Erzincan Üniversitesi Eğitim Fakültesi Dergisi
Erzincan University Journal of Education Faculty
2024 Cilt 26 Sayı 3 (400-413) https://doi.org/10.17556/erziefd.1404346
Araştırma Makalesi / Research Article

# The Effect of Sample Weighting on Hierarchical Linear Modeling in the Large-Scale Assessment Data

# Geniş Ölçekli Test Verilerinde Örneklem Ağırlıklandırmalarının Hiyerarşik Doğrusal Modellemeye Etkisi

**Metehan Güngör**[1] **Sinan M. Bekmezci**[2] **Nuri Doğan**[3]

[1] Doctoral student, Ankara University, Faculty of Educational Sciences, Ankara, Türkiye
[2] Dr., Manisa Celal Bayar University, Education Faculty, Manisa, Türkiye
[3] Professor, Hacettepe University, Department of Educational Sciences, Ankara, Türkiye

**Makale Bilgileri**

*Geliş Tarihi (Received Date)*

15.12.2023

*Kabul Tarihi (Accepted Date)*

29.06.2024

***Sorumlu Yazar***

Metehan Güngör

Ankara University Cebeci Campus, Faculty of Educational Sciences, Cebeci, Ankara

gungormetehan@gmail.com

***Abstract:*** This study examines how the different uses of sampling weights in the analysis of TIMSS 2019 data affect the ratio of variance in student achievement explained by schools and the estimation of standard errors. The research sample comprises 227,345 8th grade students from 7,636 schools in 39 countries. Mathematics achievement and science achievement are considered separately as dependent variables in all 39 countries. All plausible values are included in the analysis. Four weighting scenarios are examined: no weighting, weighting at only level 1, weighting at only level 2, and weighting at both levels. In total, 312 models are established and examined. According to the research results, the coefficients, standard errors, reliabilities, and $\chi^2$ estimations change depending on how the weighting variable is handled in the models, and as a result, the ratio of variance in the dependent variable arising from the differences between schools also changes. The ratio attributable to between-school differences can reach up to 20% in some countries. Therefore, researchers modeling hierarchical data using HLM are suggested to plan how they handle the weighting variable prior to conducting the study.

**Keywords:** Large-scale assessment, hierarchical linear modeling, weighting, TIMSS 2019

***Öz:*** Bu çalışmada TIMSS 2019 verilerinin analizinde örneklem ağırlıklarının farklı şekilde kullanımlarının öğrenci başarısındaki varyansın okullar tarafından açıklanan kısmında ve standart hataların kestiriminde nasıl bir etkiye sahip olduğu incelenmiştir. Araştırmanın örneklemini TIMSS 2019 uygulamasına katılan 39 ülkeden toplam 7636 okuldaki 227345 8. sınıf öğrencisi oluşturmaktadır. Matematik başarısı ve fen başarısı tüm ülkelerde bağımlı değişkenler olarak ayrı ayrı ele alınmıştır. Tüm olası değerler analize dahil edilmiştir. Ağırlıklandırmanın olmadığı, yalnızca 1. düzeyde ağırlıklandırmanın olduğu, yalnızca 2. düzeyde ağırlıklandırmanın olduğu ve her iki düzeyde de ağırlıklandırmanın olduğu dört farklı durum incelenmiştir. Toplamda 312 model kurulmuş ve çözümlenmiştir. Araştırmanın sonuçlarına göre katsayıların, standart hataların, güvenirlik değerlerinin ve $\chi^2$ istatistiklerinin, ağırlıklandırma değişkeninin kullanılma biçimine göre değiştiği gözlenmiştir. Bunun bir sonucu olarak da, çıktı değişkenindeki varyansın açıklanmasında okullar tarafından açıklanan kısım değişkenlik göstermektedir. Bu kısımdaki değişkenlik bazı ülkelerde %20'lere kadar çıkabilmektedir. Bu nedenle, geniş ölçekli testlerin verilerini HLM ile modelleyecek araştırmacıların ağırlıklandırma değişkenini ne şekilde ele alacaklarını araştırma öncesinde planlamaları önerilmektedir.

**Anahtar Kelimeler:** Geniş ölçekli test, hiyerarşik doğrusal modelleme, ağırlıklandırma, TIMSS 2019

Güngör, M., Bekmezci, S. M. & Doğan, N. (2024). The effect of sample weighting on hierarchical linear modeling in the large-scale assessment data. *Erzincan University Journal of Education Faculty, 26*(3), 400-413. https://doi.org/10.17556/erziefd.1404346

## Introduction

Large-scale tests such as TIMSS (Trends in International Mathematics and Science Study) and PISA (Programme for International Student Assessment) aim to facilitate international comparisons and analyze national-level trends. The influence of these tests on global education policies and reforms is continuously growing (Ababneh et al., 2016; Barber et al., 2010; Manjunath, 2021; Schmidt et al., 1997; Tobin et al., 2015). Consequently, analyzing the large-scale datasets collected in these assessments can provide valuable guidance for countries' education systems. Given that education researchers' findings can be evaluated by those involved in decision-making processes and shaping educational systems, researchers should be cautious when selecting the appropriate model for their analyses. Hierarchical linear modeling (HLM) is widely used in our country and worldwide, particularly in the analysis of data from large-scale tests. HLM is especially preferred when stratified sampling is employed, as is the case in the large-scale assessments. There is an abundance of studies utilizing HLM, conducted with data from large-scale tests (Aksu et al., 2017; Atar & Atar, 2012; Bilican & Yıldırım, 2013; Boulifa & Kaaouachi, 2022; Chu et al., 2014; Gómez & Suárez, 2020; Liang, 2010; Pacheco Diaz & Rocconi, 2021; Pong, 2009; Reinikainen, 2007; Ross, 2008; Saal et al., 2019; Sabudin et al., 2018; Sun et al., 2012; Thien et al., 2015; Valente et al., 2011; Woo & Henfield, 2016). However, some of these studies lack sufficient details on the usage of sample weight variables. In a study by Özdemir (2016), which examined the methodological aspects of analyses conducted with PISA Turkey data, it was reported that the majority of the 97 examined studies did not indicate whether plausible values and sampling weights were considered. Only five articles reported the correct usage of sampling weights, while plausible values were utilized appropriately in only 10 articles. In a similar review conducted by Liou and Hung (2015), numerous studies utilizing PISA and TIMSS data were observed not to specify whether sampling weights were used. It has been emphasized that this situation raises doubts about the reliability of the conducted analyses. Furthermore, articles have been written criticizing the impact of the analysis outputs derived from large-scale test data on education policies (Sahlberg & Hargreaves, 2015; Schleicher, 2015; Takayama, 2015). These criticisms clearly highlight the importance of exercising caution in statistical analysis.

HLM is a statistical model that is designed to take into account the hierarchical or nested structure of the data. This analysis technique allows for the examination and modeling of relationships at multiple hierarchical levels by utilizing different variables at each level (Hox, 2010; Raudenbush &

Bryk, 2002). HLM allows for the estimation of regression coefficients and standard errors for each level and variable. Additionally, in HLM, shared variances at each level are integrated into the model to facilitate predictions in subsequent stages (Woltman et al., 2012). These features render HLM a valuable tool for analyzing multilevel datasets. Consequently, HLM represents a robust technique that can be employed to investigate disparities in achievement among schools.

Researchers utilizing HLM for analyzing educational data often come across the concepts of plausible values and weighting. In large-scale tests, it is crucial to estimate students' achievement with high reliability. However, it can be challenging to administer a large number of test items to students. To address this issue, TIMSS employs the plausible value methodology, which allows for estimating students' achievement using a reduced number of items. According to OECD (2017), plausible values are random draws from posterior distribution of ability of individuals. It is important to note that plausible values do not represent individual test scores, and using them as such may lead to biased estimations of student competencies. However, when correctly grouped, plausible values may provide unbiased estimates of sample statistics, such as group means, standard deviations, and variances. In TIMSS, five plausible values are reported for the estimation of student achievement, while PISA reports ten. Laukaityte and Wiberg (2017) found in their simulation studies that utilizing multiple plausible values increased prediction accuracy and reduced errors. Therefore, it is necessary to conduct analyses using models that account for plausible values collectively.

Weighting is another concept encountered in statistical methods. When selecting a sample for large-scale tests, the aim is to choose a sample that can best represent the population. The sample should be able to accurately predict the characteristics of the entire population in order to produce meaningful results. One commonly used method to analyze the differences between the sample and the population is the utilization of sample weights (Rust, 2013). If we consider a population as a classroom consisting of 10 females and 20 males, we can provide an example of weighted and unweighted means. Let's draw 8 students from this population, with four males and four females. As observed, the selected sample does not adequately represent the population. When four females are selected from the 10 female students, the probability of each female student being selected is $^4/_{10} = 0.4$. Similarly, when four males are selected from the 20 male students, the probability of each male student being selected is $^4/_{20} = 0.2$ (assuming equal probabilities of selection for all students). In this case, the weight of each female student in the sample is 2.5, while the weight of each male student is 5. Suppose these students have obtained scores of 70, 80, 80, 90 for the female students, and 50, 60, 60, 70 for the male students, in a 100-point exam. As observed, the unweighted mean of the sample is $(70 + 80 + 80 + 90 + 50 + 60 + 60 + 70)/8 = 70$, while the weighted mean of the sample is $[2.5 * (70 + 80 + 80 + 90) + 5 * (50 + 60 + 60 + 70)]/30 = 66.67$. As observed in this example, there is a distinction between the weighted and unweighted means. Conversely, when each individual has an equal probability of selection and equal sample weights, the weighted and unweighted means will be identical. However, in TIMSS applications, a two-stage stratified sampling method is utilized, and weight variables are incorporated in the shared data sets to account for potential discrepancies between the sample characteristics and the population. In multilevel analyses, sample weights should be disaggregated based on the appropriate levels (Rutkowski et al., 2010). It is important to note that the results obtained may be less reliable if sample weights are not utilized (OECD, 2017).

In this study, it was examined how the ratio of the variance in the achievement scores of students in the countries participating in the TIMSS 2019 study that arises from differences between schools, and the standard errors, change depending on whether or not weighting variables are used. Two research questions taken into consideration in this direction were determined as follows:

1. According to the TIMSS 2019 results, what ratio of the total variation in mathematics and science achievement scores of 8th grade students in countries participating in the assessment can be attributed to differences between schools in the following situations: (a) when no weighting variables are used, (b) when only level 1 weighting variables are used, (c) when only level 2 weighting variables are used, and (d) when weighting variables are used at both levels?
2. How do the standard errors of the estimated mean mathematics and science achievement scores of students in countries participating in the TIMSS 2019 assessment change in the following situations: (a) when no weighting variables are used, (b) when only level 1 weighting variables are used, (c) when only level 2 weighting variables are used, and (d) when weighting variables are used at both levels?

## Method

In this study, the aim was to investigate the relationship between the utilization of weighting variables in various manners and the variance explained in hierarchical data derived from large-scale test applications. Therefore, this research is classified as a descriptive study (Fraenkel & Wallen, 2006).

### Population and Sample

In this research, data from students participating in the TIMSS 2019 application were used. TIMSS employs a two-stage random samples design, with a sample of schools drawn as the first stage and one or more intact classes of students selected from each of the sampled schools as the second stage. In this study, data from 227,345 8th grade students from 7,636 schools in 39 countries participating in the TIMSS 2019 application were analyzed. The number of schools in the participating countries ranges from 98 to 623, while the number of students varies between 3,265 and 22,334.

### Data Collection

The data used for the analyses are derived from the student and school questionnaires of TIMSS 2019. These data can be accessed on the internet at https://timss2019.org/international-database/. Only the BSA***M7 (ID, weighting, and plausible value variables for students) and BCG***M7 (ID and weighting variables for schools) datasets provided by TIMSS were used for the analyses.

