The Effect of the Language of Instruction on Primary Students' Performance
Evidence From Gulf States

Donia Smaali Bouhlila and Imen Hentati

We explore the impact of the language of instruction on academic performance of Arabic-speaking students who are expected to learn mathematics and science in and through a second language (L2). Using as a theoretical background Cummins's (1984) framework relating language proficiency to academic achievement, we provide some insights into the relationship between native language, second language proficiency, and academic achievement in mathematics and science among students whose "mother tongue" is Arabic. We employ a propensity score technique and use TIMSS 2019 standardized tests to explain the differences in achievement in mathematics and science between two groups of young children living in an Arabic dialect–dominated environment who get instruction in English and who receive instruction in classical Arabic. Our findings highlight that the language of instruction accounts for the variations in performance.

Keywords: TIMSS, native language, second language, students' performance, propensity score

INTRODUCTION

In Arab countries, research on factors influencing performance of students has mostly focused on socioeconomic status, school resources, teacher performance, and community characteristics (Heyneman, 1997;

Donia Smaali Bouhlila holds a Ph.D in Development Economics She is an associate professor at the Faculty of Economics and Management of Tunis-University of Tunis El Manar. She is an associate editor at the *International Journal of Educational Development*. She is a research associate at the Economic Research Forum (ERF) and a senior researcher at *Laboratoire Prospectives, Stratégies et Développement Durable* (PS2D).
Imen Hentati holds a Ph.D in Family Economics. She is an assistant professor at the Faculty of Economics and Management of Tunis-University of Tunis El Manar. She is a researcher at the Laboratory for Research on Quantitative Development Economics (LAREQUAD).

Bouhlila, D.S., & Hentati, I. (2023). The Effect of the Language of Instruction on Primary Students' Performance. *Journal of Education in Muslim Societies, 4*(2), 53–80. DOI 10.2979/jems.4.2.04 •

Chapman and Miric, 2009; Bouhlila, 2011; Salehi-Isfahani et al., 2014; Bouhlila, 2015; Bouhlila, 2017). Despite strong government support for the education sector,[1] the quality of education as measured by student performance in international evaluations such as Trends in International Mathematics and Science Study (TIMSS) remains below international standards.

The language capital is one aspect that may have an influence on student performance and deserves special attention in Arab countries (Bouhlila, 2011). Language capital, or more simply "the mother tongue," is defined in the literature as the set of skills that are acquired during childhood with no particular effort and strengthened in school (Chiswick and Miller, 1995; Chiswick, 1991). The context of Arab countries is specific, as there is a difference between the classical Arabic (known as *fusha*a) and the different dialects spoken at home (the mother tongue). While strongly rooted in standard Arabic, colloquial dialects vary greatly in pronunciation, vocabulary, and grammar (Theodoropoulou & Tyler, 2014).

Because classical Arabic differs from the mother tongue, learning classical Arabic does not appear to be an easy process for students who speak the dialect. Despite the modest linguistic gap (between the dialect and classical Arabic), these students are unlikely to be fluent in Arabic, since they live in a "dialect-dominated" environment. This, in turn, will severely limit efficiency in language acquisition (Boutieri, 2012). Furthermore, it has been demonstrated that first-language acquisition takes at least 12 years from birth (Collier, 1989). This process is not completed for Arabic-speaking students by the time they reach this age, and it may take them even longer to acquire the fundamentals of the language. In addition, the majority of Arab countries are bilingual, which adds greater difficulties for students' academic achievement.

Bilingual education in many Arab countries has emerged from the colonial era (Zakharia, 2016). The spread of the language of the colonial power was mainly to serve its political and economic interests in the colonized country (Shaaban and Ghaith, 1999). With the partitioning of the Ottoman Empire into British and French mandates and protectorates following World War I, the French, Italian, and Spanish languages gained prominence in the Maghreb region, while in the Mashreq, French, and English spread in line with the language of the colonial power (Zakharia, 2016). The Western interests, mainly American and English, in the Middle East following the discovery of oil there have strengthened the English-language dominance (Karmani, 2005). Following the oil crises of the

1970s, there was an acceleration in the teaching of English as an international language (Karmani, 2005) in the Gulf States and an integration of the teaching of English into the states' educational development plans (Brewer and Goldman, 2010).

Bilingual education has given rise to an extensive amount of research that has examined the links between bilingualism and academic performance. In his theoretical framework, Cummins's (1984) asserts that students who are fluent in their mother tongue perform better academically. Additionally, he posits that there is an interdependence between the first language (L1), or the mother tongue, and the second language (L2), which also influences academic achievement (Cummins, 1978). According to the linguistic interdependence hypothesis, the development of L2 is partially reliant on the level of development of L1, and pupils with low levels of L1 and L2 are more likely to face academic difficulties in school (Cummins, 1984).

Using the Gulf States[2] as a case study, we want to explain the disparity in academic achievements of two groups of young students living in a dialect-dominated environment. The first is taught in English, while the second is taught in classical Arabic. More precisely and based on Cummins's theoretical background, we seek to provide (a) evidence of the language interdependence in the context of Gulf countries and (b) its link with academic performance as measured by TIMSS 2019 standardized tests, which evaluate students' performance in mathematics and science.

