Science Education International
34(2), 86-95
https://doi.org/10.33828/sei.v34.i2.2

International Council of Association for Science Education

ORIGINAL ARTICLE

Multilevel Analysis of Factors that Determine the Science Achievement of Fourth-grade Students in TIMSS 2019

Pongprapan Pongsophon*

Division of Science Education, Faculty of Education, Kasetsart University, Bangkok, Thailand

***Corresponding Author:** pongprapan.p@ku.th

ABSTRACT

This study examined the factors that determined the science achievement of fourth-grade students on the Trends in International Mathematics and Science Study (TIMSS) 2019 in the USA. The data were retrieved from the TIMSS international database and imported to the R program for manipulation. The EdSurvey package was used to conduct multilevel analysis, taking plausible values of outcome variables, and weighting into account. Level 1 predictors included gender, sense of school belonging, no student bullying, instructional clarity, and attitude toward science. Level 2 predictors included instruction being affected by science resource shortages, school emphasis on academic success, a lack of school discipline problems, and students' preexisting with literacy and numeracy skills. The findings include that the variance between schools equals 33% and boys outperformed girls in science. Students' attitudes toward science, their confidence, and the school's emphasis on academic success were the two strongest positive predictors of both individual student's achievement and the school's mean science achievement. The other predictive variables were found to have only trivially positive or no effect. The utilized models could explain the variance within and between schools by 6% and 39%, respectively. The implications for policymakers and educational practitioners include boosting up confidence in learning science and by actively engaging students in inquiry-based activities and providing instant and constructive feedback, encouraging, and treating all students regardless of gender identity equally in teaching and learning science and schools should emphasize academic success.

KEY WORDS: Fourth-grade students; multilevel analysis; science achievement; TIMSS 2019

INTRODUCTION

Academic achievement is one of the most important determinants of education quality. Educational researchers agree that many factors have an impact on students' achievements (Coleman et al., 1966; Engin-Demir, 2009). International assessments can provide remarkable opportunities to policymakers and educational practitioners including school principals and teachers to assess the quality of teaching and learning of subjects at the national level. Many countries are using the data as a basis for establishing achievement goals and standards and then implementing educational reforms to support meeting the goals and standards (Mullis and Martin, 2012). The International Association for Evaluation of Educational Achievement (IEA) encourages researchers worldwide to investigate a wide variety of student, classroom, and school-level factors that may explain the variation in students' mathematics and science achievement within and between schools and the level of science achievement of different students' groups, such as students from public and private schools, boys and girls, ethnic groups, and students from rural and urban schools (Fishbein et al., 2021).

The Trends in International Mathematics and Science Study (TIMSS) is one of the most comparative assessments of fourth- and eighth-grade students' achievement in mathematics and science initiated by IEA and conducted every 4 years since 1995 (Mullis et al., 2012). TIMSS is an international assessment that monitors trends in student achievement in mathematics and science. TIMSS provides reliable and timely trend data on the mathematics and science achievement of the participating countries (Mullis et al., 2020). Several studies have been conducted to examine the factors that influence students' academic achievement in TIMSS (Neuschmidt et al., 2008; Frempong, 2010; Liou and Liu. 2015). In 2019, 70 countries participated in the assessments (Mullis et al., 2020). TIMSS provides important data about students' learning contexts that are based on questionnaires completed by students and their parents or caregivers, teachers, and school principals. These contextual factors are associated with the achievement (Baker et al., 2002; Azina, and Halimah, 2012, Caponera and Losito, 2016). For example, Caponera and Losito (2016) found that a high socioeconomic status (SES) had a significant and positive effect on student achievement compared with students from socioeconomic disadvantaged schools and students from advantaged schools performed better in mathematics achievement. Research studies indicated that student characteristics such as gender, age, motivation, and attitudes toward courses, self-efficacy, students' efforts, being bullied at school have significant impacts on academic achievement (Engin-Demir, 2009; Alkhateeb, 2001; Gevrek and Sieberlich, 2014). It has also been reported that the two

Science Education International
34(2), 86-95
https://doi.org/10.33828/sei.v34.i2.2

most influential factors on students' learning achievement are the students' SES and attitudes toward the subject (Crane, 2001; Baker et al., 2002; O'Dwyer, 2005). However, mixed results were obtained for gender. Some studies favor boys (Neuschmidt et al., 2008; Frempong, 2010; Teodorović, 2012) while other studies favor girls (Alkhateeb, 2001; Azina and Halimah, 2012). Results also indicated that school-level factors such as school climate, general school resources, school discipline and safety, and parental support, accounted for 40% of the total variance in students' achievement scores in the USA (Borman and Dowling, 2010).

These above studies suggest the need for more in-depth analyses. The multilevel model for analysis of TIMSS results - intended to explore whether the impact of student- and school-related factors on student achievement. The implications of the findings are essential for educational policymakers to monitor school activities and students' learning (Badri, 2019). They can assist policymakers to uncover the strengths and weaknesses of their educational systems. TIMSS findings have been used in a wide variety of ways in different countries such as changes in revision of curricula, curriculum reform, and teachers' professional development, and new topics and contents were added to the science curriculum.

In this present study, I investigated factors determining the science achievement of fourth-grade students in the USA in TIMSS 2019. We chose the USA because the country has an interesting education system—public education is decentralized, with each state governing its own school system. Local school districts are responsible for curriculum decisions, the implementation of standards, facilities construction and maintenance, and the operation of school programs. In addition, there is no national curriculum in the United States but standards of science contents and competencies that guide the development of a school-based curriculum. State education agencies and school districts are responsible for subject area curriculum frameworks, and local school districts are also responsible for implementing the curriculum standards. For all states and districts, the curricula for mathematics and science prescribe a series of topics, content standards, and indicators of student achievement.

In 2012, the National Research Council *published A Framework for K–12 Science Education, which provided* a vision for science education in the United States. Based on the recommendations in the NRC framework, new science standards for kindergarten through Grade 12, called Next Generation Science Standards (NGSS), were published in 2013 (NGSS Lead States, 2013). The NGSS are K–12 science content standards. The documents set the expectations for what students should know and be able to do. These standards give local educators the flexibility to design classroom learning experiences that stimulate students' interests in science and prepares them for future career and active citizenship.