### Data Analysis

In accordance with the recommendations of Arıkan and his colleagues (2020), weight variables (both at the student and

school levels) were included in datasets. In this research, the first-level (student) sample weight is calculated as the product of the variables WGTADJ2 * WGTFAC2 * WGTADJ3 * WGTFAC3 (CLASS WEIGHT ADJUSTMENT * CLASS WEIGHT FACTOR * STUDENT WEIGHT ADJUSTMENT * STUDENT WEIGHT FACTOR) in the TIMSS 2019 data set. The second-level (school) sample weight is calculated as the product of the variables WGTADJ1 * WGTFAC1 (SCHOOL WEIGHT ADJUSTMENT * SCHOOL WEIGHT FACTOR) in the TIMSS 2019 dataset. The product of the weight variables obtained at the first and second levels precisely corresponds to the TOTWGT (TOTAL STUDENT WEIGHT) shared by TIMSS. Adding variables and examining residuals for the model were performed using IBM SPSS Statistics 22 (IBM Corp., 2013), while other data analyses were conducted using HLM 6.04 program (Raudenbush et al., 2007).

Analyses are conducted using four sequential sub-models commonly used in HLM. However, since this study specifically examines the impact of the weighting variable and the standard errors of the estimations, only the first model, which is the random effect one-way ANOVA model, was established.

Before conducting the HLM analyses, certain assumptions need to be satisfied. In terms of examining these assumptions, as stated by Raudenbush and Bryk (2002), the following basic assumptions were considered: normal distributions and homogeneity of residuals, and independence of errors at both within and between-levels. The first assumption, which states that the residuals should be normally distributed, was evaluated by analyzing the residuals of the models. Graphical methods such as histograms and Q-Q plots were used to assess the distribution of the residuals. Based on the examination of these graphs, this can be concluded that there are no significant deviations from a normal distribution.

Mathematics achievement was taken as the dependent variable, and in models without weighting, the skewness values of the residuals ranged from -0.38 to 0.57, while the kurtosis values ranged from -0.23 to 0.56. In models where weighting was applied only at the first level, the skewness values ranged from -0.4 to 0.42, and the kurtosis values ranged from -0.99 to 0.42. In models where weighting was applied only at the second level, the skewness values ranged from -0.39 to 0.26, and the kurtosis values ranged from -1 to 1.4. In the models where weighting was applied at both levels, the skewness values ranged from -0.39 to 0.26, and the kurtosis values ranged from -0.24 to 1.45. Science achievement was taken as the dependent variable, and in models without weighting, the skewness values of the residuals ranged from -0.4 to 0.05, while the kurtosis values ranged from -0.2 to 0.33. In models where weighting was applied only at the first level, the skewness values ranged from -0.41 to 0.23, and the kurtosis values ranged from -0.21 to 0.52. In models where weighting was applied only at the second level, the skewness values ranged from -0.42 to 0.09, and the kurtosis values ranged from -0.24 to 1.97. In the models where weighting was applied at both levels, the skewness values ranged from -0.42 to 0.09, and the kurtosis values ranged from -0.24 to 1.98. The homogeneity of variances was tested using the $\chi^2$ statistic. However, according to the results, it was seen that the variances were not homogeneous. On the other hand, it is known that the $\chi^2$ statistic is sensitive to type 1 error and tends to make a statistically significant difference in large samples. For this reason, the normal distribution of the residuals was considered sufficient and the analyses were continued. It is also important to provide sufficient sample size at each level. The sample defined in this study has sufficient size in terms of both the number of schools and students.

To address the research questions considered, four models were established with 39 different countries' data for two different dependent variables (mathematics and science achievement). In total, 312 models were established and examined. The reliability values (directly provided by the software) for the models with mathematics achievement as the dependent variable range between .72 and .99, and the reliability values for the models with science achievement as the dependent variable range between .63 and .98. The reliability estimation values can be found in Appendix A.

## Findings

In this study, a total of 312 distinct models were established, and the aim was to address two research questions by analyzing the statistical data from these models. The results are presented in two distinct sections.

Table 1 presents the statistics for the models where mathematics achievement is considered as the dependent variable, while Table 2 provides the statistics for the models where science achievement is considered as the dependent variable. In Table 1 and Table 2, the following statistics are presented for each model: regression coefficients ($\gamma_{00}$) for each country, the standard errors of these coefficients, the $\chi^2$ values of the established model, the ratios or percentages of the total variance in the dependent variable explained by the schools ($\rho = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00}+\sigma^2}$, also known as intraclass correlation coefficient), and degrees of freedom. These statistics are provided for four different weightings: no weighting, only level 1 weighting, only level 2 weighting, and weighting at both levels.

Upon close examination of the established models (Table 1 and Table 2), it is evident that the presence of a weight variable in various forms in the USA, Morocco, and Jordan countries leads to significant changes in standard errors. As a result, the findings obtained from the data of these countries were excluded from interpretation. However, it was decided to include the statistics of these countries in both tables and figures to enable other researchers to examine the findings.

### Findings Related to the 1st Research Question

As indicated in Table 1 and Table 2, the percentage of variance in students' mathematics achievement scores explained by school differences ranges from 9.18% (Bahrain) to 60.51% (Hong Kong). Similarly, the percentage of variance in science achievement scores explained by school differences ranges from 6.30% (Korea) to 61.93% (UAE). In addition to significant variations across countries, it has been observed that the percentages reported in the data of the same country also differ depending on how the weighting is applied. Figure 1 illustrates the variation in the percentage of the variance in mathematics achievement scores attributed to school differences, depending on the treatment of the weighting variable in the models.

**Table 1.** Statistics on models using mathematics achievement as the dependent variable

| Code | Country | Coefficient | | | | Standard Error | | | | Chi-Square | | | | Variance – Between Schools (%) | | | | *df* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | No | Level 1 | Level 2 | Full | No | Level 1 | Level 2 | Full | No | Level 1 | Level 2 | Full | No | Level 1 | Level 2 | Full | |
| AUS | Australia | 520.480 | 510.810 | 519.882 | 511.155 | 3.369 | 4.128 | 6.359 | 6.258 | 5222.13 | 5777.34 | 5626.22 | 5840.00 | 35.56 | 38.74 | 35.19 | 36.25 | 283 |
| BHR | Bahrain | 482.315 | 481.842 | 482.315 | 482.330 | 3.047 | 3.356 | 3.047 | 3.049 | 686.58 | 656.84 | 686.58 | 687.21 | 9.45 | 9.18 | 9.45 | 9.47 | 111 |
| CHL | Chile | 445.068 | 460.572 | 427.024 | 427.024 | 4.429 | 6.926 | 3.770 | 3.770 | 3591.58 | 4265.29 | 2489.90 | 2489.90 | 46.50 | 51.54 | 34.02 | 34.02 | 163 |
| TWN | Chinese | 606.392 | 618.246 | 593.986 | 593.986 | 3.647 | 3.759 | 4.420 | 4.420 | 1555.73 | 1346.18 | 1542.02 | 1542.02 | 21.01 | 18.65 | 20.23 | 20.23 | 202 |
| CYP | Cyprus | 500.820 | 500.708 | 500.820 | 500.924 | 3.676 | 3.723 | 3.676 | 3.682 | 820.45 | 773.32 | 820.45 | 822.79 | 17.16 | 16.23 | 17.16 | 17.21 | 97 |
| EGY | Egypt | 413.646 | 413.498 | 417.819 | 417.819 | 4.721 | 5.306 | 9.188 | 9.188 | 4508.58 | 4365.50 | 5532.57 | 5532.57 | 39.13 | 37.66 | 46.26 | 46.26 | 168 |
| ENG | England | 509.891 | 509.207 | 507.230 | 507.236 | 5.521 | 5.855 | 7.354 | 7.353 | 2930.07 | 2989.15 | 3531.53 | 3533.91 | 49.16 | 49.30 | 53.36 | 53.37 | 135 |
| FIN | Finland | 508.673 | 510.556 | 507.156 | 507.075 | 2.420 | 2.533 | 2.854 | 2.838 | 823.80 | 780.01 | 860.26 | 843.87 | 12.38 | 11.95 | 11.59 | 11.37 | 153 |
| FRA | France | 483.270 | 480.881 | 482.023 | 482.014 | 3.046 | 3.554 | 2.933 | 2.933 | 1426.86 | 1640.19 | 1227.88 | 1227.62 | 24.85 | 27.36 | 22.05 | 22.05 | 149 |
| GEO | Georgia | 459.089 | 461.197 | 459.640 | 459.627 | 4.793 | 5.099 | 7.787 | 7.787 | 1800.69 | 1784.84 | 2118.71 | 2119.27 | 34.52 | 33.17 | 41.66 | 41.66 | 144 |
| HKG | Hong Kong | 572.600 | 581.674 | 561.817 | 561.816 | 6.114 | 6.993 | 6.821 | 6.821 | 4017.57 | 4474.92 | 3895.03 | 3894.31 | 57.64 | 60.51 | 55.97 | 55.96 | 135 |
| HUN | Hungary | 515.550 | 516.909 | 495.654 | 495.717 | 4.665 | 4.451 | 6.034 | 6.039 | 3096.97 | 2826.20 | 3678.88 | 3685.26 | 39.22 | 36.87 | 41.97 | 42.02 | 153 |
| IRN | Iran | 444.320 | 451.493 | 432.387 | 432.387 | 4.075 | 3.900 | 6.597 | 6.597 | 4144.97 | 3464.07 | 4642.35 | 4642.35 | 39.72 | 35.71 | 41.84 | 41.84 | 219 |
| IRL | Ireland | 523.342 | 522.829 | 518.853 | 518.793 | 2.986 | 3.529 | 4.063 | 4.061 | 1092.54 | 1205.66 | 1219.95 | 1220.47 | 20.40 | 22.67 | 21.68 | 21.68 | 148 |
| ISR | Israel | 511.526 | 514.261 | 517.152 | 517.152 | 5.966 | 7.079 | 6.525 | 6.525 | 4312.58 | 4973.04 | 3475.27 | 3475.27 | 53.23 | 56.72 | 47.35 | 47.35 | 156 |
| ITA | Italy | 498.990 | 497.807 | 498.009 | 498.009 | 3.029 | 3.176 | 3.120 | 3.120 | 988.13 | 1002.35 | 980.87 | 977.67 | 19.22 | 19.29 | 18.34 | 18.29 | 157 |
| JPN | Japan | 592.420 | 594.446 | 590.605 | 590.605 | 2.740 | 2.568 | 3.814 | 3.814 | 766.78 | 631.43 | 938.75 | 938.75 | 11.56 | 9.61 | 13.25 | 13.25 | 141 |
| JOR | Jordan | 416.420 | 415.841 | 425.053 | 425.053 | 3.212 | 3.707 | 6.597 | 6.597 | 2993.51 | 2958.10 | 4172.09 | 4172.09 | 29.28 | 28.98 | 36.79 | 36.79 | 234 |
| KAZ | Kazakhstan | 490.124 | 489.187 | 490.527 | 490.425 | 4.374 | 5.700 | 5.791 | 5.805 | 3122.09 | 3555.97 | 3184.12 | 3289.89 | 42.18 | 45.40 | 39.24 | 39.82 | 167 |
| KOR | Korea | 604.411 | 610.150 | 594.626 | 594.626 | 3.033 | 3.264 | 5.191 | 5.191 | 626.11 | 627.57 | 685.42 | 685.42 | 10.56 | 10.69 | 11.27 | 11.27 | 167 |
| KWT | Kuwait | 399.041 | 397.280 | 404.440 | 404.440 | 4.227 | 4.339 | 5.711 | 5.711 | 2806.81 | 2632.68 | 3168.42 | 3168.42 | 35.96 | 34.50 | 40.05 | 40.05 | 170 |
| LBN | Lebanon | 422.896 | 428.754 | 417.617 | 417.617 | 3.317 | 4.478 | 4.373 | 4.373 | 3363.43 | 3803.20 | 3700.24 | 3700.24 | 40.30 | 43.49 | 42.44 | 42.44 | 203 |
| LTU | Lithuania | 505.411 | 513.875 | 502.892 | 502.847 | 3.722 | 3.813 | 4.810 | 4.805 | 1842.54 | 1739.67 | 1582.97 | 1577.47 | 29.89 | 29.16 | 26.59 | 26.52 | 193 |
| MYS | Malaysia | 498.915 | 493.257 | 472.715 | 463.452 | 5.191 | 6.463 | 4.829 | 5.013 | 5787.50 | 8938.72 | 3967.26 | 4844.59 | 47.77 | 58.64 | 38.61 | 43.81 | 176 |
| MAR | Morocco | 386.592 | 384.900 | 398.599 | 398.599 | 2.466 | 2.444 | 6.889 | 6.889 | 3086.31 | 2726.79 | 5563.80 | 5563.81 | 27.56 | 24.19 | 42.20 | 42.20 | 250 |
| NZL | New Zealand | 496.266 | 484.911 | 484.974 | 473.634 | 5.080 | 4.835 | 7.933 | 7.715 | 2986.06 | 2406.61 | 3205.61 | 2979.94 | 34.02 | 28.87 | 37.51 | 36.01 | 133 |
| NOR | Norway | 504.543 | 505.618 | 500.182 | 500.251 | 2.384 | 2.477 | 3.809 | 3.810 | 694.58 | 704.92 | 728.72 | 728.90 | 10.04 | 9.92 | 10.06 | 10.06 | 156 |
| OMN | Oman | 411.601 | 412.462 | 411.843 | 411.843 | 3.934 | 4.375 | 4.418 | 4.418 | 3092.71 | 3115.10 | 2587.84 | 2587.51 | 29.77 | 29.75 | 25.93 | 25.93 | 227 |
| PRT | Portugal | 499.205 | 497.489 | 497.858 | 497.783 | 3.365 | 3.498 | 4.304 | 4.303 | 1369.58 | 1307.75 | 1478.70 | 1473.49 | 26.80 | 25.87 | 28.40 | 28.34 | 155 |
| QAT | Qatar | 451.356 | 443.669 | 452.508 | 452.520 | 5.523 | 6.910 | 5.522 | 5.521 | 3537.38 | 3710.18 | 3556.46 | 3551.89 | 47.54 | 50.19 | 47.55 | 47.52 | 151 |
| ROM | Romania | 477.798 | 487.470 | 460.716 | 460.716 | 4.652 | 4.865 | 5.437 | 5.437 | 3380.50 | 3081.77 | 3270.30 | 3268.76 | 40.74 | 39.18 | 39.40 | 39.39 | 197 |
| RUS | Russia | 542.196 | 549.700 | 535.053 | 535.053 | 3.880 | 4.978 | 4.817 | 4.817 | 3040.16 | 3181.88 | 2686.88 | 2686.88 | 41.82 | 44.07 | 36.74 | 36.74 | 203 |
| SAU | Saudi Arabia | 408.619 | 410.467 | 390.219 | 390.218 | 3.909 | 4.311 | 4.061 | 4.061 | 3188.90 | 3007.81 | 2092.76 | 2092.89 | 37.76 | 36.00 | 26.83 | 26.83 | 208 |
| SGP | Singapore | 609.480 | 614.958 | 609.480 | 608.208 | 5.047 | 5.115 | 5.047 | 5.100 | 4154.15 | 4139.18 | 4154.15 | 4230.20 | 45.63 | 45.51 | 45.63 | 46.12 | 152 |
| ZAF | South Africa | 408.602 | 417.843 | 378.823 | 378.823 | 2.971 | 3.364 | 4.665 | 4.665 | 22525.12 | 23923.03 | 17005.42 | 17005.42 | 55.99 | 56.97 | 51.02 | 51.02 | 518 |
| SWE | Sweden | 502.465 | 501.581 | 501.797 | 501.708 | 3.050 | 3.208 | 4.660 | 4.658 | 1012.40 | 1005.75 | 1137.10 | 1135.89 | 18.85 | 18.56 | 22.30 | 22.28 | 149 |
| TUR | Türkiye | 489.358 | 498.498 | 477.252 | 477.252 | 5.111 | 5.096 | 8.468 | 8.468 | 2276.40 | 1840.36 | 2731.98 | 2731.98 | 34.99 | 30.29 | 39.08 | 39.08 | 180 |
| ARE | UAE | 467.908 | 472.292 | 468.179 | 468.112 | 3.105 | 4.076 | 3.039 | 3.042 | 30235.27 | 34427.62 | 28822.38 | 28878.57 | 56.52 | 59.62 | 55.25 | 55.29 | 622 |
| USA | United States | 516.409 | 520.033 | 482.754 | 483.021 | 4.161 | 5.320 | 22.947 | 23.066 | 7333.55 | 8095.30 | 9591.03 | 9585.68 | 45.69 | 48.52 | 62.30 | 62.52 | 272 |