TIMSS tests are extremely interesting measures of language proficiency because they test the student's ability to use the language in different contexts. The proposed items in TIMSS refer to students' ability to apply their knowledge in various contexts, as well as analyze, reason, and communicate when they state, solve, and interpret problems in different contexts. The goal is not to have the student replicate what she or he has learned at school. Hence, the tests provide a measure of students' abilities to think in the language and to demonstrate content knowledge and cognitive knowledge. As a consequence, our research allows us to not only shed light on the role of language as an explanatory factor for student achievement, but also to consider future language policy in multilingual Arab states.

Our work adds to the empirical literature associating academic performance with school language. It is conducted with three goals in mind. The first is to expand research on variables influencing academic achievement in the Gulf States. The second is to investigate the linguistic interdependence

hypothesis in the context of Gulf countries. The third is to consider future language policy in multilingual Arab countries.

CONTEXT OF THE RESEARCH STUDY

The Gulf States are renowned for having diverse educational systems that educate in languages other than Arabic. Colonization and "oil" discovery provided a fertile environment for the expansion of English, which disproportionately served the economic interests of the English-speaking nations of the West (Karmani 2005; Brewer & Goldman, 2010; Sabic-El-Rayess, 2020). The tremendous industrial expansion that followed the discovery of oil resulted in bilingual education in the region (Brewer & Goldman, 2010), a consequence of the increase in the number of foreign workers (Winckler, 2010) resulting in a high demand for foreign schools offering international curricula (GCC Education, 2020). For many years, public schools in the Gulf region have had the lion's share of enrollment (GCC Education 2020, p. 13). In Oman, for instance, total primary school enrollment in 2012 was 248,859 students, compared to 50,612 students enrolled in private schools (GCC Education 2020, p. 64) in Bahrain, primary and secondary enrollments were around 72 percent in 2012, compared to 28 percent in private schools (GCC Education 2020, p. 32). However, attitudes have shifted in favor of private schools (GCC Education, 2020) which provide foreign language instruction (Zakharia, 2016). Public schools, however, teach all subjects in Arabic and introduce the English language as a subject from the primary grades (Zakharia, 2016). The rapid expansion of English in the Gulf region, as well as its penetration not just in formal education but also in almost every major public and private institution, has pushed for more incorporation of English and the privatization of schools (Asmi, 2013;[3] Brewer & Goldman, 2010).

Local nationals mainly from high socioeconomic status prefer English instruction for their children because they believe it will help them get global exposure (Zakharia, 2016). Arabic is only regarded as the language of the Qur'an and the heritage language (Hamidaddin, 2008). For students, English is viewed as more valuable in scientific and business disciplines than Arabic (Zakharia, 2009). Furthermore, a big number of local students want to study overseas to obtain better jobs and governments frequently provide scholarships to local students, encouraging them to seek higher education abroad (GCC Education, 2020).

The use of English and/or a second language as a medium of instruction has long been debated in the literature. In terms of job opportunities, a second language is seen to improve students' likelihood of securing employment and help in job mobility (Belhiah & Elhami, 2015; Chiswick, 1991; Angrist & Lavy, 1997).

Because investments in education pay off in the form of job opportunities and higher future incomes, most parents think that learning English as a second language and developing the necessary communication skills should commence as early as possible (Djiwandono, 2005; Tavil, 2009; Ching-Ying & Hsiang-Chun, 2016). In line with the adage "the earlier the better," and as English's relevance in the global economy has grown, policymakers throughout the world are also demanding earlier beginning ages for English language learning in schools (Enever, 2012; European Commission, 2012; Sayer, 2018; Song, 2018). The assumption that "the earlier the better" is inextricably linked to achieving linguistic competency in terms of communicating, reading, and writing skills, in other words to ensure native-like proficiency of L2 (Muñoz, 2014a; Muñoz, 2014b; Butler, 2015; De Wilde et al., 2021). However, research has demonstrated that second language mastering and its relationship to academic achievement is a complicated process that extends much beyond mere verbal communication abilities. This topic is covered in the next section.

THEORETICAL BACKGROUND

We refer to Cummins's (1984) conceptual and theoretical framework in our study, which distinguishes between the conversational dimension of English language proficiency known as basic interpersonal communicative skills (BICS) and the academic dimension of English language proficiency known as cognitive academic language proficiency (CALP). Cummins (2008, p. 71) defines them as follows:

> BICS refers to conversational fluency in a language while CALP refers to students' ability to understand and express, in both oral and written modes, concepts and ideas that are relevant to success in school.

BICS is cognitively undemanding, and nonspecialized. It takes the learner from six months to two years to develop BICS. CALP focuses on proficiency in academic language or language used in the school in the various content areas. The language needed for school is complex, context specific, and specialized. It includes not only all the domains of language (phonetics, syntax, vocabulary, discourse, etc.) but also the four language

skills of listening, speaking, reading, and writing (Cummins, 2008) to be mastered in each domain as well as the language to be mastered for each subject area (mathematics, science, economics, etc.). In addition to acquiring the language, learners need to develop skills such as analyzing, creating, applying, comparing, classifying, synthesizing, evaluating, and inferring when developing academic competence.

Empirical research has shown that it takes significantly longer for students receiving second language instruction to develop age-appropriate academic skills in L2 than it does to develop some aspects of age-appropriate English face-to-face communication skills (Collier, 1989).

Disentangling conversational skills from academic skills in a second language makes it possible to understand the academic difficulties faced by L2 learners. The distinction that was made between CALP and BICS is elaborated into a theoretical framework for relating language proficiency to academic achievement among bilingual students. The theoretical framework emphasizes the interrelationships between academic performance and language proficiency in both LI and L2, known as the linguistic interdependence hypothesis (Cummins, 1984). This hypothesis, when combined with a second hypothesis known as the threshold hypothesis, implies that a bilingual child must acquire certain levels of language competence (Cummins, 1979). More precisely, a lack of continuing L1 cognitive development during second language learning may result in lower levels of second language competency and academic retardation.