The NGSS focuses on a limited set of core ideas in the natural sciences and in engineering, technology, and applications of science that build coherently across the grade levels. The NGSS also emphasizes the importance of crosscutting concepts that apply across disciplines, as well as the practices used by scientists and engineers that K–12 students should develop. The intent is a set of standards that provides a coherent, internationally benchmarked science education program for all K–12 students.

The present study aims to model explaining variation in science achievement scores of American fourth graders nationwide within and between schools by student and school-level predictors based on TIMSS 2019 data using the multilevel linear modeling methodology.

Research Questions

1. How much variance is there in science achievement at the student and school levels?
2. How much of that variance in science achievement at both levels is explained using multilevel analysis?

Hypothetical Model

In this study, a literature review (Crane, 2001; Baker et al., 2002; O'Dwyer, 2005) and highlighted results from TIMSS 2019 (Mullis et al., 2020) guided the construction of a hypothetical model of multilevel regression of the factors influencing fourth-grade students' science achievement in the USA. From the results across all participating countries, fourth-grade students with many home resources for learning had substantially higher achievement than did students with few home resources for learning. Students whose parents often engaged them in literacy and numeracy activities during their early childhood had much higher achievement in fourth grade than did students whose parents never or almost never initiated these activities. Attending a school where instruction was not affected by science resource shortages was associated with higher average achievement, and students attending schools with a higher emphasis on academic success also had higher average achievement. Students with a high sense of school belonging demonstrated higher achievement than did students with little sense of school belonging. Students who attended schools with fewer school discipline problems had higher average achievement. Most students were in schools that were reported to be very or somewhat safe and orderly; when bullying was present, it had a negative relationship with student achievement. Regarding affective factors of fourth-grade students, enjoying learning science and being very confident in science was strongly associated with higher average achievement, as was higher clarity of instruction. The hypothetical model for multilevel regression of the US data is illustrated in Figure 1.

METHODS

Nest Data and Multilevel Analysis

In social science research, data at a lower level are usually nested in a higher-level unit. For example, students are grouped in a class, the classes are grouped in schools, and

Science Education International
34(2), 86-95
https://doi.org/10.33828/sei.v34.i2.2

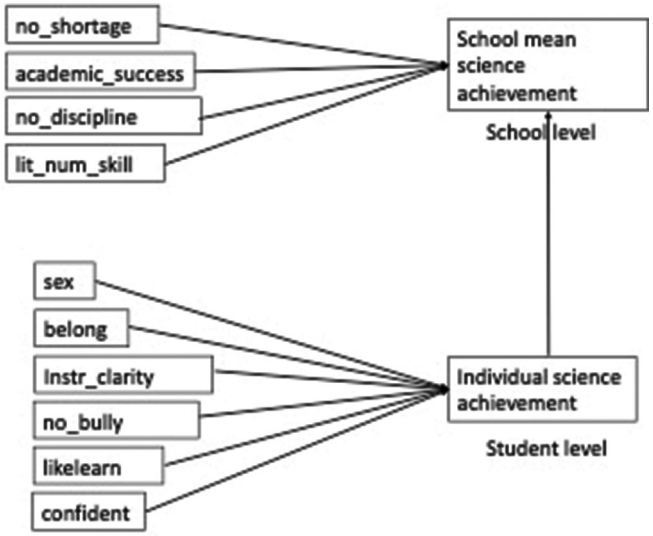

Figure 1: Hypothetical model of multilevel regression of fourth-grade students' achievement in the USA

the schools are grouped in an educational district. Data in the same unit are more likely similar than those in other units at the same level. In a school context, for example, class A is taught by teacher A, while class B is taught by another teacher; therefore, the students in class A are similarly influenced by a number of factors. This means that their achievements will vary less among their cohort (class) than they do in comparison to class B. If data analysis of nested data does not take into account the structure of the data, the result will be inaccurate and biased. Therefore, nested data should be analyzed by multilevel analysis. The hierarchical linear model is a type of regression analysis for multilevel data, in which the dependent variable is at the lowest level and explanatory variables can be defined at any level. This study investigates the effect of explanatory variables on the science achievement of fourth-grade students nested in a school. TIMSS 2019 data are clustered. Selected explanatory variables were added into a regression model at both levels to explain the variability of the scores at each level. Issues related to the analysis of a large-scale survey such as TIMSS are sampling technique and plausible values. This study resolves these issues using the R program, a free software environment for statistical computing and graphics. R contains Packages are the fundamental units of reproducible R code. To analyze multilevel regression in this study, we used a package named EdSurvey, which was developed to analyze large-scale data, taking into account sampling, weighting, and plausible values (Bailey et al., 2020).

Data Retrieval, Plausible Values, and Weighting

TIMSS 2019 data are available from the TIMSS 2019 International Database, including student achievement data and student, teacher, school, and curricular contextual data. Due to the study's multistage sample design and use of imputed scores (also known as plausible values), the data are complex. In TIMSS, to measure achievement in a large population, it is more efficient to use a matrix sampling design, in which each subject responds to relatively few items than it is to create long assessments for each test-taker. Although a matrix sampling design does not facilitate the creation of precise statements about individuals, it allows for more efficient estimation of population characteristics (Lorah, 2019). The implication of this design is that individual scores contain a large amount of uncertainty; plausible values, which are represented by five scores for each student, are utilized to model this uncertainty. These plausible values represent multiple imputations of the latent construct (Wu, 2005). When conducting analyses, it is important to note although it may be possible to recover population parameters based on only one plausible value, this practice is discouraged, as is averaging the plausible values (Wu, 2005; Rogers and Stoeckel, 2008).