- The degrees of freedom have different values in the models fitted with the data of South Africa (ZAF) and Italy (ITA).

**Table 2.** Statistics on models using science achievement as the dependent variable

| Code | Country | Coefficient | | | | Standard Error | | | | Chi-Square | | | | Variance – Between Schools (%) | | | | *df* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | No | Level 1 | Level 2 | Full | No | Level 1 | Level 2 | Full | No | Level 1 | Level 2 | Full | No | Level 1 | Level 2 | Full | |
| AUS | Australia | 533.620 | 524.344 | 533.640 | 525.830 | 2.981 | 3.823 | 5.494 | 5.505 | 3610.68 | 4034.26 | 3764.35 | 3977.29 | 27.51 | 30.67 | 26.77 | 27.94 | 283 |
| BHR | Bahrain | 488.153 | 489.116 | 488.153 | 488.149 | 5.375 | 5.823 | 5.375 | 5.376 | 2195.56 | 1969.06 | 2195.56 | 2197.12 | 27.62 | 25.58 | 27.62 | 27.63 | 111 |
| CHL | Chile | 465.736 | 478.533 | 450.707 | 450.707 | 4.129 | 6.176 | 3.672 | 3.672 | 2569.35 | 2962.89 | 1882.80 | 1882.80 | 38.03 | 42.20 | 27.85 | 27.85 | 163 |
| TWN | Chinese | 570.061 | 578.912 | 560.043 | 560.043 | 2.570 | 2.600 | 3.414 | 3.414 | 1130.17 | 975.40 | 1156.34 | 1156.34 | 15.44 | 13.42 | 15.25 | 15.25 | 202 |
| CYP | Cyprus | 484.057 | 483.706 | 484.057 | 484.156 | 3.756 | 3.549 | 3.756 | 3.757 | 703.43 | 624.96 | 703.43 | 705.49 | 15.66 | 13.76 | 15.66 | 15.70 | 97 |
| EGY | Egypt | 390.194 | 392.185 | 391.046 | 391.046 | 5.343 | 6.260 | 9.291 | 9.291 | 3815.81 | 3951.49 | 4185.47 | 4185.47 | 34.99 | 35.31 | 39.02 | 39.02 | 168 |
| ENG | England | 512.230 | 511.424 | 511.939 | 511.944 | 5.243 | 5.519 | 7.328 | 7.329 | 2173.52 | 2197.30 | 2718.17 | 2720.65 | 41.29 | 41.22 | 46.67 | 46.68 | 135 |
| FIN | Finland | 542.441 | 544.262 | 540.500 | 540.418 | 2.930 | 3.031 | 3.007 | 2.996 | 871.03 | 838.01 | 879.06 | 864.97 | 13.16 | 12.96 | 11.65 | 11.48 | 153 |
| FRA | France | 489.209 | 486.502 | 488.147 | 488.140 | 3.223 | 3.700 | 3.333 | 3.333 | 1194.82 | 1345.25 | 1113.88 | 1114.46 | 21.85 | 23.75 | 20.56 | 20.57 | 149 |
| GEO | Georgia | 444.202 | 445.102 | 448.187 | 448.170 | 4.188 | 4.307 | 6.085 | 6.086 | 1174.56 | 1144.69 | 1413.76 | 1414.98 | 25.51 | 24.29 | 29.79 | 29.79 | 144 |
| HKG | Hong Kong | 497.591 | 504.854 | 495.949 | 495.947 | 5.973 | 7.039 | 6.649 | 6.649 | 3345.06 | 3932.45 | 3236.40 | 3236.34 | 50.96 | 55.45 | 49.68 | 49.68 | 135 |
| HUN | Hungary | 528.699 | 529.855 | 511.704 | 511.740 | 4.080 | 3.831 | 5.484 | 5.486 | 2517.96 | 2281.12 | 3098.97 | 3100.65 | 34.84 | 32.30 | 38.65 | 38.67 | 153 |
| IRN | Iran | 448.035 | 455.058 | 437.222 | 437.222 | 3.954 | 3.786 | 6.622 | 6.622 | 4128.23 | 3372.59 | 4880.42 | 4880.42 | 39.85 | 35.24 | 43.54 | 43.54 | 219 |
| IRL | Ireland | 522.761 | 522.326 | 517.956 | 517.885 | 3.313 | 3.946 | 4.868 | 4.866 | 1030.91 | 1157.54 | 1176.80 | 1177.08 | 19.47 | 21.96 | 21.16 | 21.16 | 148 |
| ISR | Israel | 506.699 | 510.286 | 508.115 | 508.115 | 5.508 | 6.381 | 6.017 | 6.017 | 3290.84 | 3637.68 | 2853.73 | 2853.73 | 46.42 | 48.85 | 41.93 | 41.93 | 156 |
| ITA | Italy | 502.171 | 501.309 | 500.108 | 500.096 | 2.864 | 2.985 | 2.963 | 2.960 | 952.20 | 949.86 | 946.36 | 944.19 | 18.86 | 18.68 | 17.88 | 17.85 | 157 |
| JPN | Japan | 568.553 | 569.829 | 567.036 | 567.036 | 2.030 | 2.007 | 2.960 | 2.960 | 544.11 | 460.95 | 654.75 | 654.75 | 7.91 | 6.64 | 8.98 | 8.98 | 141 |
| JOR | Jordan | 448.670 | 448.207 | 455.005 | 455.005 | 3.665 | 4.141 | 7.159 | 7.159 | 2995.82 | 2867.38 | 4271.49 | 4271.49 | 28.94 | 28.18 | 36.65 | 36.65 | 234 |
| KAZ | Kazakhstan | 480.557 | 476.640 | 480.877 | 480.603 | 4.183 | 5.506 | 5.403 | 5.417 | 2362.56 | 2843.02 | 2251.74 | 2323.89 | 35.21 | 39.78 | 30.44 | 30.95 | 167 |
| KOR | Korea | 559.234 | 562.541 | 556.040 | 556.040 | 2.236 | 2.473 | 2.713 | 2.713 | 480.86 | 499.72 | 462.32 | 462.32 | 7.45 | 8.01 | 6.30 | 6.30 | 167 |
| KWT | Kuwait | 441.803 | 441.390 | 443.740 | 443.740 | 4.737 | 4.870 | 6.238 | 6.238 | 2754.04 | 2654.54 | 3059.62 | 3059.62 | 36.08 | 35.07 | 39.55 | 39.55 | 170 |
| LBN | Lebanon | 366.924 | 377.412 | 356.138 | 356.138 | 5.561 | 7.106 | 7.101 | 7.101 | 4370.13 | 4916.06 | 4793.00 | 4793.00 | 48.12 | 51.00 | 50.22 | 50.22 | 203 |
| LTU | Lithuania | 516.301 | 525.349 | 517.918 | 517.861 | 3.684 | 3.688 | 4.365 | 4.360 | 1930.56 | 1752.33 | 1508.75 | 1502.08 | 32.05 | 30.09 | 25.34 | 25.25 | 193 |
| MYS | Malaysia | 499.153 | 490.967 | 475.223 | 464.661 | 4.924 | 6.308 | 4.983 | 5.260 | 4301.78 | 7063.54 | 3156.91 | 4120.12 | 40.91 | 52.60 | 33.85 | 39.99 | 176 |
| MAR | Morocco | 391.557 | 389.379 | 398.629 | 398.629 | 3.310 | 3.662 | 5.686 | 5.686 | 3907.71 | 3929.98 | 4960.62 | 4960.55 | 31.49 | 31.28 | 36.74 | 36.74 | 250 |
| NZL | New Zealand | 513.099 | 502.137 | 505.137 | 494.740 | 4.990 | 4.817 | 8.421 | 8.432 | 2759.86 | 2338.80 | 2984.33 | 2886.95 | 32.34 | 28.25 | 36.28 | 35.77 | 133 |
| NOR | Norway | 496.591 | 496.891 | 495.666 | 495.717 | 3.033 | 3.122 | 4.636 | 4.638 | 706.68 | 714.22 | 767.44 | 769.21 | 10.35 | 10.29 | 10.99 | 11.01 | 156 |
| OMN | Oman | 456.748 | 457.950 | 457.185 | 457.185 | 3.950 | 4.399 | 4.392 | 4.392 | 2863.40 | 2842.69 | 2504.62 | 2504.48 | 28.11 | 27.99 | 24.75 | 24.75 | 227 |
| PRT | Portugal | 517.757 | 516.611 | 516.343 | 516.269 | 3.061 | 3.205 | 3.934 | 3.934 | 1097.38 | 1040.89 | 1194.81 | 1192.00 | 22.43 | 21.43 | 24.02 | 23.98 | 155 |
| QAT | Qatar | 477.698 | 472.839 | 478.460 | 478.474 | 5.384 | 6.661 | 5.393 | 5.390 | 2198.60 | 2280.94 | 2212.84 | 2207.51 | 37.01 | 38.85 | 37.06 | 37.02 | 151 |
| ROM | Romania | 468.637 | 476.607 | 455.056 | 455.055 | 4.362 | 4.351 | 5.313 | 5.313 | 3051.46 | 2705.29 | 3047.75 | 3046.90 | 38.50 | 35.93 | 38.10 | 38.10 | 197 |
| RUS | Russia | 541.492 | 547.619 | 537.044 | 537.044 | 3.366 | 4.096 | 4.454 | 4.454 | 2272.34 | 2306.32 | 2099.82 | 2099.82 | 35.00 | 36.36 | 30.99 | 30.99 | 203 |
| SAU | Saudi Arabia | 444.952 | 446.555 | 427.183 | 427.181 | 3.980 | 4.150 | 4.868 | 4.868 | 2686.17 | 2435.58 | 2285.76 | 2286.12 | 33.48 | 30.84 | 28.75 | 28.76 | 208 |
| SGP | Singapore | 601.119 | 606.727 | 601.119 | 599.898 | 4.964 | 5.021 | 4.964 | 5.023 | 4147.61 | 4114.03 | 4147.61 | 4226.61 | 45.60 | 45.36 | 45.60 | 46.10 | 152 |
| ZAF | South Africa | 398.010 | 411.841 | 353.598 | 353.598 | 3.972 | 4.427 | 6.446 | 6.446 | 23478.83 | 24255.84 | 18450.98 | 18450.98 | 56.66 | 56.93 | 53.42 | 53.42 | 518 |
| SWE | Sweden | 521.261 | 520.481 | 518.783 | 518.685 | 3.749 | 3.971 | 5.768 | 5.765 | 980.98 | 957.65 | 1175.38 | 1173.00 | 18.66 | 18.06 | 23.12 | 23.08 | 149 |
| TUR | Türkiye | 510.218 | 518.682 | 499.498 | 499.498 | 4.440 | 4.612 | 7.634 | 7.634 | 2135.14 | 1755.79 | 2592.50 | 2592.50 | 33.44 | 29.12 | 38.06 | 38.06 | 180 |
| ARE | UAE | 464.708 | 470.069 | 466.090 | 466.003 | 3.932 | 5.066 | 3.832 | 3.836 | 33220.96 | 37948.12 | 31535.70 | 31627.73 | 58.85 | 61.93 | 57.47 | 57.54 | 622 |
| USA | United States | 523.727 | 527.188 | 490.639 | 490.786 | 3.883 | 5.052 | 22.871 | 22.923 | 6037.26 | 6654.61 | 8358.99 | 8345.63 | 40.65 | 43.59 | 58.22 | 58.29 | 272 |