Collier (1989) analyzed different research findings on academic achievement in a second language. The findings give support for the linguistic interdependence and the need for maintaining cognitive development in L1 for young children with little or no L1 education. Other research studies corroborate Cummins's (1979) hypothesis. For instance, Cuevas (1984) argued that language fluency positively influences mathematics achievement. He argued that students must be proficient in both L1 and L2 in order to cope with the variety of language activities necessary for mathematics. Papanastasiou (2000) highlighted the fact that students who took TIMSS evaluations in a second language are the most disadvantaged. In the same vein, Herbert et al. (2002) showed that students who received their instruction in a second language (English instead of Chinese) experienced lower achievements than native students. Brock-Utne (2007) stressed that English as a medium of instruction in Tanzania likely slows

down the learning process. According to Samuelson and Freedman (2010), using English as the only medium of instruction would not necessarily enable students to engage successfully in the global economy, since many of them will lack a strong command of academic literacy in either their native language or English.

In sum, the expansion of English as a medium of instruction in the Gulf States is not solely tied to colonization but is also contingent to a large degree on serving the American and English interests. From a theoretical perspective, teaching in a second language has benefited from the theoretical framework developed by Cummins (1979). His central thesis posits that adequately developed LI skills can lead to cognitively and academically beneficial bilingualism. In this paper, we will investigate the linguistic interdependence in the context of Gulf countries by comparing the academic performance of two groups of students whose mother tongue is "Arabi" (Dialect). One group is receiving instruction in English, while the other is receiving instruction in Arabic.

TIMSS DATA AND METHODS

TIMSS Data

TIMSS assessments are intended to offer valid measurement of mathematics and science content and skills that are valued by the international education community and are incorporated in participating countries' curriculum. TIMSS was first administered in 45 countries in 1994/1995, with only Kuwait of Arab countries taking part. However, the number of Arab countries participating in TIMSS evaluations, whether at the fourth or eighth grade levels, has increased since then. The most recent TIMSS survey was conducted in 2019. It had 64 countries and 8 benchmarking participants, with 9 Arab countries participating at the fourth grade and 11 at the eighth grade. Testing was carried out at the end of the school year to guarantee cross-country comparability.

TIMSS reveals two realities: underachievement of Arab countries across all TIMSS cycles (Bouhlila, 2011) and underachievement of Arab students taking the tests in English[4] in comparison to native English speaking students. While the international average for both evaluations is set at 500, students in the Gulf States (Bahrain, Oman, Qatar, and the United Arab Emirates) who took the tests in English lag far behind the international average, as shown in Table A1.

Sample Design and Exclusion

We use data from TIMSS 2019 of students in the fourth grade. The samples are designed and conducted so that they provide reliable estimates about their representative population. TIMSS used some exclusion criteria at the school and student levels over its various cycles (Martin et al., 2020). Schools are excluded if they are located in remote areas, are extremely small (e.g., four or fewer students in the target grade), offer a curriculum that is substantially different from the mainstream educational system, or solely educate children with special needs. The international within-school exclusion rules are as follows: students with functional disabilities, students with intellectual disabilities and students who are nonnative language speakers. For our sample, the overall exclusion rate did not exceed 5.6% (Table A2).

Our research focuses on fourth-grade students[5] who took TIMSS assessments in Arabic and English. Four Gulf States (Bahrain, Oman, Qatar, and the UAE) provided the tests in both languages. The reason for selecting fourth-grade students is that at this grade the predominant use of language is cognitively demanded and context-reduced. This means that students have to rely primarily on linguistic cues to meaning and may in some cases require suspending knowledge of the "real world" in order to appropriately interpret the logic of the communication (Cummins & Swain, 1986). Since our aim is to test the linguistic interdependence hypothesis and its link to students' performance, we define two groups of students: those who took the tests in Arabic known as the control group and those who took the tests in English known as the treatment group. In order to get more accurate results and knowing that only native speakers are allowed to participate in TIMSS (according to the students' exclusion criteria discussed above), we restricted the participants who took the test in English to those whose both parents were born in the country. By doing so, we excluded students who were born from mixed marriages (having a native parent), and speak Arabic and English at home.

With 4,047 students, the UAE has the largest sample size, followed by Oman with 3,592 students and Bahrain with 2,787 students, while Qatar has the smallest sample size with 1,476 students. TIMSS tests in Arabic and English were completed by 9,790 students and 2,112 students, respectively.

EMPIRICAL MODEL AND TECHNIQUE

Methods

OLS regression is first used to estimate the average treatment effect of the use of English as a medium of instruction. The variable treatment is introduced in Equation (1) as a dummy variable. Recall that the treated group consists of students receiving English instruction, while the control group consists of students receiving Arabic instruction.

$$P_{i,c,s}=\alpha_0+\alpha_1 F_{i,c,s}+\alpha_2 T_{i,c,s}+\varepsilon_{i,c,s} \qquad (1)$$

Where P_{ics} is the score of student i in class c at school s. F_{ics} is a vector of individual and family background characteristics. T_{ics} is the treatment variable, which is a binary variable that takes the value 1 for treated observations and 0 for control observations. To draw valid inferences, we use the students' sampling weights.