TIMSS is an international large-scale assessment in education. TIMSS data also include sampling weights that adjust for unequal probability of selection and are included in the analysis to avoid bias. Sampling weights are available at multiple levels (for example, students and schools) and must be scaled appropriately (Rutkowski and Delandshere, 2016; Kwiek, 2018; Laukaityte and Wiberg, 2018; Wagemaker, 2020; Hernández-Torrano and Courtney, 2021). With a multilevel model, these sampling weights can be included in the analysis, and there are software options that can do this, such as the BIFIEsurvey package, WeMix package, EdSurvey package, and Rstan package in R. This study used the EdSurvey package. EdSurvey is an R statistical package designed for the analysis for national and international education data from the National Center for Education Statistics (NCES). EdSurvey was developed by the American Institutes for Research and commissioned by the NCES. Many datasets using complex survey designs include replicate weights, which can be used to adjust for cluster sampling and the implied non-independence of individual observation. Failure to account for non-independence of observation could lead to downwardly biased standard errors, which would inflate Type I error rates (Snijders and Bosker, 2012). The use of replicate weights essentially represents a resampling method that can empirically derive unbiased standard error estimates (Martin and Mullis, 2012).

The multilevel model is particularly well-suited for analyzing complex survey data such as TIMSS because it directly models different levels of data that can correspond to a cluster sampling design; as such, the multilevel model is frequently used for the analysis of complex survey data (Lorah, 2019). When questions related to the connection of variables at multiple levels (such as students and schools) are investigated, multilevel models can be used (Snijders and Bosker, 2012).

Questionnaire and Scales

All explanatory variables in this study, except gender, are TIMSS 2019 Context Questionnaire transformed scale scores (Table 1). For example, Instructional Clarity in Science Lessons seeks to measure students' perceptions about the clarity of instruction in their science lessons based on their responses to six statements. For each of the six statements,

Science Education International
34(2), 86-95
https://doi.org/10.33828/sei.v34.i2.2

Table 1: Descriptive statistics of level 1 and level 2 weighted predictors

Variable names	Labels (abbreviations)	Weighted means/cutoff points	Weighted SD
ID student	School ID		
ID school	Student ID		
Level 1 predictors			
It sex	Gender (sex)		
asbgssb	Students sense of school belonging (belong)	9.67 high sense of school belonging 9.6 ≥some sense of school belonging ≤_7.2 little sense of school belonging	2.19
asbgsb	No student bullying (no_bully)	9.99 never or almost never 9.2 ≥ about monthly ≤7.4 about weekly	1.90
asbgics	Instructional clarity in science lessons (instr_clarity)	10.13 high clarity 8.8 ≥moderate clarity ≤6.9 low clarity	1.99
asbgsls	Students like learning science (likelearn)	10.16 very much like 9.7 ≥Somewhat Like ≤7.6 Do Not Like	2.34
asbgscs	Students are confident in science (confident)	9.88 very confident 10.2 ≥somewhat confident ≤8.2 not confident	1.88
totwgt	Overall weight		
Level 2 predictors			
Acbgsrs	Instruction not affected by science resource shortage (no_shortage)	11.19 not affected 11.4 ≥somewhat affected ≤7.0 affected a lot	2.26
Acbgeas	School emphasis on academic success (academic_success)	10.06 very high emphasis 13.0 ≥high emphasis ≤9.2 medium emphasis	2.35
Acbgdas	No school discipline problems (no discipline)	9.83 hardly any problems 9.7 ≥minor problems ≤7.6 moderate-to-severe problem	1.46
Acbglns	Students enter with literacy and numeracy skills lit_num_skills	11.62 more than 75% 11.5 ≥25–75% ≤25%	1.99
Schwgt	Overall school weight		

students were asked to indicate a degree of agreement with the statement: agree a lot, agree a little, disagree a little, or disagree a lot. Using the item response theory partial credit model, the data from student responses were placed on a scale constructed so that the scale center point was located at the combined mean score of all fourth graders who took the TIMSS in 2019. TIMSS developed a system of production programs for calibration of the items on each scale using ConQuest and production of scale scores for each scale respondent.

Multilevel Analysis

We incorporated data from J schools, with a different number of students (n_j) in each school. On the student level, the outcome variable "weighted science score" (Y) was measured by a test. We utilized six explanatory variables at level 1, which were measured by a student questionnaire. The first was gender (sex; X_1: 0 = boy, 1 = girl) and the other five were continuous variables: students' sense of school belonging (belong; X_2), no student bullying (no_bully; X_3), instructional clarity (instr_clarity; X_4), student enjoyment of science learning (likelearn; X_5), and student confidence in science (confident; X_6). This study incorporated data on 7,047 students in 256 schools, and the average class size was 28 students. There is no ethical issue since it is secondary data in the public domain with the identity of participants and informants protected.

To analyze these data, we set up separate regression equations in each school to predict the outcome variables using the explanatory variables, as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X5_{ij} + \beta_{6j}X_{6ij} + e_{ij}$$

Using variable labels instead of algebraic symbols, the equation read:

$$Y_{ij} = \beta_{0j} + \beta_{1j}sex_{1ij} + \beta_{2j}belong_{ij} + \beta 3jno\_bullyij + \beta_{4j}instr\_clarity_{4ij} + \beta_{5j}likelearn_{ij} + \beta_{6j}confident_{6ij} + e_{ij}$$

In this regression equation, β_{0j} is the intercept, β_{1j} is the regression coefficient for the dichotomous explanatory variable gender, and β_{2j}-β_{6j} is the regression coefficients for the continuous explanatory variables (belong, no_bully, instr_clarity, likelearn, and confident, respectively). The subscript j is for the schools ($j = 1 \ldots J$) and subscript i is for the individual students ($i = 1 \ldots nj$). Each class had a different intercept β_{0j} and different slope coefficients β_{2j}-β_{6j}. The residual error, e_{ij}, was assumed to have a mean of zero and a variance (σ^2_e) to be estimated. It was assumed to be the same in all schools. Since the intercept and slope coefficients are random variables that varied across the schools, they are referred to as random coefficients.

In this study, we attempted to explain the variation of the intercepts β_{0j} by only introducing explanatory variables at the school level (level 2). These included the effects of science instruction resource shortages (no_shortage; Z_1), school emphasis on academic success (academic_success; Z_2), school discipline problems (no_discipline; Z_3), and students' preexisting literacy and numeracy skills (lit_num_skill; Z_4). u_{oj} is the random residual error at the school level. It was assumed to have a mean of 0 and to be independent from the residual errors e_{ij} at the student level. The variance of the residual errors u_{oj} was specified as $\sigma^2_{u_0}$.