- The degrees of freedom have different values in the models fitted with the data of Morocco (MAR), South Africa (ZAF), Lebanon (LBN) and Norway (NOR).

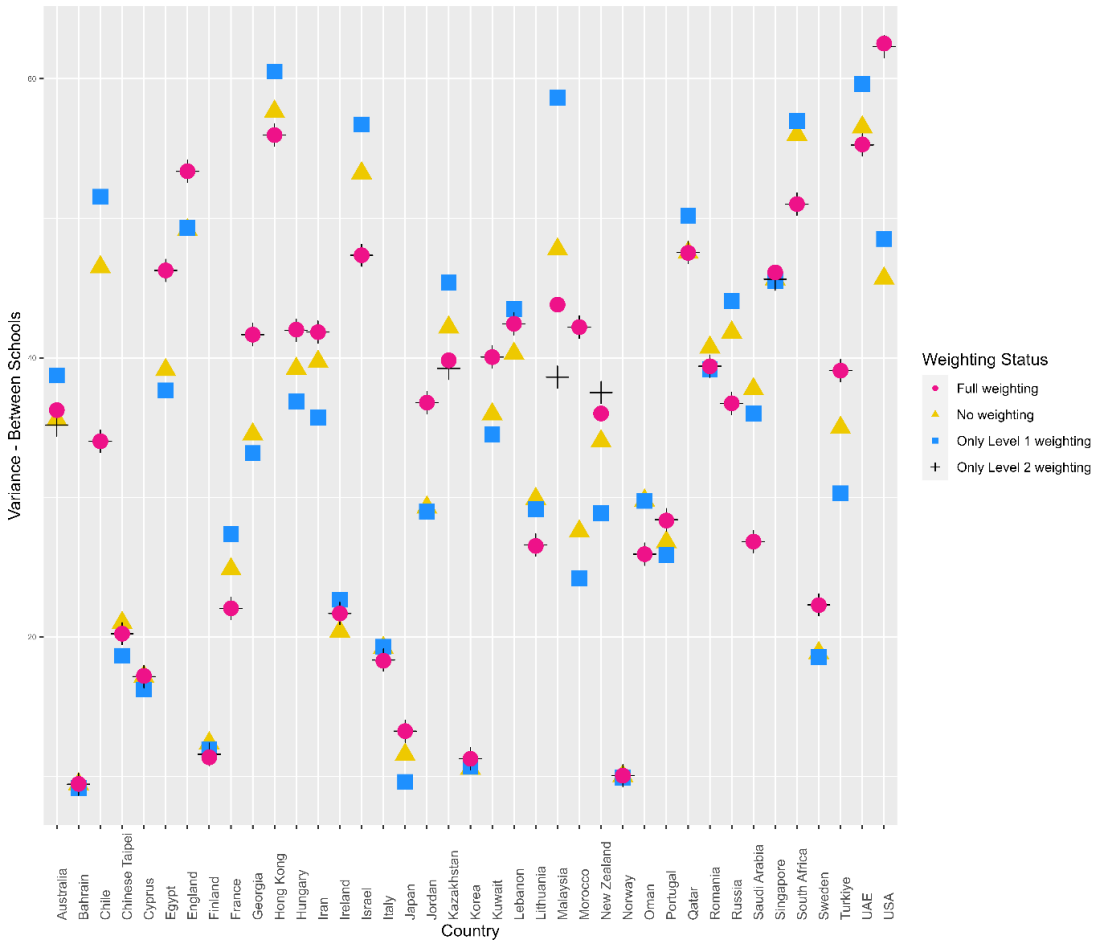


**Figure 1**. The ratio of variance in mathematics achievement attributed to schools

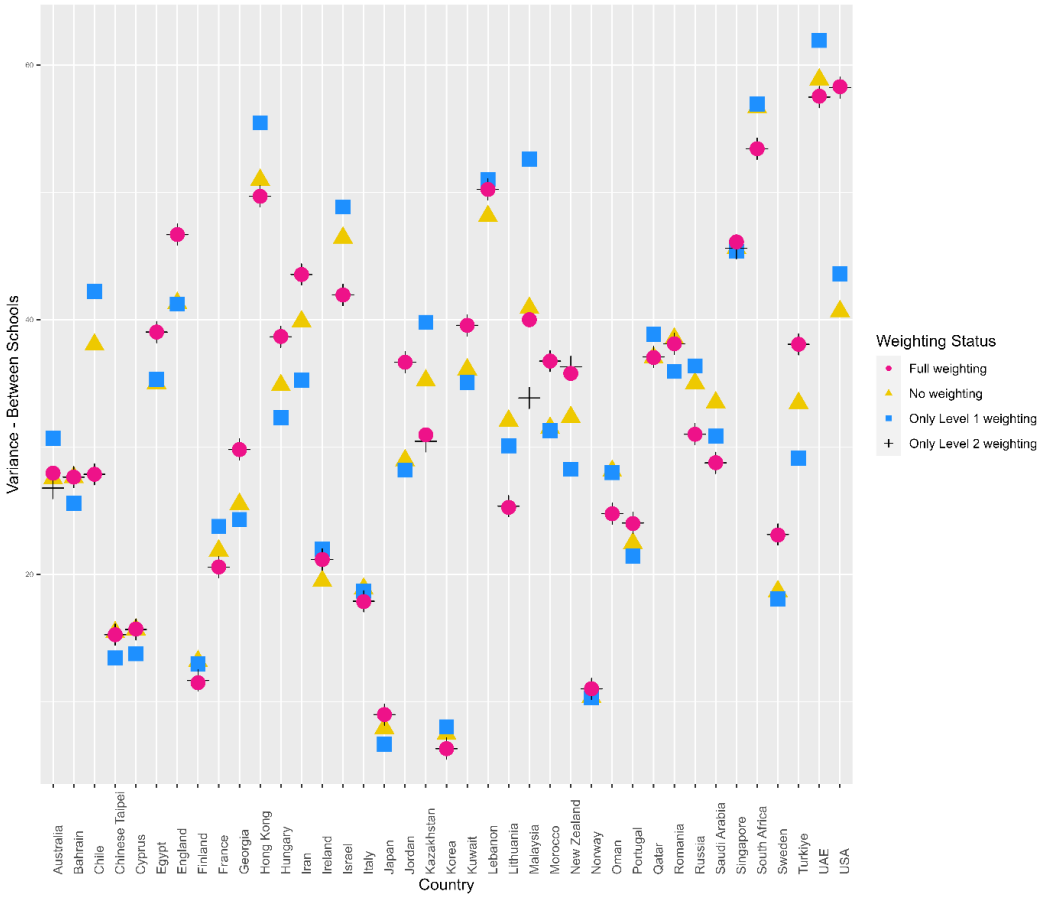


**Figure 2**. The ratio of variance in science achievement attributed to schools

As illustrated in Figure 1, it is evident that the instances where the ratio of variance attributed to schools is calculated to be highest exhibit a relatively balanced distribution. The highest percentages were obtained in 6 countries with no weighting, 14 countries with only level 1 weighting, 10 countries with only level 2 weighting, and 12 countries with both levels of weighting. It is noteworthy that the highest percentages among the six countries are exactly equal to each other, regardless of their weighting status.

Figure 2 depicts the variation in the percentage of the variance in science achievement scores attributed to school differences, depending on the treatment of the weighting variable in the models.

As observed from Figure 2, the instances where the ratio of variance attributed to schools is calculated to be highest exhibit a relatively balanced distribution. The highest rates were obtained in seven countries with no weighting, 14 countries with only level 1 weighting, nine countries with only level 2 weighting, and 12 countries with both levels of weighting. It is noteworthy that the highest rates among the six countries are exactly equal to each other, irrespective of their weighting status.

**Findings Related to the 2nd Research Question**

As shown in Table 1 and Table 2, the standard errors of estimations in models where mathematics achievement is the dependent variable range from 2.38% (Norway) to 9.19% (Egypt). Similarly, in models where science achievement is the dependent variable, the standard errors of estimations range between 2.01% (Japan) and 9.29% (Egypt). It has been observed that the standard errors vary based on the treatment of the weighting variable in the data of different countries. Figure 3 displays the standard errors obtained depending on how the weighting variable is handled in models where mathematics achievement is the dependent variable.

