevel analysis: An introduction to basic and applied multilevel analysis.* Thousand Oaks, CA: Sage Publications.

Stipek, D., & Valentino, R. A. (2015). Early childhood memory and attention as predictors of academic growth trajectories. *Journal of Educational Psychology*, *107*(3), 771-788.

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

Warm, T. A. (1985). *Weighted maximum likelihood estimation of ability in item response theory with tests q/jinite length.* (Technical Report No. CGI-TR-85-08). Oklahoma, OK: Coast Guard Institute.

Woltman, H., Feldstain, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchical linear modeling. *Tutorials in Quantitative Methods for Psychology*, *8*(1), 52-69.

Wright. B.D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

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

Yang, H. (2013). The case for being automatic: Introducing the automatic linear modeling (LINEAR) procedure in SPSS statistics. *Multiple Linear Regression Viewpoints*, *39*(2), 27–37.