Science Education International
34(2), 86-95
https://doi.org/10.33828/sei.v34.i2.2

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + \gamma_{04}Z_{4j} + u_{0j}$$

Using variable labels instead of algebraic symbols, the equation read:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}shortage_j + \gamma_{02}academic\_success_j \gamma_{03}discipline_j + \gamma_{04}lit\_num\_skills_j + u_{oj}$$

Our model, including student-level and school-level explanatory variables, can be written as a single complex regression equation, as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + \gamma_{04}Z_{4j} + u_{0j} + \beta_{1j}X_{1ij} + + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X5_{ij} + \beta_{6j}X_{6ij} + e_{ij}$$

The segment $[\gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + \gamma_{04}Z_{4j} + + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X5_{ij} + \beta_{6j}X_{6ij}]$ contains the fixed coefficients and is the fixed part of the model. The segment $[u_{0j}+e_{ij}]$ contains the random error terms, called the random part of the model. Since grouped data observations from the same group were more similar to each other than to observations from different groups, the amount of dependence was expressed as the intraclass correlation (ICC) (ρ). ICC is a ratio of the amount of variance due to groups relative to the total variance of Y_{ij}. The resulting value is between 0 and 1.0, where higher values reflect greater between-group variability. It can be calculated as follows:

$$\text{ICC} = \rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e}$$

The model used to estimate ICC was a model that contained no explanatory variables, called the intercept-only model or unconditional means model. This is a null model that serves as a benchmark against which other, more complex, models are compared. The two more complicated models were that with level 1 explanatory variables and the model with both level 1 and level 2 explanatory variables. The variance of error terms of both levels was used to determine variance compared to the baseline model.

Centering is the rescaling of predictors by subtracting the mean. Centering makes this value more interpretable, as the expected value of Y when x (centered X) is zero represents the expected value of Y when X is at its mean. There are two different versions of centering in multilevel regression, grand-mean centering, and group-mean centering. Grand-mean centering subtracts the grand mean of the predictors using the mean from the full sample. Group-mean centering subtracts the individual's group mean from the individual's score. In this study, all level 1 predictors except gender were group-mean centered, while level 2 predictors were grand-mean centered.

RESULTS

Descriptive Statistics

The results, as shown in Table 1, indicated that fourth-grade American students had a high sense of school belonging and never or almost never experienced bullying in school. The students perceived that their teachers delivered high instructional clarity in science lessons, and they liked learning science very much. However, they had somewhat low confidence in their ability to learn science. As reported by the school principals, although instruction was somewhat affected by science resource shortages, the schools highly emphasized academic success. Regarding school discipline, the schools had hardly any problems. More than 75% of students possessed adequate preexisting literacy and numeracy skills. Descriptive statistics of the weighted and unweighted outcome variables are shown in Table 2.

Multilevel Analysis

Model 1: The intercept-only model.

It is written as: $Y_{ij}=\gamma_{00}+u_{0j}+e_{ij}$. In Table 3, the intercept-only model estimates the intercept as 543.69, which is simply the average science score across all schools and students. To measure the magnitude of the variation among schools in the mean achievement levels, we calculated the plausible values range for these means based on the between variance we obtained from the model: $543.69 \pm 1.96*(2{,}353^{1/2}) = (448.61, 638.76)$. The variance of student-level residual error (σ^2_e) was estimated as 4,779, and the variance of the school-level residual errors ($\sigma^2_{u_0}$) was estimated as 2,353. All estimated parameters were much larger than the corresponding standard errors, and they were all significant at $\rho < 0.01$. The interclass correlation, calculated as $\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e}$, was 2,353/7,132, which equals 0.33. Thus, 33% of the variance of the science scores was at the school level, which is very high. In the intercept-only model, the residual variance represents unexplained error variance.

Model 2: A Random Coefficient Model.

We ran a regression of science scores on group-centered student-level predictors for each school; in other words, we ran 256 regressions. Following is the equation that motivates this model:

$$Y_{ij} = \gamma_{00} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X5_{ij} + \beta_{6j}X_{6ij} + u_{0j} + e_{ij}$$

The 95% plausible value range for school mean scores was $430.91 \pm 1.96*(2{,}138^{1/2}) = (340.28, 521.54)$. The significant unstandardized predictors were gender (-8.26), no_bully (3.71), and confidence (8.42). The boys outperformed the girls by 8.26. The positive coefficient of no_bully indicates that as the value of no_bully increased, the means of science score also tended to increase. With one unit shift in no_bully,

Science Education International
34(2), 86-95
https://doi.org/10.33828/sei.v34.i2.2

Table 2: Descriptive statistics of weighted and unweighted outcome variables

		Weighted outcome variable				
Variable name	**Label**	**n**	**Min**	**Max**	**Mean**	**SD**
ssci	science score	8776	175.72	784.40	538.64	84.27
Unweighted outcome variables						
asssci01	First possible value science	8776	175.37	791.57	537.47	83.97
asssci02	Second possible value science	8776	194.83	774.30	535.68	84.68
asssci03	Third possible value science	8776	174.77	786.84	535.43	85.30
asssci04	Forth possible value science	8776	150.05	780.41	534.93	85.65
asssci05	Fifth possible value science	8776	183.59	788.85	537.16	84.20

Table 3: Multilevel models and their parameter estimates

Model	**Model 1**	**Model 2**	**Model 3**	
Fixed part	Coefficients (s.e.)	Coefficients (s.e.)	Coefficients (s.e.)	Standardized coefficients
Intercept	543.69 (4.62)	430.91 (12.90)	430.85 (12.45)	-
Sex		−8.26 (2.22)	−8.20 (2.2182)	−0.05*
Belong		0.54 (0.57)!	0.51 (0.57)	0.01
No_Bully		3.71 (0.56)	3.69 (0.56)	0.08*
Instr_clarity		−0.51 (0.61)!	−0.52 (0.61)	−0.01
Likelearn		−0.32 (0.57)!	−0.31 (0.57)	−0.01
Confident		8.42 (0.74)	8.41 (0.74)	0.19**
No_shortage			−2.99 (1.34)	−0.08*
Academic_success			9.11 (1.27)	0.26**
Discipline			5.52 (2.33)	0.10*
Lit_num_skill			4.34 (1.64)	0.10*
Random part				
σ_e^2	4,779 (110.4)	4483 (100.9)	4482 (101)	
σ_{u0}^2	2,353 (313.7)	2138 (292.2)	1422 (190.1)	

*ρ<0.05, **ρ<0.01

the science score increased by 3.71; with one unit shift in confidence, the science score increased by 8.41. Notice that the residual variance becomes 4,483, whereas the residual variance in the intercept-only model was 4,779. We can compute the proportion variance explained at level 1 as (4,779–4,483)/4,779 = 0.06. This suggests that using student-level predictors of science scores reduced the within-school variance by 6.2%.