The aim of OLS regression is to examine if there are any significant differences in performance between the treatment and control groups.

To gain a better understanding of performance differences, and because it is not straightforward to directly compare the outcomes for these two groups because those who choose English education may differ from those who choose Arabic, we use the propensity score matching technique. The propensity score matching (PSM) is a quasi-experimental method in which the researcher creates an artificial control group by matching each treated unit with a nontreated unit with similar characteristics. PSM, in particular, computes the probability of a unit enrolling in a program based on observed characteristics. This is the propensity score. Then, based on the propensity score, PSM assigns treated units to untreated units. PSM is based on the assumption that untreated units can be compared to treated units based on some observable characteristics, as if the treatment had been fully randomized (Rubin, 2001).

The use of the matching technique is to control the potential confounding influence of pretreatment variables (individual and family variables). We utilize a logit model to predict children's propensity for the treatment group. Following that, we use the Nearest Neighbor Matching to pair cases in both groups based on their likelihood of experiencing a treatment. We use matching with replacement to identify neighbor cases (Frisco et al., 2007).

Variables

This section describes the outcome variables used for the purpose of this study as well as the covariates.

Outcome variables. TIMSS standardized tests are interesting measures of language proficiency because they assess the student's ability to think in the language and to use the language in order to demonstrate content knowledge and cognitive knowledge in mathematics and science rather than replicating what was done in class.

The dependent variables are overall performance in mathematics and science as well as cognitive and content performance in mathematics and science. In sum, we have six dependent variables: Overall achievement in mathematics and science, achievement in mathematics and science content domains, and achievement in mathematics and science cognitive domains.

Number, measurement and geometry, and data are the three content areas of mathematics. Expressions, basic equations, and relationships, as well as fractions and decimals, are all part of the number content domain. The computation of whole numbers was required in the number domain. Pre-algebra principles included the concept of variables (unknowns) in basic equations, as well as preliminary understandings of quantity connections. Using a ruler to measure length, calculating areas and perimeters of basic polygons, and calculating volumes using cubes, as well as determining the qualities and characteristics of lines, angles, and a variety of two- and three-dimensional objects, were all part of the measurement process. Geometry entailed explaining and drawing a wide range of geometric forms, as well as exploiting geometric relationships to solve issues (Table 1).

The data content domain was divided into two sections: reading and interpreting and representing data and using data to solve problems (Martin et al., 2020). Life science, physical science, and Earth science are the examples of the science content domains. Topics in life science include organism features and life processes, life cycles, reproduction, and heredity, organisms, environment, and their interactions, ecosystems, and human health. Students were required to understand general features of organisms, how they work, and how they interact with other species and their environment, as well as basic scientific topics such as life cycles, heredity, and human health (Table 1).

The physical science content domain covered topics such as matter categorization and properties, as well as changes in matter; energy forms

Table 1. Content Performance

Subject	Areas	Percentage	Sections
Mathematics	Number	50%	- Expressions - Basic equations - Relationships - Fractions and decimals
	Measurement and Geometry	30%	For Measurement: - Using a ruler to measure length. - Calculating areas and perimeters of simple polygons - Using cubes to determine volumes - Identifying the properties and characteristics of lines, angles, and a variety of two- and three-dimensional shapes For Geometry: - Describing and drawing a variety of geometric figures - Using geometric relationships to solve problems
	Data	20%	- Reading, interpreting, and representing data - Using data to solve problems
Science	Life Science	45%	- Characteristics and life processes of organisms - Life cycles, reproduction, and heredity - Organisms, environment, and their interactions; ecosystems. - Human health
	Physical Science	35%	- Classification and properties of matter and changes in matter - Forms of energy and energy transfer - Forces and motion
	Earth Science	20%	- Earth's physical characteristics, resources, and history - Earth's weather and climates - Earth in the Solar System

Note. The percentage represents how much the specific area contributes to content performance. Source: Mullis et al. (2020).

and energy transmission; and forces and motion. Students were quizzed on physical states of matter (solid, liquid, and gas), as well as typical changes in the state and shape of matter; common forms and sources of energy and their practical applications; and fundamental concepts of light, sound, electricity, magnetism, and forces and motion. The Earth science domain includes the physical properties, resources, and history of the Earth, as well as the weather and climates of the Earth and its location in the Solar System. Students were required to explain the structure and physical attributes of the Earth's surface, as well as the utilization of the planet's most essential resources. They were asked to define some of Earth's processes in terms of visible changes, as well as identify the time span over which such changes happened. They were also questioned about Earth's position in the Solar System based on observations of changing patterns on Earth and in the sky (Table 1). Moreover, students were required to use a variety of cognitive abilities throughout the above-mentioned topic categories.

The cognitive abilities were divided into three general categories: knowing, applying, and reasoning (Table 2). The knowing domain focuses on the facts, concepts, and processes that students must grasp, whereas the applying domain focuses on students' capacity to apply information and conceptual understanding to solve problems or answer questions. The reasoning domain extends beyond the solution of basic issues learned in math or science classes to include novel scenarios, complicated settings, and multistep problems (Martin et al., 2020).

Table 2. Cognitive Performance

Subjects	Areas	Percentage	Sections
Mathematics and Science	Knowing	40%	- The facts, concepts, and procedures students need to know
	Applying	40%	- The ability of students to apply their knowledge and conceptual understanding to solve practical problems or provide answers
	Reasoning	20%	- The solution of familiar problems to encompass unfamiliar situations, complex contexts, and multistep problems

Note. The percentage represents how much the specific area contributes to cognitive performance. Source: Mullis et al. (2020).