Figure 3 illustrates that the lowest standard errors were obtained in 26 countries with no weighting, in five countries with only level 1 weighting, in seven countries with only level 2 weighting, and in three countries with both levels of weighting. Notably, in five countries (Bahrain, Chile, Cyprus, France and Singapore), the minimum standard errors remain the same across different weighting conditions.

Figure 4 displays the standard errors obtained depending on how the weighting variable is handled in models where science achievement is the dependent variable.

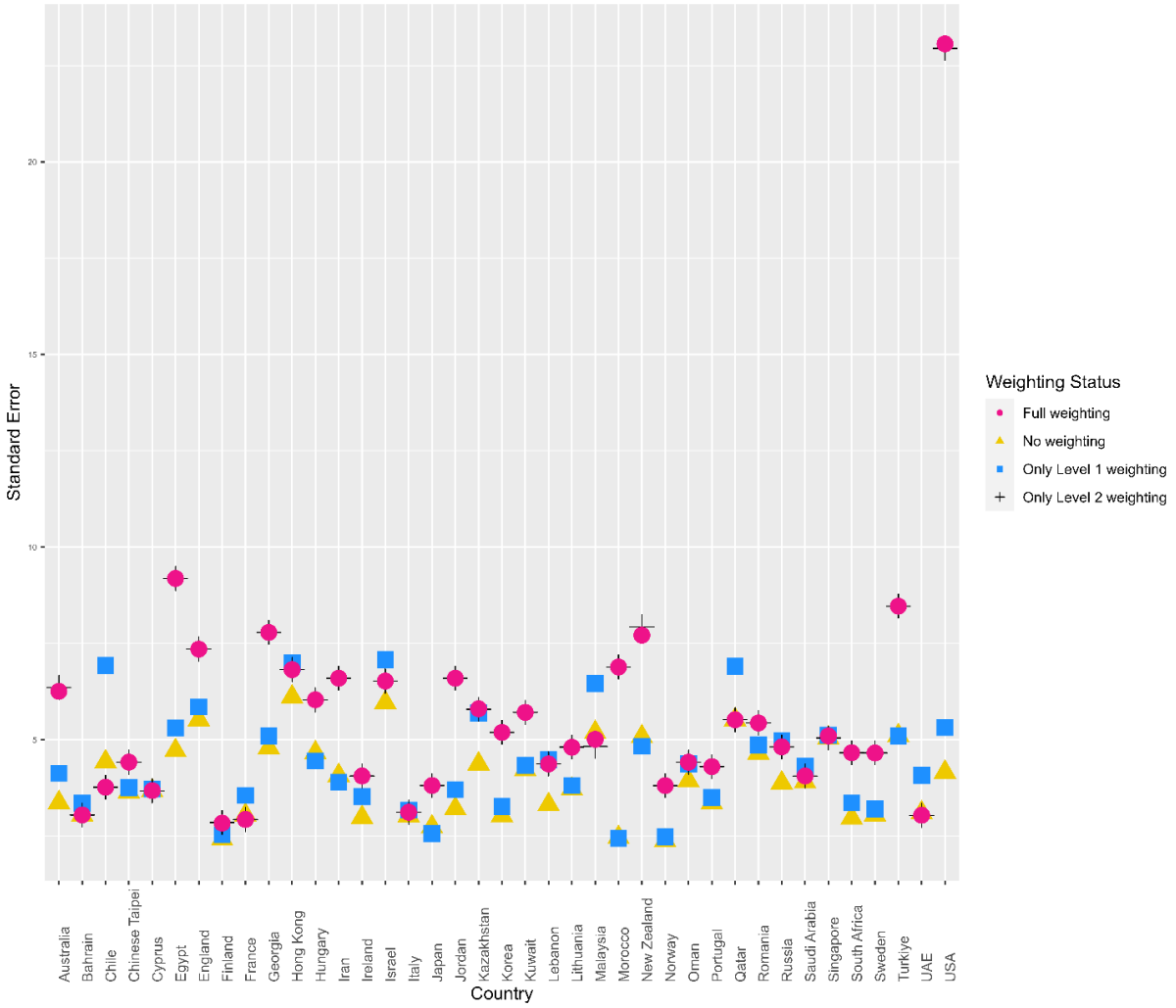


**Figure 3**. Standard errors in mathematics achievement models based on weighting variable handling

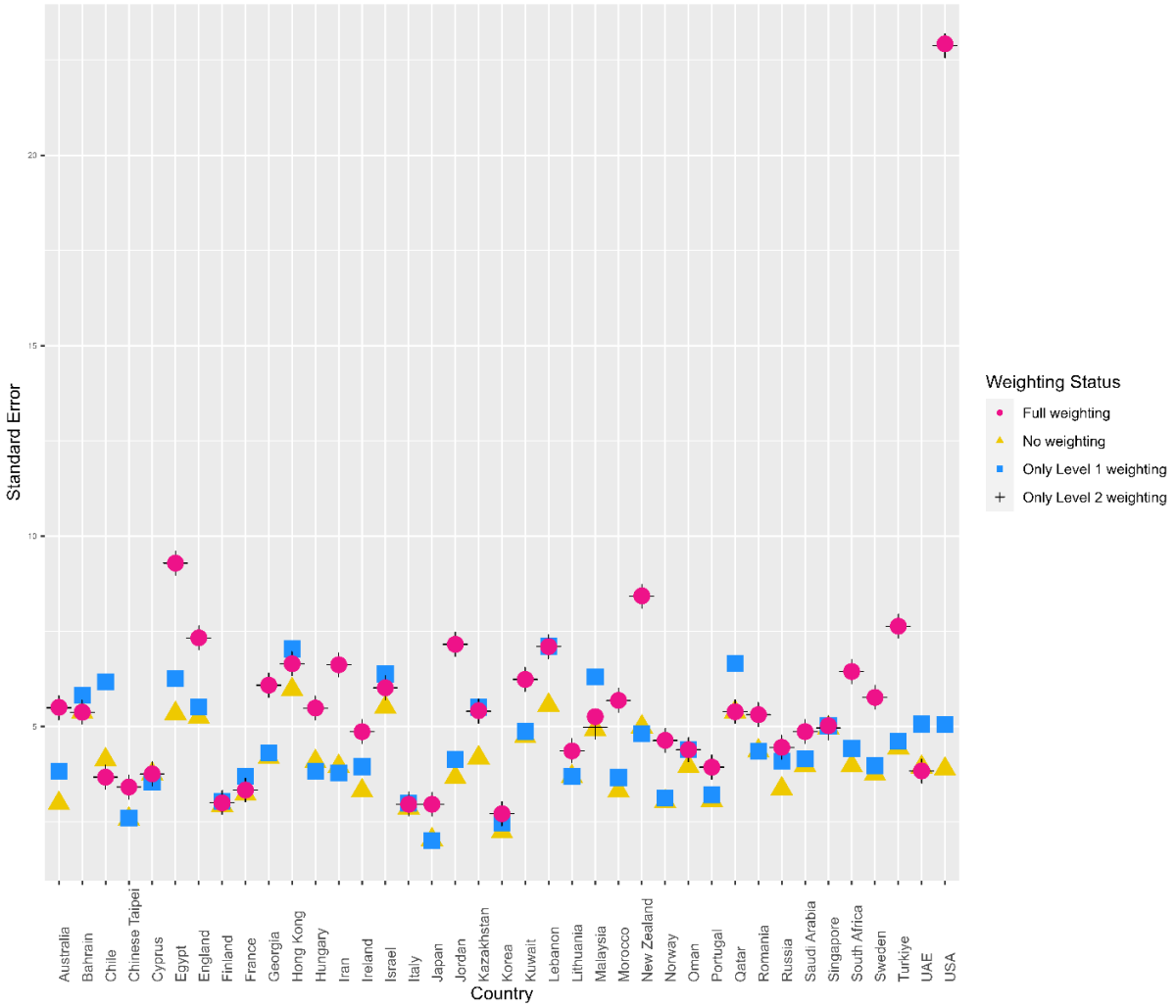


**Figure 4**. Standard errors in science achievement models based on weighting variable handling

As depicted in Figure 4, the lowest standard errors are observed in 28 countries without weighting, six countries with only level 1 weighting, four countries with only level 2 weighting, and one country with both levels of weighting when modeling science achievement as the dependent variable. Notably, in three countries (Bahrain, Chile and Singapore), the minimum standard errors remain the same across different weighting conditions.

## Results, Discussion and Recommendations

This study examines the impact of weighting in different ways on estimations (particularly the ratio of variance in achievement scores attributed to differences between schools and standard errors) when analyzing the data obtained from large-scale tests where plausible values are considered as the dependent variable.

In cases where both mathematics and science achievement were considered as dependent variables, it was observed that coefficients were estimated to be higher when only 1st level weighting was applied. The highest model reliability was achieved when using either only level 2 weighting or both levels of weighting. The lowest standard errors were observed in the unweighted condition, with 26 countries in mathematics and 28 countries in science. This increase in standard errors when weighting is applied aligns with the findings reported in the studies of Carle (2009) and Tat et al. (2019), indicating a parallelism in the results. As Meinck and Vandenplas (2012) stated, the increase in standard errors due to the use of weighting variables in analyses should not be interpreted as making weighting inappropriate. This increase could, however, depend on the distribution of the weights and perhaps even their correlation with the dependent variable. The lowest $\chi^2$ values were obtained when only level 1 weighting was used, with 16 countries in mathematics and 15 countries in science.

In this study, it was observed that in models where mathematics achievement is considered as the dependent variable, the differences between the rates calculated based on different approaches to weighting can reach up to 20% for a country (Malaysia). Furthermore, it was noted that these rates can vary significantly in models constructed using data from Morocco, Chile, and the USA. Similarly, in models where science achievement is the dependent variable, it was observed that the differences between the rates calculated according to various weighting approaches can reach up to 18.75% for a country (Malaysia). Additionally, significant variations in these rates were observed in models based on data from the USA and Chile. Despite the use of weighting variables in the same manner, the differing results observed in the data from these countries may indicate that weighting should be applied differently in these contexts. Additionally, this variation could be due to the distribution of the weighting variable within the data of the respective countries or the correlation between the weighting variable and the dependent variables (Meinck & Vandenplas, 2012).

The most crucial conclusion that can be drawn from the research findings is that the use, or lack thereof, and the manner in which the weighting variable is employed when conducting HLM analysis on data from large-scale tests can greatly impact statistical inferences. Although the importance of handling the weighting variable is emphasized in studies involving hierarchical linear modeling on multilevel data (Rutkowski et al., 2010), empirical results demonstrate that treating weighting variables differently affects many statistics (Meinck & Vandenplas, 2012; Tat et al., 2019). For instance, Laukaityte and Wiberg (2017) explicitly state that not applying weighting or applying weighting only at the first level (student) can lead to misleading results. In a study conducted by Mang et al. (2021) using large-scale test data, they examined which weighting method yielded better results. The researchers reported that weighting at only the second level (school) was appropriate. In this study, it was observed that adding weighting variables increased the standard errors of estimates while positively contributing to model reliability coefficients. Therefore, based on the findings of this study and the knowledge in the literature, it is recommended that researchers conducting analyses with HLM consider using weighting variables either at the second level only or at both levels. However, it should be kept in mind that in studies attempting to explain the variance in the dependent variable by variables at the second level, adding weighting variables may inflate the ratio of the variance in the dependent variable attributed to differences between schools. Hence, it is crucial for researchers to transparently disclose the steps taken in the analysis process.

In this study, only the data from the 2019 cycle of TIMSS were utilized as the large-scale test data. Examining situations in other large-scale tests and in different cycles of TIMSS can provide an opportunity to strengthen the interpretations in the existing literature.

## Author Contributions

MG wrote the first draft of the manuscript, along with undertaking the analysis. MG, SB and ND refined the analytical methods for the modeling. MG, SB and ND refined the discussion. All authors contributed to the editing of the paper and the development of the final manuscript. All authors have read and approved the final manuscript.