Model 3: Level 1 and Level 2 Predictors.

This is an explanatory model that we built to account for variability. We wanted to model the following:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + \beta_{5j}X5_{ij} + \beta_{6j}X_{6ij} + e_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + \gamma_{04}Z_{4j} + u_{0J}$$

The 95% plausible value range for school mean scores was 430.84 ± 1.96*($1422^{1/2}$) = (356.09, 504.75). The level 1 predictors were group-centered, while the level 2 predictors were grand-centered. As in model 2, the significant unstandardized level 1 predictor were sex (−8.20), no_bully (3.68), and confidence (8.41). All level 2 predictors were significant: no_shortage (−2.99), academic_success (9.11), discipline (5.52), and Lit_num_skill (4.34). The negative coefficient of shortage suggests that, as no_shortage increased, the science score tended to decrease. The mean of the science score changed by −2.99 given a one-unit increase in no_shortage while holding other variables in the model constant.

The coefficients in Table 2 are all unstandardized regression coefficients. To interpret them properly, we must take the scale of the explanatory variables into account. One can derive the standardized regression coefficients from the unstandardized coefficients (Hox, 2002):

$$\text{Standardized coefficients} = \frac{\textit{unstandardized coefficient} * \textit{stand.dev.explanatory var.}}{\textit{stand.dev.outcome var.}}$$

The standardized coefficients in Table 3 indicate that among level 1 predictors, confidence (0.19) is the strongest explanatory variable, while academic_success (0.26) is the strongest explanatory variable of all the level 2 predictors.

One of the first explained variance measures for multilevel models was based on the reduction of unexplained variance

Science Education International
34(2), 86-95
https://doi.org/10.33828/sei.v34.i2.2

when predictors are added (Bryk and Raudenbush 1992). Explanation of variance is achieved by the subtraction of the residual variance of a baseline model (model 1) from the residual variance of a full model (model 3), which is then divided by the residual variance of the baseline model (model 1). Using the intercept-only as the baseline model, the variance explained at each level was calculated: at level 1, (4,779–4,482)/4,779 = 0.06 or 6%; at level 2, (2,353–1,422)/2,353 = 0.39 or 39%. The remaining variance indicated that there must have been other explanatory variables that were omitted from the full model.

DISCUSSION

The Interclass Correlation

The interclass correlation (ICC) coefficient equals 0.33, meaning the variance between schools is 33%, while the variance within schools is 67%. The variance at both levels is large enough for multilevel analysis (Bryk and Raudenbush, 1992). Focusing on the variation between schools, the US education system is decentralized; each state is eligible to adopt or develop their own curriculum standards. As mentioned earlier, the USA has no national curriculum but standards that specify appropriate content to be taught at a particular grade level. Curriculum frameworks provide guidance for implementing the content standards adopted by the State Board of Education (SBE). The selection of textbooks and learning materials is another source of variation of science achievement across and within a stage. Textbooks are influential in determining what many teachers teach and, in turn, what students learn. For grades one through eight, most of the textbook and material selection occurs at the state level by the SBE, which is assisted by committees of volunteers who specialize in each discipline. The board selects several texts for the same subject and grade level, and individual school districts choose from this list. Local schools, however, can petition for permission to use textbooks that do not appear on the official lists.

This study found that science achievement varies by 67%, even within a school. This large variance is rooted in the history of education in the USA, which has revolved around inclusivity and diversity in the classroom. It has been a little more than 60 years since Brown vs. Board of Education, one of the most important supreme court cases in the history of the United States, which made it illegal to segregate public schools on the basis of race. The National Education Association reports that 2014 was the 1st year, in which the majority of students in American public schools represented racial and ethnic minorities. Diversity in a classroom means that students can contribute different and divergent perspectives. This leads to increased cultural understanding, stronger critical thinking skills, and enhanced creativity, all of which better prepare students for adulthood (Konan et al., 2010).

Strong Student and School Predictors

From the standardized coefficients in Table 3, the two highest values are students' confidence in science (confident) and school emphasis on academic success (academic_success). Confidence is belief in one's ability to successfully perform a task. Positive responses may indicate past successes in the science classroom and students' authentic science performance, which is transferable to the TIMSS 2019 assessment. There are a number of studies that found a strong relationship between students' confidence and their academic achievement (Whitesell et al., 2009; Booth and Gerard, 2011; Cvencek et al., 2018). Chang and Cheng (2008) found that there was a statistically significant correlation, with a moderate effect size, between Taiwanese senior high school students' science achievement and their self-confidence and interest in science. Students with low self-confidence regarding science can be assisted by eliciting their initiative, encouraging them, scaffolding their problem-solving, and providing instant and constructive feedback. Teachers should create a respectful and encouraging atmosphere in the classroom, in which students support each other.

Principals' responses were used to complete the school emphasis on Academic Success Scale. The principals were largely focused on how well teachers understood the curriculum's goal and could effectively implement it in their classrooms. They also believed that parental engagement and involvement in their children's education are essential aspects of healthy learning, as a high level of parental engagement creates a supportive learning environment at home. Parents who pay attention to and follow up on the progress of their children's learning can complement learning at school (Badri, 2019).

When a school's emphasis on academic success was mentioned, the importance of collective efficacy was often expressed. According to Lezotte (2001), one of the characteristics of an effective school is high expectations, bolstered by the belief and persuasion of school staff, that lead students to obtain mastery of the school's essential curriculum. Moreover, significant relationships have been found between school emphasis on academic success and students' performances in both science and mathematics achievement tests across all countries involved in TIMSS 2011 (Mullis et al., 2012).