In TIMSS, student achievement is represented by sets of five plausible values (Martin et al., 2016; Mislevy, 1991). Plausible values are imputed values

drawn from the estimated ability distributions. Plausible values are generated by making use of all available background data of the students. Plausible values are not intended to be estimates of individual student scores, but rather are imputed scores for like students—students with similar response patterns and background characteristics in the sampled population—that may be used to estimate population characteristics correctly (Martin et al., 2020). A detailed review of the plausible values methodology is given in Mislevy (1991).

Covariates. The core explanatory variables are individual and family background characteristics. These variables and their codings are listed in Table A3. Controlling for socioeconomic status (SES) and pre-education is important as they may influence the rate at which a second language is learned. Individual factors include the students' age, gender, and the number of years in pre-primary education. Family factors include parents' education and educational resources at home. The latter was constructed based on the data reported by students and their parents regarding the number of books and other study materials in their homes, the parents' levels of education (National Academies of Sciences, Engineering, and Medicine, 1997), and the parents' employment. Cut scores were used to define students into three categories: students with many resources, students with some resources, and students with few resources. All the nominal variables were introduced in the model as dummies. In terms of gender, females are assigned as 1, and males as 0. Parent's education levels are presented in four categories: some primary, lower secondary, upper secondary, post secondary and university or higher. The category some primary is considered as a reference category. For home resources, students are assigned to three categories, which are the following: students with many resources, students with some resources, and students with few resources. The category few resources is considered as a reference category.

Descriptive Statistics

This section presents the descriptive statistics of the outcome variables and the covariates used. Table 3 displays descriptive statistics for students.

The mean of the outcome variables in both the treated and control groups is less than 500 points, which is the international average. The average age of the tested students is 9.8 years, and the sample is roughly divided into boys and girls. Students typically received two years of pre-primary education. Furthermore, 50 percent of parents have a university degree or higher and more than 80 percent of students have access to some home educational resources.[6]

Table 3. Descriptive Statistics

Variables	Mean of Total Sample	Mean of Control Group (Arabic)	Mean of Treatment Group (English)	Effect Size (Cohen's d)
Math1	460.913	461.989	455.923	-0.0653
Math2	460.249	461.301	455.369	-0.0641
Math3	461.279	462.144	457.266	-0.0531
Math4	460.247	461.223	455.724	-0.0594
Math5	460.143	461.481	453.945	-0.0811
Science1	467.79	473.490	441.364	-0.306
Science 2	466.356	472.555	437.623	-0.332
Science 3	466.1598	471.981	439.174	-0.313
Science 4	465.37	471.388	437.474	-0.319
Science 5	467.4827	473.472	439.717	-0.320
Content_ Math1	458.665	459.631	454.188	-0.0576
Content_ Math 2	457.2484	458.074	453.418	-0.0488
Content_ Math 3	458.873	460.070	453.325	-0.0715
Content_ Math 4	457.8296	459.095	451.962	-0.0751
Content_ Math 5	457.447	458.496	452.582	-0.0623
Cognitive_ Math1	458.9122	459.8861	454.397	-0.0580
Cognitive_ Math 2	458.765	459.851	453.728	-0.0650
Cognitive_ Math 3	459.625	460.249	456.731	-0.0376
Cognitive_ Math 4	457.961	458.922	453.506	-0.0574
Cognitive_ Math 5	458.229	459.363	452.968	-0.0679
Content_ Science1	466.585	472.679	438.333	-0.321
Content _Science 2	464.148	470.346	435.420	-0.324
Content _ Science 3	464.439	470.272	437.400	-0.304
Content _Science 4	463.175	469.103	435.700	-0.310
Content _ Science 5	466.731	472.44	440.267	-0.299
Cognitive _ Science1	478.573	491.080	438.510	-0.518
Cognitive _Science 2	477.419	489.250	439.518	-0.488
Cognitive _ Science 3	477.617	489.885	438.318	-0.505
Cognitive _ Science 4	476.566	488.902	437.047	-0.503
Cognitive_ Science 5	478.568	490.706	439.683	-0.498
Age	9.693	9.689	9.709	0.0473
Sex	0.534	0.539	0.508	-0.0623
Pre-Education	1.723	1.700	1.828	0.109
Parents Education				
Some primary	0.0409	0.0451	0.0214	-0.135
Lower secondary	0.0461	0.0529	0.0146	-0.222

(*Continued*)

Table 3. Continued

Upper secondary	0.287	0.327	0.104	-0.578
Post-secondary	0.157	0.159	0.142	-0.0467
University or higher	0.469	0.416	0.718	0.640
Resources				
Few resources	0.0467	0.055	0.01	-0.273
Some resources	0.881	0.888	0.847	-0.123
Many resources	0.0723	0.057	0.143	0.296
Number of Observations				
Total sample	11,902			
Control group	9,790			
Treatment group	2,112			

Note. Data on cognitive performance is not available for Oman. $Math_i$ $Science_i$ (i=1 . . . 5) represent overall performance in mathematics and science respectively.

Table 3 reveals that students who took the tests in Arabic outperformed those who took the tests in English. The average effect size in mathematics overall evaluation as well as content and cognitive performances is negative. Though the effect size is small in mathematics and medium in science (for more detailed explanations for the range of variation of *d,* see Borenstein, 2009; Hattie, 2009), the negative sign of *d* indicates that learning in English is less effective than learning in Arabic.