## Ethical Approval

The authors declare that their work is not subject to ethics committee approval and that the rules set by the Committee on Publication Ethics (COPE) were followed throughout the study.

## Conflict of Interest

The authors declare that there is no conflict of interest with any institution or person within the scope of the study.

## References


Ababneh, E., Al-Tweissi, A., & Abulibdeh, K. (2016). TIMSS and PISA impact – the case of Jordan. *Research Papers in Education*, *31*(5), 542-555, http://dx.doi.org/10.1080/02671522.2016.1225350

Aksu, G., Güzeller, C. O., & Eser, M. T. (2017). Analysis of maths literacy performances of students with hierarchical linear modeling (HLM): The case of PISA 2012 Turkey. *Education and Science*, *42*(191), 247-266. http://dx.doi.org/10.15390/EB.2017.6956

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. https://doi.org/10.21031/epod.602765

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

Barber, M., Chijioke, C., & Mourshed, M. (2010). *How the World's most improved school systems keep getting better*. McKinsey and Company.

Bilican, S., & Yıldırım, Ö. (2013). The effects of approaches to learning on student's reflective and evaluative reading performance in Turkey: The results from PISA 2009. *Procedia - Social and Behavioral Sciences*, *116*, 2437-2442. https://doi.org/10.1016/j.sbspro.2014.01.588

Boulifa, K., & Kaaouachi, A. (2022). The relationship between the school resources index, gender, age and mathematics achievement in TIMSS 2019 survey: Multilevel analysis. *Procedia Computer Science*, *201*, 738-745. https://doi.org/10.1016/j.procs.2022.03.100

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

Chu, M. W., Babenko, O., Cui, Y., & Leighton, J. P. (2014). Using HLM to explore effects of perceptions of learning environments and assessments on students' test performance. *International Journal of Testing*, *14*(2), 95-121. https://doi.org/10.1080/15305058.2013.841702

Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th Ed.). McGraw-Hill.

Gómez, R. L., & Suárez, A. M. (2020). Do inquiry-based teaching and school climate influence science achievement and critical thinking? Evidence from PISA 2015. *International Journal of STEM Education*, *7*(43), 1-11. https://doi.org/10.1186/s40594-020-00240-5

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Routledge.

IBM Corp. (2013). *IBM SPSS Statistics for Windows* (Version 22.0). IBM Corp.

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. https://doi.org/10.1080/03610926.2016.1267764

Liang, X. (2010). Assessment use, self-efficacy and mathematics achievement: comparative analysis of PISA 2003 data of Finland, Canada and the USA. *Evaluation & Research in Education*, *23*(3), 213-229. https://doi.org/10.1080/09500790.2010.490875

Liou, P. Y., & Hung, Y. C. (2015). Statistical techniques utilized in analyzing PISA and TIMSS data in science education from 1996 to 2013: A methodological review. *International Journal of Science and Mathematics Education*, *13*, 1449-1468. https://doi.org/10.1007/s10763-014-9558-5

Mang, J., Küchenhoff, H., Meinck, S., & Prenzel, M. (2021). Sampling weights in multilevel modelling: An investigation using PISA sampling structures. *Large-scale*

*Assess Educ.*, *9*(6), 1-39. https://doi.org/10.1186/s40536-021-00099-0

Manjunath, M. (2021). Using large-scale assessments to inform education policy and governance, Assessment resources, Azim Premji University. Retrieved from https://cdn.azimpremjiuniversity.edu.in/apuc3/media/resources/Assessments-and-Education-Policy.f1640171199.pdf

Meinck, S., & Vandenplas, C. (2012). *Sample size requirements in HLM: An empirical study*. IERI monograph series issues and methodologies in large-scale assessments. IER Institute, special issue 1, Educational Testing Service and International Association for the Evaluation of Educational Achievement.

Organisation for Economic Co-operation and Development [OECD]. (2017). *PISA 2015 technical report.* Paris: OECD Publishing.

Özdemir, C. (2016). A methodological review of research using OECD PISA Turkey data. *Education Science Society Journal*, *14*(56), 10-27.

Pacheco Diaz, N., & Rocconi, L. M. (2021). Examining science achievement in Chile: A multilevel approach using PISA 2015 data. *Journal of Research in STEM Education*, *7*(2), 93-116. https://doi.org/10.51355/jstem.2021.100

Pong, S. (2009). Grade level and achievement of immigrants' children: academic redshirting in Hong Kong. *Educational Research and Evaluation: An International Journal on Theory and Practice*, *15*(4), 405-425. http://dx.doi.org/10.1080/13803610903087078

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

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2007). *HLM for Windows* (Version 6.04). Scientific Software International.

Reinikainen, P. (2007). *Sequential explanatory study of factors connected with science achievement in six countries: Finland, England, Hungary, Japan, Latvia and Russia: Study based on TIMSS 1999*. Institute for Educational Research, University of Jyväskylä.

Ross, S. P. (2008). *Motivation correlates of academic achievement: Exploring how motivation influences academic achievement in the PISA 2003 dataset* [Doctoral Dissertation, University of Victoria]. UVicSpace. https://dspace.library.uvic.ca/handle/1828/3209

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* (pp. 117-154). 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. https://doi.org/10.3102/0013189X10363170

Saal, P. E., van Ryneveld, L., & Graham, M. A. (2019). The relationship between using information and communication technology in education and the mathematics achievement of students. *International Journal of Instruction*, *12*(3), 405-424. https://doi.org/10.29333/iji.2019.12325a

Sabudin, S., Mansor, A. N., Meerah, S. M., & Muhammad, A. (2018). Teacher-level factors that influence students' science and technology culture: HLM analysis. *International Journal of Academic Research in Business and Social Sciences*, *8*(5), 977-985. http://dx.doi.org/10.6007/IJARBSS/v8-i5/4243

Sahlberg, P., & Hargreaves, A. (2015). "The Tower of PISA is Badly Leaning. An Argument for Why It Should Be Saved", The Washington Post. https://www.washingtonpost.com/news/answer-sheet/wp/2015/03/24/the-tower-of-pisa-is-badly-leaning-an-argument-for-why-it-should-be-saved/ (8 May 2023).

Schleicher A. (2015). "Attacks on PISA are Entirely Unjustified", tes Magazine. https://www.tes.com/magazine/archive/attacks-pisa-are-entirely-unjustified-0 (8 May 2023).

Schmidt, W. H., McKnight, C. C., & Raizen, S. A. (1997). *A splintered vision: An investigation of US science and mathematics education*. U. S. National Research Center.

Sun, L., Bradley, K. D., & Akers, K. (2012). A multilevel modelling approach to investigating factors impacting science achievement for secondary school students: PISA Hong Kong sample. *International Journal of Science Education*, *34*(14), 2107-2125. https://doi.org/10.1080/09500693.2012.708063

Takayama, K. (2015). "Has PISA Helped or Hindered?" https://headfoundation.org/2015/04/15/has-pisa-helped-or-hindered/ (8 May 2023).

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. https://doi.org/10.21031/epod.486999

Thien, L. M., Darmawan, I. G. N., & Ong, M. Y. (2015). Affective characteristics and mathematics performance in Indonesia, Malaysia, and Thailand: What can PISA 2012 data tell us?. *Large-Scale Assessments in Education*, *3*(3), 1-16. https://doi.org/10.1186/s40536-015-0013-z

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). *Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region*. Melbourne: ACER and Bangkok: UNESCO.

Valente, M. O., Fonseca, J., & Conboy, J. (2011). Inquiry science teaching in Portugal and some other countries as measured by PISA 2006. *Procedia - Social and Behavioral Sciences*, *12*, 255-262. https://doi.org/10.1016/j.sbspro.2011.02.034

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. https://doi.org/10.20982/tqmp.08.1.p052

Woo, H., & Henfield, M. S. (2016). Student and teacher factors' impact on fourth grade students' mathematics achievement: An HLM analysis of TIMSS 2007. *Journal of Mathematics Education*, *9*(1), 69-87.

**APPENDIX A**. Reliability Values of Models Constructed Based on the Use of Weighting Variable

| Country | Models where Mathematics Achievement is the Dependent Variable | | | | Models where Science Achievement is the Dependent Variable | | | |
|---|---|---|---|---|---|---|---|---|
| | No | Level 1 | Level 2 | Full | No | Level 1 | Level 2 | Full |
| Australia | 0.93 | 0.94 | 0.94 | 0.94 | 0.91 | 0.92 | 0.91 | 0.92 |
| Bahrain | 0.83 | 0.81 | 0.83 | 0.83 | 0.94 | 0.93 | 0.94 | 0.94 |
| Chile | 0.95 | 0.95 | 0.93 | 0.93 | 0.93 | 0.93 | 0.91 | 0.91 |
| Chinese Taipei | 0.86 | 0.84 | 0.87 | 0.87 | 0.81 | 0.78 | 0.82 | 0.82 |
| Cyprus | 0.86 | 0.85 | 0.86 | 0.86 | 0.85 | 0.82 | 0.85 | 0.85 |
| Egypt | 0.96 | 0.96 | 0.97 | 0.97 | 0.96 | 0.95 | 0.96 | 0.96 |
| England | 0.96 | 0.96 | 0.97 | 0.97 | 0.94 | 0.94 | 0.96 | 0.96 |
| Finland | 0.81 | 0.80 | 0.82 | 0.82 | 0.82 | 0.81 | 0.82 | 0.82 |
| France | 0.89 | 0.90 | 0.88 | 0.88 | 0.87 | 0.88 | 0.87 | 0.87 |
| Georgia | 0.89 | 0.88 | 0.94 | 0.94 | 0.84 | 0.83 | 0.91 | 0.91 |
| Hong Kong | 0.97 | 0.97 | 0.97 | 0.97 | 0.96 | 0.97 | 0.96 | 0.96 |
| Hungary | 0.94 | 0.93 | 0.95 | 0.95 | 0.92 | 0.91 | 0.95 | 0.95 |
| Iran | 0.94 | 0.93 | 0.96 | 0.96 | 0.94 | 0.92 | 0.96 | 0.96 |
| Ireland | 0.86 | 0.88 | 0.88 | 0.88 | 0.85 | 0.87 | 0.88 | 0.88 |
| Israel | 0.96 | 0.97 | 0.96 | 0.96 | 0.95 | 0.95 | 0.95 | 0.95 |
| Italy | 0.83 | 0.83 | 0.83 | 0.83 | 0.82 | 0.82 | 0.83 | 0.83 |
| Japan | 0.80 | 0.76 | 0.84 | 0.84 | 0.72 | 0.68 | 0.77 | 0.77 |
| Jordan | 0.92 | 0.91 | 0.95 | 0.95 | 0.92 | 0.91 | 0.95 | 0.95 |
| Kazakhstan | 0.93 | 0.94 | 0.95 | 0.95 | 0.91 | 0.92 | 0.93 | 0.93 |
| Korea | 0.73 | 0.72 | 0.76 | 0.76 | 0.64 | 0.66 | 0.63 | 0.63 |
| Kuwait | 0.94 | 0.93 | 0.95 | 0.95 | 0.94 | 0.93 | 0.95 | 0.95 |
| Lebanon | 0.93 | 0.94 | 0.94 | 0.94 | 0.95 | 0.95 | 0.96 | 0.96 |
| Lithuania | 0.86 | 0.84 | 0.86 | 0.86 | 0.87 | 0.85 | 0.85 | 0.85 |
| Malaysia | 0.97 | 0.98 | 0.96 | 0.96 | 0.96 | 0.97 | 0.95 | 0.96 |
| Morocco | 0.92 | 0.91 | 0.96 | 0.96 | 0.94 | 0.94 | 0.96 | 0.96 |
| New Zealand | 0.95 | 0.94 | 0.96 | 0.96 | 0.95 | 0.94 | 0.96 | 0.96 |
| Norway | 0.73 | 0.73 | 0.78 | 0.78 | 0.73 | 0.74 | 0.80 | 0.80 |
| Oman | 0.91 | 0.91 | 0.91 | 0.91 | 0.91 | 0.90 | 0.91 | 0.91 |
| Portugal | 0.87 | 0.87 | 0.89 | 0.89 | 0.84 | 0.84 | 0.87 | 0.87 |
| Qatar | 0.95 | 0.95 | 0.95 | 0.95 | 0.92 | 0.93 | 0.92 | 0.92 |
| Romania | 0.93 | 0.92 | 0.94 | 0.94 | 0.92 | 0.91 | 0.94 | 0.94 |
| Russia | 0.92 | 0.92 | 0.93 | 0.93 | 0.90 | 0.90 | 0.91 | 0.91 |
| Saudi Arabia | 0.93 | 0.93 | 0.91 | 0.91 | 0.92 | 0.91 | 0.92 | 0.92 |
| Singapore | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 | 0.96 |
| South Africa | 0.98 | 0.98 | 0.97 | 0.97 | 0.98 | 0.98 | 0.98 | 0.98 |
| Sweden | 0.84 | 0.84 | 0.88 | 0.88 | 0.84 | 0.83 | 0.89 | 0.89 |
| Türkiye | 0.92 | 0.89 | 0.94 | 0.94 | 0.91 | 0.88 | 0.94 | 0.94 |
| UAE | 0.97 | 0.98 | 0.97 | 0.97 | 0.98 | 0.98 | 0.98 | 0.98 |
| United States | 0.95 | 0.96 | 0.99 | 0.99 | 0.94 | 0.95 | 0.98 | 0.98 |