Regarding the relationship between science achievement and gender, this study found that boys outperformed girls at the fourth-grade level in both mathematics and science. This is consistent with PISA's 2015 results, which also found that boys have more confidence in their science skills than do girls. When a student believes in their ability to solve a scientific problem, they are said to have a high level of self-efficacy. Students who have low self-efficacy in science do not perform as well as students who trust in their ability to use their scientific knowledge in their daily lives.

Regarding no_bully, students who reported never experiencing bullying, or only experiencing it a few times a year, gained higher scores than did those who reported experiencing it once or twice a month or at least once a week. A high score, then, indicated a rarity of student bullying in a given school. This study found that the no_bully variable was positively

associated with fourth-grade science achievement. Previous studies (Juvonen et al., 2011; Ladd et al., 2017) have also found that school bullying was linked to lower academic achievement. Children who suffered chronic levels of bullying during their school years had lower academic achievement, a greater dislike of school, and less confidence in their academic abilities. Schools should have anti-bullying programs, and parents should ask their children if they are being bullied or excluded at school.

Issues related to school discipline served as a level 2 predictor in the model. These were reported by the school principal responses signifying not a problem or only a minor problem received a high score, while responses indicating a moderate or severe problem got a low score. Consistent with previous studies (Mullis et al., 2012), this study found that no school discipline problem was positively related to a school's mean science achievement. Fewer school discipline problems generally imply a safe and supportive learning environment, and when students feel they are secure and protected, they can concentrate and more fully engage in the learning process. Safety and order in schools were strongly related with students' physical and emotional security. This finding is consistent with that of the study conducted by Ceylan and Sever (2020), who found that, in schools with few discipline problems, there is a tendency to prioritize academic success. In addition, allocating time to emphasizing academic success is strongly related with having teaching time that is not interrupted by undesirable behaviors.

To determine the preexisting skills variable (lit_num_skill), principals were asked how many of the students in their school are able to perform 12 tasks that represent literacy and numeracy skills when they begin the first grade of primary/elementary school. The more students that can perform these activities, the more one can assume an emphasis on literacy and numeracy skills before they entered the school. This study found a positive association between this variable and science achievement. Students who start with these abilities have more potential to successfully learn science when they move up to higher grades. This result is consistent with many previous studies, such as Chen et al. (2020) and Pace et al. (2018). At the school level (Chen et al., 2020), it was found that school emphasis on academic success was a strong predictor of science achievement among students in most Asian regions.

Unexpectedly, this study found that the instruction not affected by science resource shortages (no_shortage) variable had a negative association with science achievement. To determine this, principals were asked if their school's capacity to provide instruction had been affected by shortages. If they reported no effect, the score would be high; if they reported a significant effect, the score would be low. Previous studies (Greenwald et al., 1966; Hedges et al., 2016) conducted meta-analyses to assess the direction and magnitude of the relationships between a variety of school inputs and students' achievement. These studies found that a broad range of resources was positively related to student outcomes, with effect sizes large enough to suggest that moderate increases in spending may be associated with significant increases in achievement. This unexpected result should be further investigated in future studies.

Weak 2-Level Predictors

This study found that several variables had no effect on science performance in TIMSS 2019, including the sense of belonging, liking to learn science, and instructional clarity.

Goodenow (1992) defines a sense of belonging as the feeling of being included, accepted, and supported by other persons in a school social environment. When a student believes there is a personal connection to the school, engagement is more likely to occur. Furthermore, this attachment involves caring about what others think and trying to fulfill those expectations (Cothran and Ennis, 1997). Perceived friendliness from others and a sense of being personally valued are necessary but not solely sufficient for success. Belonging in a class must also include participation in the shared educational goals of that class (Goodenow, 1992).

Liking to learn science or having a positive attitude toward science was found to have no effect on science achievement. Mao et al. (2021) conducted a meta-analysis on the relationship between one's attitude toward science and academic achievement in science, and they found that empirical studies have produced inconsistent findings regarding this relationship. On the one hand, many studies have shown that students' attitudes toward science and their science achievement are moderately positively correlated (Wang and Liou, 2017; Zheng et al., 2019). On the other hand, other studies have found that this relationship is either quite weak, statistically non-significant, or even negative (Rennie and Punch, 1991; Gardner, 1995). For instance, based on a sample of sixth-grade students from Finland, Estonia, Latvia, and Belgium, Salmi et al. (2016) reported that the correlation between students' societal attitudes (the value of science in their society) and performance was positive, yet weak ($r = 0.11$), but the relationship between students' attitudes toward engineering (interest in computer design) and performance was negative ($r = -0.11$).

The meta-analysis performed by Mao et al. (2021) suggests that these inconsistencies may be due to a moderation effect on the relationship between attitude toward science and science achievement; these effects may be related to, for example, grade, geographical region, or publication type. A previous primary study (Liou and Liu, 2015) on attitudes toward science suggested that the relationship between attitude and achievement could vary across grade levels. Based on the TIMSS 2011 Taiwanese data, they noted that the correlations between students' self-concept and science scores, and between intrinsic interest and science scores, were stronger for the eighth-grade students than for the fourth-grade students.

This study surprisingly found that instructional clarity had

no effect on fourth-grade students' science achievement. This finding is inconsistent with previous studies, such as that by Zheng (2021). Zheng found that both teacher clarity and credibility are strong predictors of students' academic engagement and willingness to attend classes. In addition, the association between teacher clarity and students' willingness to attend classes can be reasonably justified by the fact that those students who experience organized and clear instruction are naturally more inclined to attend their classes.