RESULTS AND DISCUSSION

This section discusses first the OLS results (Tables 4 and 5), then propensity score matching results (Table 6). OLS regression was conducted separately five times for each specific performance (overall performance, content performance, and cognitive performance) and for each evaluation. Rubin's rules (Rubin, 1987) were used to produce the final parameter of interest. TIMSS, like any other survey data, has missing data. This issue arises when students and school principals fail to complete certain questionnaire items. Missing values in Stata are handled by default using "listwise deletion" which means that Stata will remove any observation that is missing on the outcome variable or any of the predictor variables. Because missing data in the students' background variables were generally low, "listwise deletion" was used to exclude missing data (Tabachnik & Fidell, 2007).

In this study, we use a version of Rubin's Rules to obtain the final estimates drawn from the five plausible values. The details of the procedure are shown in Figure 1.

Step 1: Estimate the statistic/model of interest five times, once using each of the plausible values. This will generate five separate parameter estimates (β_{pv}) and five estimates of the sampling error (σ_{pv}).

Step 2: To produce the final parameter and sampling error estimates, one simply takes the average of the five estimates produced in step 1: $\beta_* = \frac{\sum_{pv=1}^{5} \beta_{pv}}{n_{pv}}$ and. $\sigma_* = \frac{\sum_{pv=1}^{5} \sigma_{pv}}{n_{pv}}$

Where:

β_*= Final estimate of the parameter of interest

σ_* = Final estimate of the sampling error

n_{pv} = The number of plausible values (five)

Step 3: Estimate the magnitude of the imputation error, based upon the following formula:

$$\delta_* = \frac{\sum_{pv=1}^{5} (\beta_{pv} - \beta^*)^2}{n_{pv} - 1}$$

Where:

δ_* = The magnitude of the imputation error.

Step 4: Calculate the value of the final standard error by combining the sampling error (σ_*) and the imputation error (δ_*) via the following formula:

Standard error = $\sqrt{\sigma_*^2 + (1 + \frac{1}{n_{pv}}).\delta_*^2}$

One can then use the final parameter estimate (β_*) and its standard error to conduct hypothesis tests and construct confidence intervals following the usual methods.

Figure 1. Rubin's Rules

In both evaluations, family variables have a significant positive influence on overall, cognitive, and content performance. This result is consistent with the majority of earlier investigations (Ammermüller et al., 2005; Wömann, 2003, 2004; Chiu & Khoo, 2005). In terms of individual student characteristics, and regarding mathematics evaluation, the coefficient of sex is not significant. However, girls outperform boys in science. Preschool education participation has a positive and significant influence on overall performance, content performance, and cognitive performance, which is in line with previous studies (Holla et al., 2021).

Table 4. OLS Results for Science

Variables	**Overall Performance**	**Content Performance**	**Cognitive Performance**
TREAT	-53.878*** (2.798)	-54.423*** (2.811)	-63.503*** (2.959)
Age	19.668*** (2.605)	19.316*** (2.677)	21.955*** (2.808)
Sex	14.820*** (1.933)	15.995*** (2.0778)	11.863*** (2.381)
Pre-education	8.475*** (0.918)	8.403*** (0.891)	5.662*** (1.00548)
Parents' education			
Lower secondary	1.432 (7.394)	1.0742 (7.307)	-9.833 (8.887)
Upper secondary	24.989*** (6.795)	26.904*** (6.369)	11.456 (7.520)
Post-secondary	46.425*** (6.717)	49.196*** (6.921)	27.222*** (8.1435)
University	68.624*** (6.466)	71.569*** (6.346)	47.417*** (7.422)
Resources			
Some resources	32.881*** (6.832)	33.752*** (5.827)	20.379** (8.135)
Many resources	61.943*** (7.442)	63.856*** (6.732)	46.497*** (9.0715)
Constant	182.767*** (25.623)	180.938*** (26.432)	208.290*** (28.612)
R-Squared	0.11698	0.11698	0.11
Number of Observations	11,902	11,902	8,310

Notes. Standard errors are in parentheses. Significance levels: *** p<.01, ** p<.05, * p<.1. Data on cognitive performance is not available for Oman.

Table 5. OLS Results for Mathematics

Variables	**Overall Performance**	**Content Performance**	**Cognitive Performance**
TREAT	-22.0150*** (4.107)	-22.275*** (2.519)	-21.443*** (2.601)
Age	14.727*** (4.336)	14.924*** (2.321)	15.0190*** (2.294)
Sex	1.477 (3.406)	1.965 (1.808)	2.365 (1.808)
Pre-education	5.882*** (1.495)	6.0776*** (0.779)	6.528*** (0.763)
Parents' education			
Lower secondary	-4.364 (13.257)	-6.170 (6.879)	-4.848 (6.628)
Upper secondary	11.177 (11.809)	12.198* (6.0104)	12.101* (5.683)
Post-secondary	29.070** (12.143)	29.548*** (5.933)	29.00265*** (5.708)
University	46.298*** (11.725)	47.158*** (5.807)	46.557*** (5.542)
Resources			
Some resources	27.0560** (11.515)	28.311*** (5.371)	28.450*** (5.159)
Many resources	52.0494*** (12.7648)	54.701*** (6.340)	53.621*** (6.279)
Constant	254.0285*** (45.925)	247.0467*** (23.0469)	245.953*** (23.461)
R-Squared	0.07166	0.07194	0.07202
Number of Observations	11,902	11,902	11,902

Note. Standard errors are in parentheses, significance levels: *** p<.01, ** p<.05, * p<.1

Additionally, OLS results show that students taking the tests in English have significantly lower achievement in mathematics and science than their peers who passed the tests in Arabic. To provide more evidence for this finding, we estimate a propensity score model. The results are displayed in Table 6.