## Genişletilmiş Türkçe Özet

### Giriş

Uluslararası Matematik ve Fen Eğilimleri Araştırması (TIMSS) ve Uluslararası Öğrenci Değerlendirme Programı (PISA) gibi geniş ölçekli testler, uluslararası düzeyde karşılaştırmayı ve ulusal düzeyde eğilimleri incelemeyi amaçlamaktadır. Bu uygulamaların küresel eğitim politikaları ve eğitim alanında yapılacak reformlara olan etkisi her geçen gün artmaktadır (Ababneh ve diğerleri, 2016; Barber ve diğerleri, 2010). Eğitim araştırmacılarının geniş ölçekli testlerin verilerinin analizleriyle ortaya koyduğu bulgular eğitim sistemleri açısından karar alma mekanizmalarına yakın kişilerce incelenip aksiyonlar alınabileceği için araştırmacılar analizlerinde kullanacağı modeli seçme konusunda dikkatli olmalıdır.

Türkiye'de ve dünyanın diğer ülkelerinde, özellikle geniş ölçekli testlerin verilerinin analizinde sıklıkla kullanılan istatistiksel modellemelerden biri hiyerarşik doğrusal modellemedir (HLM). HLM özellikle geniş ölçekli test uygulamalarındaki gibi örneklem seçiminin tabakalı olduğu durumlarda tercih edilmektedir. Alan yazında geniş ölçekli testlerin verileri ile yürütülen, HLM'nin kullanıldığı çok sayıda araştırma bulunmaktadır. HLM'nin işe koşulduğu araştırmalarda örneklem ağırlık değişkenlerinin kullanıldığı, kullanılmadığı ve bu konu ile ilgili yeterli detayın yazarlar tarafından verilmediği durumlar mevcuttur. Özdemir'in (2016) PISA Türkiye verileri ile yapılan analizleri yöntembilimsel açıdan incelediği çalışmasında, incelenen 97 araştırmanın büyük kısmında olası değerlerin ve örneklem ağırlıklarının ele alınıp alınmadığının belirtilmediği raporlanmıştır. Ele alınan sadece 5 makalede örneklem ağırlıklarının doğru kullanımı rapor edilirken, olası değerler de sadece 10 makalede uygun şekilde kullanılmıştır. Liou ve Hung'un (2015) yürüttüğü benzer bir tarama çalışmasında PISA ve TIMSS verisi kullanılarak yapılan çok sayıda çalışmada örneklem ağırlıklarının kullanılmadığı veya kullanılıp kullanılmadığının belirtilmediği gözlemlenmiştir. Bu durumun da gerçekleştirilen analizlerin güvenilirliğini şüpheli hale getirdiğinin altı çizilmiştir. Dahası, geniş ölçekli testlerin verileri ile yürütülen analizlerin çıktılarının eğitim politikalarına etkisini eleştiren yazılar da kaleme alınmıştır (Sahlberg ve Hargreaves, 2015; Schleicher, 2015; Takayama, 2015). Bu eleştiriler, istatistiksel analizlerde özenli olunması gerektiğinin açık bir göstergesidir.

Örneklemin sahip olduğu özelliklerin, örneklemin alındığı evrenden farklı olma durumunu çözümlemek için kullanılabilecek en bilindik yöntemlerden biri örneklem ağırlıklarının kullanılmasıdır (Rust, 2013). Ağırlıklandırma, özellikle HLM ile yürütülen çalışmalarda göz önünde bulundurulması gereken önemli bir kavramdır. TIMSS uygulamalarında toplanan tüm verilerin yanı sıra örneklem ağırlıkları değişkenleri de veri seti içinde paylaşılmaktadır. Bu ağırlık değişkenlerinin analizlerde nasıl kullanılacağı araştırmacılara bırakılmaktadır. Çok düzeyli analizlerde örneklem ağırlıkları ilgili düzeylere göre ayrıştırılarak kullanılması (Rutkowski ve diğerleri, 2010) ve örneklem ağırlıklarının kullanılmadığı durumlarda elde edilen sonuçların daha az güvenilir olacağının farkında olunması gerekmesine (OECD, 2017) karşın örneklem ağırlıklarının nasıl kullanılacağı ve analizlere etkilerinin ne olacağı konusunda alan yazında tartışmalar bulunmaktadır.

Bu araştırmada, TIMSS 2019 uygulamasına katılan ülkelerdeki 8. sınıf öğrencilerinin matematik ve fen başarı puanlarındaki varyansın okullar arasındaki farktan kaynaklanan kısmının ve kestirimlerdeki standart hataların ağırlıklandırma değişkenlerinin kullanılıp kullanılmama durumlarına göre ne şekilde değiştiği incelenmiştir. Bu doğrultuda dikkate alınan iki araştırma sorusu aşağıdaki şekilde belirlenmiştir.

1. TIMSS 2019 sonuçlarına göre uygulamaya katılan ülkelerdeki okullar arası fark, (a) ağırlıklandırma değişkenlerinin kullanılmadığı, (b) yalnızca birinci düzey ağırlıklandırma değişkeninin kullanıldığı, (c) yalnızca ikinci düzey ağırlıklandırma değişkeninin kullanıldığı ve (d) her iki düzeyde de ağırlıklandırma değişkenlerinin kullanıldığı durumlarda 8. sınıf öğrencilerinin matematik ve fen başarı puanlarındaki toplam varyansın ne kadarını açıklamaktadır?
2. TIMSS 2019 sonuçlarına göre uygulamaya katılan ülkelerdeki öğrencilerin ortalama matematik ve fen başarı puanları kestiriminin standart hataları, (a) ağırlıklandırma değişkenlerinin kullanılmadığı, (b) yalnızca birinci düzey ağırlıklandırma değişkeninin kullanıldığı, (c) yalnızca ikinci düzey ağırlıklandırma değişkeninin kullanıldığı ve (d) her iki düzeyde de ağırlıklandırma değişkenlerinin kullanıldığı durumlarda nasıl değişmektedir?

### Yöntem

Bu araştırmada, geniş ölçekli test uygulamalarından elde edilen hiyerarşik yapıdaki verilerde, ağırlıklandırma değişkenlerinin farklı şekillerde kullanılması ile açıklanan varyans arasındaki ilişki belirlenmeye çalışılmıştır. Bu nedenle araştırma ilişkisel araştırma türündedir (Fraenkel ve Wallen, 2006).

### Evren ve Örneklem

Bu araştırmada TIMSS 2019 uygulamasına katılan 39 ülkeden toplam 7636 okuldaki 227345 8. sınıf öğrencisi araştırma örneklemini oluşturmaktadır. Katılımcı ülkelerdeki uygulamaya katılan okul sayıları 98 ile 623; öğrenci sayıları 3265 ile 22334 arasında değişmektedir.

### Veri Toplama

Bu çalışmanın analizlerinde kullanılan veriler TIMSS 2019 öğrenci ve okul anketleri ile elde edilen verilerdir. Bu verilere internet üzerinden https://timss2019.org/international-database/ adresinden erişilebilmektedir.

### Veri Analizi

Arıkan ve arkadaşlarının (2020) önerileri dikkate alınarak verilerin analizinden önce araştırma için seçilen veri setlerine ağırlık değişkenleri (hem öğrenci, hem okul düzeyinde) eklenmiştir. Birinci düzey (öğrenci) örneklem ağırlığı TIMSS 2019 veri setlerinde yer alan WGTADJ2 * WGTFAC2 * WGTADJ3 * WGTFAC3 (CLASS WEIGHT ADJUSTMENT * CLASS WEIGHT FACTOR * STUDENT WEIGHT ADJUSTMENT * STUDENT WEIGHT FACTOR) değişkenlerinin çarpımı; ikinci düzey (okul) örneklem ağırlığı WGTADJ1 * WGTFAC1 (SCHOOL WEIGHT ADJUSTMENT * SCHOOL WEIGHT FACTOR) değişkenlerinin çarpımı ile elde edilerek veri setlerine eklenmiştir. 1 ve 2. düzeyde elde edilen ağırlık değişkenlerinin çarpımı tam olarak TIMSS tarafından paylaşılan TOTWGT (TOTAL STUDENT WEIGHT) değerini vermektedir.

HLM'de sıklıkla kullanılan sıralı dört alt model ile analizler gerçekleştirilmektedir. Ancak bu çalışmada, ağırlıklandırma değişkeninin etkisine ve kestirimlerin standart hatalarına odaklanıldığı için yalnızca ilk model olan tesadüfi etkili tek yönlü ANOVA modelleri kurulmuştur. Bu model aracılığı ile yapılacak analizlerden önce bazı varsayımların karşılanması gerekmektedir. Varsayımların incelenmesi noktasında, Raudenbush ve Bryk (2002) tarafından ifade edilen; artık değerlerin (residuals) 0 ortalama ile normal dağılımlı ve homojen olması, düzeylere yönelik hataların hem kendi içlerinde hem de birbirleri arasında ilişkisiz olması temel varsayımları dikkate alınmıştır. İlk varsayımın incelenmesi, kurulan modellere yönelik artık değerlerin kontrolü ile gerçekleştirilmiştir. Artıkların dağılımı grafiksel yöntemlerle (histogram, Q-Q grafiği) incelenmiştir ve buna göre grafikler, normal dağılımdan ciddi bir sapma olmadığını göstermektedir. Ayrıca kurulan modellerden elde edilen hataların çarpıklık ve basıklık değerleri incelenmiş, çarpıklık değerlerinin -,42 ile ,57 arasında, basıklık değerlerinin -1 ile 1,98 arasında dağıldığı gözlenmiştir. Buna göre ciddi bir normallik ihlali bulunmadığı değerlendirilmiştir. Varyansların homojenliğinin testi, $\chi^2$ istatistiğine dayalı olarak gerçekleştirilmiştir. Ancak sonuçlara göre varyansların homojen olmadığı görülmüştür. Buna karşın, $\chi^2$ istatistiğinin tip I hataya duyarlı ve büyük örneklemlerde istatistiksel olarak anlamlı fark çıkarmaya eğilimli olduğu bilinmektedir. Bu nedenle, artıkların normal dağılım gösteriyor oluşu yeterli görülmüş ve analizlere devam edilmiştir. Son olarak analizler için yeterli örneklem büyüklüğünün her bir düzeyde sağlanması önemlidir. Bu araştırmada tanımlanan örneklem, hem okul hem öğrenci sayısı açısından yeterli büyüklüğe sahiptir.