CONCLUSIONS AND IMPLICATIONS

The findings of this study include confirmation of a gender gap in science achievement. Boys may have outperformed girls in science because the boys received more attention from their teachers. Boys might also have had more opportunities to engage in science or they might have dominated the class. To close the gender gap in science achievement, girls must receive equal treatment. It is also important that teachers introduced girls to role models, such as outstanding female scientists, be assigned to play key roles in group work, and be praised by their teachers. This study also found that confidence in science is one of the two strongest predictors of science achievement. To boost self-confidence in science, sources of self-efficacy (Bandura, 1997) can be applied such as mastery experiences, vicarious experiences, verbal persuasion, and physiological and affective states. Mastery experiences are those that students gain when they take on a new task in science assigned by their teacher and succeed. Vicarious experiences occur when students see classmates who are similar to themselves succeed in scientific investigations or quizzes by sustained effort. Social persuasion occurs when students receive positive verbal feedback from their teachers and friends while undertaking complex tasks such as designing an experiment, interpreting the data, and drawing a conclusion. Finally, emotional and physiological states refer to the way, in which the emotional, physical, and psychological well-being of students can influence how they feel about their personal abilities in a particular situation. Science teachers must ensure the learning environment supportive and friendly, so the students feel safe and have fun to learn science. The final strongest predictor at the school level is the school emphasis on academic success. There is a strong and direct link between school principals and student achievement. Principals' practices influence school conditions, teacher quality, instructional quality, and student achievement. Effective principals are responsible for establishing a schoolwide vision of commitment to high standards and the success of all students.

The Limitation of the Study and Suggestions for Future Research

With multilevel analysis framework, we could not examine the relationship among variables in the same levels. I, therefore, suggest future research to use more advanced statistical analysis called multilevel structural equation modeling when the units of the observation form a hierarchy of nested clusters and some variables of interest are measured by a set of items. Multilevel structural equation modeling combines the advantages of multilevel modeling and structural equation modeling and enables researchers to scrutinize complex relationships between latent variables on different levels.

Data Availability Statement

The data are in the public domain and retrievable on https://timss2019.org/international-database/the TIMSS 2019 International Database is available to researchers, analysts, and any individuals interested in the data collected and analyzed as part of TIMSS 2019.

Ethics Statement

As only secondary data were used in this study, ethical approval was not required.

REFERENCES

Alkhateeb, H.M. (2001). Gender differences in mathematics achievement among high school students in United Arab Emirates, 1991-2000. *School Science and Mathematics,* 101(1), 5-9.

Azina, I.N., & Halimah, A. (2012). Student factors and mathematics achievement: Evidence from TIMSS 2007. *EURASIA Journal of Mathematics, Science and Technology Education,* 8(4), 249-255.

Badri, M. (2019). School emphasis on academic success and TIMSS science/math achievements. *International Journal of Research in Education and Science*, 5(1), 176-189.

Bailey, P., Lee, M., Nguyen, T., & Zhang, T. (2020). Using EdSurvey to analyse PIAAC data. In Maehler, D., & Rammstedt, B. (Eds.), *Large-Scale Cognitive Assessment: Methodology of Educational Measurement and Assessment*. Germany: Springer. pp. 209-237.

Baker, D.P., Goesling, B., & Letendre, G.K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman-Loxley Effect" on mathematics and science achievement. *Comparative Education Review,* 46(3), 291-312.

Bandura, A. (1997). *Self-efficacy: The Exercise of Control*. New York: W. H. Freeman.

Booth, M.Z., & Gerard, J.M. (2011). Self-esteem and academic achievement: A comparative study of adolescent students in England and the United States. *Compare,* 41(5), 629-648.

Borman, G.D., & Dowling, M. (2010). Schools and inequality: A multilevel analysis of Coleman's equality of educational opportunity data. *Teachers College Record,* 112, 1201-1246.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical Linear Models: Applications and Data* Analysis Methods. United States: Sage Publications, Inc.

Caponera, E., & Losito, B. (2016). Context factors and student achievement in the IEA studies: Evidence from TIMSS. *Large-scale Assessment Education*, 4(12), 1-22.

Ceylan, E., & Sever, M. (2020). Schools' emphasis on academic success in TIMSS 2015 across Finland, Singapore, and Turkey. *International Journal of Psychology and Educational Studies,* 7(4), 203-212.

Chang, C.Y., & Cheng, W.Y. (2008). Science achievement and students' self-confidence and interest in science: A Taiwanese representative sample study. *International Journal of Science Education,* 30(9), 1183-1200.

Chen, Y., Guo, C., Lim, K.M., Mun, K., Otsuji, H., Park, Y., Sorrell, D., & So, W. (2020). The influence of school entry skills in literacy and numeracy on the science achievement of fourth grade students and schools in Asian regions. *Eurasia Journal of Mathematics, Science and Technology Education,* 16(9), em1877.

Coleman, J.S., Compbell, E.Q., Hobson, C.J., McPartland, J., Mood, A.M., Weinfeld, F.D., & York, R.L. (1966). *Equality of Educational Opportunity*. United States: Department of Health, Education and Welfare Office of Education.

Cothran, D.J., & Ennis, C.D. (1997). Students' and teachers' perceptions of conflict and power. *Teaching and Teacher Education,* 13, 541-553.

Crane, J. (2001). Effects of home environment, SES, and maternal test scores on mathematics achievement. *The Journal of Educational Research,* 89(5), 305-314.

Cvencek, D., Fryberg, S.A., Covarrubias, R., & Meltzoff, A.N. (2018). Self-concepts, self-esteem, and academic achievement of minority and majority north American elementary school children. *Child Development,* 89(4), 1099-1109.

Engin-Demir, C. (2009). Factors influencing the academic achievement of the Turkish urban poor. *International Journal of Educational Development,* 29, 17-29.

Fishbein, B., Foy, P., & Yin, L. (2021). *TIMSS 2019 User Guide for the International Database*. United States: TIMSS and PIRLS International Study Center, Lynch School of Education and Human Development, Boston College. Available from: http://hdl.handle.net/2345/bc-ir:109292 [Last accessed on 2023 Jan 19].

Frempong, G. (2010). Equity and quality mathematics education within schools: Findings from TIMSS data for Ghana. *African Journal of Research in Mathematics, Science and Technology Education,* 14(3), 50-62.

Gardner, P.L. (1995). Measuring attitudes to science: Unidimensionality and internal consistency revisited. *Research in Science Education,* 25, 283-289.

Gevrek, Z.E., & Seiberlich, R.R. (2014). Semiparametric decomposition of the gender achievement gap: An application for Turkey. *Labour Economics,* 31, 27-44.

Goodenow, C. (1992). *School Motivation, Engagement, and Sense of Belonging among Urban Adolescent Students*. San Francisco, CA: Paper Presented at the American Educational Research Association Convention.