Table 6. ATET Nearest Neighbor Results

ATE TREAT (1 vs. 0)	Mathematics	Science
Overall Performance	-27.277*** (3.0246)	-60.981*** (3.509)
Cognitive Performance	-27.459*** (3.176)	-62.0945*** (3.269)
Content Performance	-26.782*** (3.464)	-72.470*** (3.752)

Note. Standard errors are in parentheses significance levels: *** p<.01, ** p<.05, * p<.1.

Students who received English-language mathematics instruction (Treatment group) had their overall, cognitive, and content scores reduced by 27.2; 27.4, and 26.7 points, respectively, when compared to students who passed the Arabic-language Mathematics test (Control group). Similarly, students who received science instruction in English (Treatment group) had their overall, cognitive and content scores fall by 60.9, 62.1, and 72.4 points, respectively, as compared to students in the control group. This suggests that performance at this grade level is lower when education is offered in English. All the differences are statistically significant at the 1% level.

Our findings reveal that students whose mother tongue is Arabic and with early English learning experience score significantly lower in TIMSS evaluations than their peers who take the tests in Arabic. The differences are more noticeable in science than in mathematics. The learning of mathematics and science requires a variety of linguistic skills that second language learners may not have mastered by the age of nine. Even though mathematics language seems to be abstract, the study of mathematics (especially at the primary level) begins with the study of real world problems and requires the application of the language. Therefore, the language plays an important role in conveying mathematical knowledge to students and in knowing how abstraction is interpreted (Ferrari, 2003). Likely, the difference in science proficiency between the treatment and control group is manifest. This provides evidence that a first language enhances more science learning than a second language. At grade 4, science context domains are more linked to real-world situations and thus to "home language." Students need to develop a deep understanding of science concepts, make connections among concepts, and apply concepts in explaining natural phenomena or real-world situations. Students in science classes must be engaged in science inquiry, have to negotiate ideas, and justify claims based on evidence. It has been shown that sometimes,

and for effective instruction in science, teachers focus on students' home language as an instructional support. They use students' home language to explain science terms (Goldenberg, 2013).

A key premise in the literature is that age is significant in child language acquisition, whether in L1 or L2 (Collier, 1989; Oliver and Azkarai, 2017). Given that the critical period of 12 years is required to be fully proficient in the first language (Collier, 1989), students in Arab countries at the age of nine (fourth grade) have not yet begun to complete full cognitive development in the first language and so do their peers in their native languages. Nonetheless, the difference is that English native speakers are approaching the completion of L1 acquisition, whereas students whose "*mother tongue*" is Arabic (Dialect) require additional years to improve their Arabic language skills. So they lag behind their peers in terms of years of L1 acquisition. The mismatch between the language of the home and the language of the school makes those students less proficient in "*modern Arabic*."

Having English instruction does not boost performance. Furthermore, in these standardized assessments, students assessed in English do not achieve native-speaker norms. The difference in results between native speakers and bilingual students is around 100 points, as indicated in Table A1. The linguistic mismatch, which results in low levels of L1 proficiency, along with bilingualism is identified in this research as a factor of academic and cognitive retardation (Table 6). This finding supports the interdependence hypothesis, which predicts that the development of L2 school language is partly reliant on the level of development of L1 school language (Cummins, 1978).

CONCLUSION

In this study, we explore the impact of the language of instruction (English in this case) on academic performance of Arabic speaking students who are expected to learn mathematics and science in and through a second language (L2) before they have become adequately proficient in that language. Having as a theoretical background Cummins's (1984) framework relating language proficiency to academic achievement, we provide some insights into the relationship between L1, L2 proficiency and academic achievement in mathematics and science among Arab students. We employ propensity score technique and used TIMSS 2019 standardized tests to explain the difference in cognitive and content

academic results in mathematics and science between two groups of young children living in Gulf States and living in a dialect-dominated environment who got instruction in English and who received instruction in Arabic. Our findings highlight that the language of instruction in the two groups accounts for the variations in performance. Children who began studying L2 at a young age without having fully developed L1 exhibited deficits in overall, content, and cognitive performance as compared to those who began their Arabic instruction at the same age. The finding of this research is consistent with Cummins's (1984) linguistic interdependence hypothesis. This study helps us to think about future language policy not only in the Gulf States, but also in Arab countries that are all multilingual. Strengthening early exposure to the Arabic language (modern Arabic) will minimize the retardation produced by the mismatch between students' mother tongue and the school language. Arabic should be taught and learned in schools in a more practical manner. Furthermore, audio-visual tools for learning Arabic should be provided, which have the advantage of not only developing fluency in the language but also developing cognitive and academic aspects of the language. Lastly, bilinguals with sufficient competency in one of their languages would experience no such academic disadvantages and students fully proficient in both languages would enjoy cognitive and academic advantages associated with bilingualism.

Notes

1. https://data.worldbank.org/indicator/SE.XPD.TOTL.GB.ZS?locations=ZQ.

2. Arab countries are Algeria, Bahrain, Comoros, Djibouti, Egypt, Iraq, Jordan, Kuwait, Lebanon, Libya, Mauritania, Morocco, Oman, Palestine, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tunisia, the United Arab Emirates, and Yemen. The Gulf States are a subset of Arab countries.