## Bulgular

Bu çalışmada toplamda 312 farklı modelleme yapılmış ve kurulan modellere ait istatistikler aracılığı ile iki araştırma sorusu yanıtlanmaya çalışılmıştır. Elde edilen bulgular bu bölümde iki ayrı başlık halinde verilmiştir. Bu bölümde yorumlar yapılırken ağırlıklandırma değişkeninin kullanımının ABD, Fas ve Ürdün puanlarına ilişkin standart hatalarda ciddi farklılaşmaya (%100'den fazla bir artış) yol açması nedeniyle bu ülkelerin sonuçları dikkate alınmamıştır. Yorumlar 36 ülkenin verileri üzerinden yapılmıştır. Buna karşın, bulguların diğer araştırmacılar tarafından incelenebilmesi adına bu ülkelere ait istatistiklerin hem tablolarda hem de grafiklerde bulunmasına karar verilmiştir.

### 1. Araştırma Sorusuna İlişkin Bulgular

Öğrencilerin matematik başarı puanlarındaki varyansın okullar arası farklar tarafından açıklanan kısımları yüzde olarak 9,18 (Bahreyn) ile 60,51 (Hong Kong) arasında; fen başarı puanlarındaki varyansın okullar arası farklar tarafından açıklanan kısımları ise 6,30 (Kore) ile 61,93 (BAE) arasında değişmektedir. Bu farkların ülkeden ülkeye önemli değişiklikler göstermesinin yanı sıra aynı ülkenin verilerinde ağırlıklandırmanın nasıl yapıldığına göre açıklanan yüzdelerin de değiştiği gözlenmiştir. Matematik başarısının çıktı değişkeni olarak ele alındığı modellerde varyansın okullardan kaynaklanan kısmının en yüksek olarak hesaplandığı durumlar görece dengeli bir dağılım göstermiştir. En yüksek yüzdeler, 6 ülkede ağırlıklandırma yapılmadığı durumda, 14 ülkede yalnızca 1. düzey ağırlıklandırma yapıldığı durumda, 10 ülkede yalnızca 2. düzey ağırlıklandırma yapıldığı durumda ve 12 ülkede her iki düzeyde de ağırlıklandırma yapıldığı durumda elde edilmiştir. 6 ülkenin en yüksek yüzdeleri ise ağırlıklandırma durumlarına göre birebir birbirine eşittir. Bu halde denebilir ki, matematik başarısındaki varyansın okullardan kaynaklanan kısmı, yalnızca 1. düzey ağırlıklandırmanın yapıldığı ve her iki düzeyde de ağırlıklandırmanın yapıldığı durumda daha yüksek olarak hesaplanmaktadır. Fen başarısının çıktı değişkeni olarak ele alındığı modellerde varyansın okullardan kaynaklanan kısmının en yüksek olarak hesaplandığı durumlar yine görece dengeli bir dağılım göstermiştir. En yüksek oranlar, 7 ülkede ağırlıklandırma yapılmadığı durumda, 14 ülkede yalnızca 1. düzey ağırlıklandırma yapıldığı durumda, 9 ülkede yalnızca 2. düzey ağırlıklandırma yapıldığı durumda ve 12 ülkede her iki düzeyde de ağırlıklandırma yapıldığı durumda elde edilmiştir. 6 ülkenin en yüksek oranları ise ağırlıklandırma durumlarına göre birebir birbirine eşittir. Bu halde denebilir ki, fen başarısındaki varyansın okullardan kaynaklanan kısmı, yalnızca 1. düzey ağırlıklandırmanın yapıldığı ve her iki düzeyde de ağırlıklandırmanın yapıldığı durumda daha yüksek olarak hesaplanmaktadır.

### 2. Araştırma Sorusuna İlişkin Bulgular

Matematik başarısının çıktı değişkeni olduğu modellerde kestirimlerin standart hataları 2,38 (Norveç) ile 9,19 (Mısır) arasında; fen başarısının çıktı değişkeni olduğu modellerde kestirimlerin standart hataları ise 2,01 (Japonya) ile 9,29 (Mısır) arasında değişmektedir. Standart hataların ülkelerin verilerinde ağırlıklandırmanın nasıl yapıldığına göre değişiklik gösterdiği gözlenmiştir.

Matematik başarısının çıktı değişkeni olduğu modellerde en düşük standart hatalar, 26 ülkede ağırlıklandırma yapılmadığı durumda, 5 ülkede yalnızca 1. düzey ağırlıklandırma yapıldığı durumda, 7 ülkede yalnızca 2. düzey ağırlıklandırma yapıldığı durumda ve 3 ülkede her iki düzeyde de ağırlıklandırma yapıldığı durumda elde edilmiştir. 5 ülkede ise en düşük standart hatalar farklı ağırlıklandırma durumlarına göre birbirine eşittir. Fen başarısının çıktı değişkeni olduğu modellerde en düşük standart hatalar, 28 ülkede ağırlıklandırma yapılmadığı durumda, 6 ülkede yalnızca 1. düzey ağırlıklandırma yapıldığı durumda, 4 ülkede yalnızca 2. düzey ağırlıklandırma yapıldığı durumda ve 1 ülkede her iki düzeyde de ağırlıklandırma yapıldığı durumda elde edilmiştir. 3 ülkede ise en düşük standart hatalar farklı ağırlıklandırma durumlarına göre birbirine eşittir.

## Sonuçlar, Tartışma ve Öneriler

Bu çalışmada geniş ölçekli testlerden elde edilen verilerle yapılan analizlerde olası değerlerin çıktı değişkeni olarak birlikte ele alındığı durumda, farklı şekillerde ağırlıklandırma yapılmasının kestirimlere (özellikle başarı puanlarındaki varyansın okullar arasındaki farklılıktan kaynaklanan kısmı ve standart hatalar) etkisi incelenmiştir. Hem matematik, hem fen başarısının çıktı değişkeni olduğu durumlarda, sadece 1. düzey ağırlıklandırma yapılması durumunda katsayılar daha yüksek kestirilmiştir. En yüksek model güvenirlikleri yalnızca 2. düzey ağırlıklandırma ya da her iki düzeyde de ağırlıklandırma yapılması durumunda elde edilmiştir. En düşük standart hatalar, ağırlıklandırma yapılmayan durumda elde edilmiştir (matematikte 26 ülkede, fende 28 ülkede). Ağırlıklandırmanın yapıldığı durumda standart hataların yükselmesi durumu Carle (2009) ve Tat ve diğerlerinin (2019) çalışmalarında da raporlanan bir bulgudur. Bu açıdan bulgular paralellik göstermektedir. En düşük $\chi^2$ değerleri, yalnızca 1. düzeyde ağırlıklandırma yapıldığı durumda elde edilmiştir (matematikte 16 ülkede, fende 15 ülkede).

Bu çalışmada, matematik başarısının çıktı değişkeni olduğu modellerde ağırlıklandırmanın farklı şekillerde ele alınması durumuna göre hesaplanan oranlar arasındaki farkların bir ülke (Malezya) için %20'ye kadar çıktığı gözlenmiştir. Ayrıca, Fas, Şili ve ABD ülkelerinin verileriyle kurulan modellerde de bu oranların ciddi farklılık gösterebileceği görülmüştür. Yine çalışmada, fen başarısının çıktı değişkeni olduğu modellerde ağırlıklandırmanın farklı şekillerde ele alınması durumuna göre hesaplanan oranlar arasındaki farkların bir ülke (Malezya) için %18,75'e kadar çıktığı gözlenmiştir. Ayrıca, ABD ve Şili ülkelerinin verileriyle kurulan modellerde de bu oranların ciddi farklılık gösterebileceği görülmüştür. Ek olarak, kurulan modeller yakından incelendiğinde (tablolardan görülebilir) ABD, Fas ve Ürdün ülkelerinde ağırlık değişkeninin farklı şekillerdeki varlığı standart hatalarda ciddi değişikliklere yol açmaktadır. Bu nedenle bu ülkelerin verileri ile ulaşılan bulgular yorum dışı bırakılmıştır. Ancak bulguların diğer araştırmacılar tarafından incelenebilmesi adına bu ülkelere ait istatistiklerin hem tablolarda hem de grafiklerde bulunmasına karar verilmiştir. Araştırma bulgularından ulaşılabilecek en önemli sonuç, geniş ölçekli testlerin verilerinin analizinde HLM kullanılırken ağırlıklandırma değişkeninin kullanılıp kullanılmamasının (veya nasıl kullanıldığının) istatistiksel çıkarımları ciddi derecede farklılaştırabileceğidir.

Çok düzeyli verilerin HLM ile analizinde örneklem ağırlıklandırmalarının kullanılmasının yararlarının vurgulanmasına (Rutkowski ve diğerleri, 2010) karşın, ampirik kanıtlar ağırlık değişkenlerinin pek çok istatistiği ciddi şekilde etkilediğini göstermektedir. Ağırlık değişkenlerinin kullanımı ile ilgili olarak Laukaityte ve Wiberg (2017) ağırlık değişkeninin kullanılmamasının ya da yalnızca 1. düzeyde (öğrenci) kullanılmasının yanıltıcı sonuçlara yol açabileceğini ifade etmektedir. Mang ve arkadaşları (2021) tarafından geniş ölçekli test verileri ile yürütülen bir çalışmada hangi ağırlıklandırma senaryosunun en iyi çalıştığı incelenmiştir. Araştırma sonucunda, ağırlık değişkenlerinin eklenmesinin model güvenirliklerini arttırırken standart hataları da arttırdığı raporlanmıştır. Araştırmacılar, sonuç olarak ağırlıklandırmanın ikinci düzeyde (okul) ya da her iki düzeyde de yapılmasını önermektedir. Ancak, bu araştırmanın sonuçlarından hareketle, çıktı değişkenindeki varyansın ne kadarlık kısmını ikinci düzeydeki değişkenler tarafından açıklanabileceğine odaklanılan çalışmalarda ağırlıklandırmanın çıktı değişkenindeki varyansın okullar tarafından açıklanan kısmını artırabileceğinin akılda tutulması gerektiği söylenebilir.

Araştırma bulgularından görüldüğü üzere çözümlemeler yapılırken ağırlıklandırma değişkeninin kullanıp kullanılmama durumuna göre farklı sonuçlar elde edilebilmektedir. Bu nedenle araştırmacılar çözümlemedeki basamakları açıkça paylaşmalıdır. Özetle, bu çalışmada geniş ölçekli testlerden TIMSS'in 2019 döngüsüne katılan 39 ülkenin verileriyle 312 model kurulmuş, modeller çözümlenmiş ve 36 ülkenin verilerinin yorumlanması ile şu sonuçlara ulaşılmıştır:

- Araştırmacılar alan yazının da önerdiği üzere olası değerleri birlikte ele alabilecekleri modeller kurmalı ve bu modellerde ağırlık değişkenini ne şekilde ele alacaklarını planlamalıdır.
- Ağırlıklandırma değişkeninin modellemelerde ele alınış biçimine bağlı olarak katsayı, standart hata, güvenirlik ve $\chi^2$ kestirimleri değişmekte ve bunlara bağlı olarak çıktı değişkenindeki varyansın okullar arası farklılıktan kaynaklanan kısmı değişmektedir.
- Başarı varyansındaki okullar arası farklılıktan kaynaklanan kısmın olduğundan yüksek hesaplanması araştırmacıları yanıltabilir. Örnek vermek gerekirse, başarıdaki varyansı incelerken Malaysia verileri ile bir çalışma yürütüldüğünde başarıdaki varyansın okullar arasındaki farktan kaynaklanan kısmı, olduğundan %20 düşük –ya da %20 yüksek– hesaplanabilir. Bu durumda, araştırmacı devam modellemelerde 1 ve 2. düzeyde ele alacağı değişkenleri seçmekte zorlanabilir.

Bu çalışmada geniş ölçekli test verisi olarak yalnızca TIMSS'in 2019 döngüsünün verileri kullanılmıştır. Diğer geniş ölçekli testlerde ve TIMSS'in diğer döngülerinde durumun ne olduğuna bakmak alan yazındaki yorumları güçlendirme imkanı sunabilir.