Greenwald, R. Hedges, L.V., & Laine, R.D. (1996). The effect of school resources on student achievement. *Review of Educational Research,* 66(3), 361-396.

Hedges, L., Pigott, T., Polanin, J., Ryan, A., Tocci, C., & Williams, R. (2016). The question of school resources and student achievement: A history and reconsideration. *Review of Research in Education Centennial Edition,* 40,143-168.

Hernández-Torrano, D., & Courtney, M.G.R. (2021). Modern international large-scale assessment in education: An integrative review and mapping of the literature. *Large-scale Assessment Education,* 9, 17.

Hox, J. (2002). Multilevel Analysis Techniques and Applications. United States: Lawrence Erlbaum Associates Publishers.

Juvonen, J., Wang, Y., & Espinoza, G. (2011). Bullying experiences and compromised academic performance across middle school grades. *The Journal of Early Adolescence*, 31(1), 152-173.

Konan, P., Chatard, A., Selimbegovic, L., & Gabriel, M. (2010). Cultural diversity in the classroom and its effects on academic performance: A cross-national perspective. *Social Psychology*, 41, 230-237.

Kwiek, M. (2018). International research collaboration and international research orientation: Comparative findings about European academics. *Journal of Studies in International Education*, 22(2), 136-160.

Ladd, G.W., Ettekal, I., & Kochenderfer-Ladd, B. (2017). Peer victimization trajectories from kindergarten through high school: Differential pathways for children's school engagement and achievement? *Journal of Educational Psychology,* 109(6), 826-841.

Laukaityte, I., & Wiberg, M. (2018). Importance of sampling weights in multilevel modeling of international large-scale assessment data. *Communications in Statistics-Theory and Methods,* 47(20), 4991-5012.

Lezotte, L. (2001). *Revolutionary and Evolutionary: The Effective Schools Movement*. United States: Effective Schools Products.

Liou, P.Y., & Liu, E.Z. (2015). An analysis of the relationships between Taiwanese eighth and fourth graders' motivational beliefs and science achievement in TIMSS 2011. *Asia Pacific Education Review,* 16(3), 433-445.

Lorah, J. (2019). Estimating a multilevel model with complex survey data: Demonstration using TIMSS. *Journal of Modern Applied Statistical Methods*, 18(2), eP3155.

Martin, M.O., & Mullis, I.V.S. (Eds.), (2012). *Methods and Procedures in TIMSS and PIRLS 2011*. Boston: TIMSS and PIRLS International Study Center. Available from: https://timssandpirls.bc.edu/methods [Last accessed on 2022 Jan 13].

Mao, P., Cai, Z., He, J., Chen, X., & Fan, X. (2021). The relationship between attitude toward science and academic achievement in science: A three-level meta-analysis. *Frontiers in Psychology,* 12, 784068.

Mullis, I.V., Martin, M.O., Foy, P., & Arora, A. (2012). *TIMSS 2011 International Results in Mathematics.* Boston: TIMSS & PIRLS International Study Center, Boston College.

Mullis, I.V.S., Martin, M.O., Foy, P., Kelly, D.L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science.* Available from: https://timssandpirls.bc.edu/timss2019/international-results [Last accessed on 2022 Jan 21].

Mullis, I., & Martin, M. (2012). Using TIMSS and PIRLS to improve teaching and learning. *Recherches en Education [Educational Research],* 14, 5835.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. United States: The National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: For States, By States*. United States: The National Academies Press.

Neuschmidt, O., Barth, J., & Hastedt, D. (2008). Trends in gender differences in mathematics and science (TIMSS 1995-2003). *Studies in Educational Evaluation,* 34(2), 56-72.

O'Dwyer, L.M. (2005). Examining the variability of mathematics performance and its correlates using data from TIMSS '95 and TIMSS '99. *Educational Research and Evaluation*, 11(2), 155-177.

Pace, A., Alper, R., Burchinal, M., Golinkoff, R., & Hirsh-Pasek, K. (2018). Measuring success: Within and cross-domain predictors of academic and social trajectories in elementary school. *Early Childhood Research Quarterly,* 46(1), 112-125.

Rennie, L.J., & Punch, K.F. (1991). The relationship between affect and achievement in science. *Journal of Research in Science Teaching*, 28(2), 193-209.

Rogers, A.M., & Stoeckel, J.J. (2008). *NAEP 2008 Arts: Music and Visual Arts Restricted-Use Data Files Data Companion (NCES no. 2011470)*. Washington, D.C.: National Center for Education Statics.

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large scale assessment data: Using a validity framework. *Large-Scale Assessments in Education,* 4(1), 1-18.

Salmi, H., Thuneberg, H., & Vainikainen, M. P. (2016). Do engineering attitudes vary by gender and motivation? Attractiveness of outreach science exhibitions in four countries. *European Journal of Engineering Education,* 41(6), 638-659.

Snijders, T.A.B., & Bosker, R.J. (2012). *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. United States: Sage Publishing.

Teodorović, J. (2012). Student background factors influencing student achievement in Serbia. *Educational Studies,* 38(1), 89-110.

Wang, C.L., & Liou, P.Y. (2017). Students' motivational beliefs in science learning, school motivational contexts, and science achievement in Taiwan. *International Journal Science Education,* 39(7), 898-917.

Wagemaker, H. (2020). *Reliability and Validity of International Large-Scale Assessment: Understanding IEA's Comparative Studies of Student Achievement.* United Kingdom: Springer Open.

Whitesell, N.R., Mitchell, C.M., & Spicer, P. (2009). A longitudinal study of self-esteem, cultural identity, and academic success among American Indian adolescents. *Cultural Diversity and Ethnic Minority Psychology,* 15(1), 38-50.

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

Zheng, A., Tucker-Drob, E.M., & Briley, D.A. (2019). National gross domestic product, science interest, and science achievement: A direct replication and extension of the Tucker-Drob, Cheung, and Briley (2014) study. *Psychological Science,* 30(5), 776-788.

Zheng, J. (2021). A functional review of research on clarity, immediacy, and credibility of teachers and their impacts on motivation and engagement of students. *Frontier Psychology*, 12, 712419.