3. Reference cited in Zakharia (2016, pp. 1–13).

4. Most of the Arab countries administered TIMSS evaluations in Arabic (Mullis et al., 2020).

5. All students enrolled in the grade that represents four years of schooling counting from the first year of ISCED Level 1, providing the mean age at the time of testing is at least 9.5 years. It is worth noting that eighth-grade students were excluded from our analysis because they do not have data on pre-primary education that corresponds to the number of years in pre-primary education for each student. Excluding this variable may bias the results for eighth-grade students since pre-education provides opportunities for children to acquire certain knowledge about letters and language.

6. Resources are not correlated with parents' education since it is not a linear combination of the variables used to construct it. The variance inflation factor (VIF) of the model =4.84<10, so there is no problem of multicollinearity.

References

Ammermüller, A., Heijke, H., & Wößmann, L. (2005). Schooling quality in eastern Europe: educational production during transition. *Economics of Education Review, 24*(5), 579–599. http://www.sciencedirect.com/science/article/pii/S0272-7757(04)00124-4.

Angrist, J. D., & Lavy, V. (1997). The Effect of a change in language of instruction on the returns to schooling in Morocco. *Journal of Labor Economics*, *15*(1), S48–S76. https://doi.org/10.1086/209856.

Asmi, R. (2013). Language in the mirror: Arabic, Islam and schooling in Qatar. Unpublished doctoral dissertation, Columbia University, New York. https://doi.org/10.7916/D8W383JK.

Belhiah, H. & Elhami, M. (2015). English as a medium of instruction in the Gulf: When students and teachers speak. *Language Policy*, *14*, 3–23. 10.1007/s10993-014-9336-9.

Borenstein (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta analysis* (pp. 279–293). Russell Sage Foundation. Available at: https://www.daneshnamehicsa.ir/userfiles/files/1/9-%20The%20Handbook%20of%20Research%20Synthesis%20and%20Meta-Analysis.pdf.

Bouhlila, S. D. (2011). The quality of secondary education in the Middle East and North Africa: What can we learn from TIMSS' results? *A Journal of Comparative and International Education*, *41*(3), 327–352. https://doi.org/10.1080/03057925.2010.539887.

Bouhlila, S. D. (2015). The Heyneman-Loxley effect revisited in the context of MENA countries: Analysis using TIMSS 2007 database. *International Journal of Educational Development*, *42*, 85–95. 10.1016/j.ijedudev.2015.02.014.

Bouhlila, S. D. (2017). Parents' education and literacy skills: Evidence on inequality of socioeconomic status in Arab countries. *World Development Perspectives*, *5*, 34–43. 10.1016/j.wdp.2017.02.006.

Boutieri, C. (2012). In two speeds (A deux vitesses): Linguistic pluralism and educational anxiety in contemporary Morocco. *International Journal for Middle East Studies*, *44*(3), 443–464. https://doi.org/10.1017/S0020743812000414.

Brewer, D. J., & Goldman, C. A. (2010). An introduction to Qatar's primary and secondary education reform. In O. Abi-Mershed (Ed.), *Trajectories of education in the Arab World: Legacies and challenges* (pp. 226–246). Routledge. 10.4324/9780203873755.

Brock-Utne, B. (2007). Learning through a familiar language versus learning through a foreign language—a look into some secondary school classrooms in Tanzania. *International Journal of Educational Development*, *27*(5), 487–498. 10.1016/j.ijedudev.2006.10.004.

Butler, Y. (2015). English language education among young learners in East Asia: A review of current research (2004–2014). *Language Teaching, 48*(3), 303–342. https://doi.org/10.1017/S0261444815000105.

Chapman, D. W., & Miric, S. L. (2009). Education quality in the Middle East. *International Review of Education, 55,* 311–344. 10.1007/s11159-009-9132-5.

Ching-Ying, L., & Hsiang-Chun, C. (2016). Parental perceptions of early childhood English education. *International Journal on Studies in English Language and Literature (IJSELL), 4*(11), 62–70. http://dx.doi.org/10.20431/2347-3134.0411011.

Chiswick, B. R. (1991). Speaking, reading and earnings among low-skilled immigrants. *Journal of Labor Economics*, *9*(2),149–170. http://dx.doi.org/10.1086/298263.

Chiswick, B. R., and Miller, P. W. (1995). The endogeneity between language and earnings: International analysis. *Journal of Labor Economics*; *13*, 246–288. http://dx.doi.org/10.1086/298374.

Chiu, M. M., & Khoo, L. (2005). Effects of resources, inequality, and privilege bias on achievement: Country, school and student level analyses. *American Education Research Journal, 42*, 575603. https://doi.org/10.3102/00028312042004575.

Collier, V. P. (1989). How long? A synthesis of research on academic achievement in a second language. *TESOL Quarterly, 23,* 509–531. https://doi.org/10.2307/3586923.

Cuevas, G .J. (1984). Mathematics learning in English as a second language. *Journal of Research in Mathematics Education, 15*(2): 134–144. https://doi.org/10.2307/748889.

Cummins, J. (1978). Educational implications of mother tongue maintenance in minority-language groups. *The Canadian Modern Language Review*, *34*, 395-416. https://doi.org/10.3138/cmlr .34.3.395.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*(2), 222-251. https://doi.org/10.3102/00346543049002222.

Cummins, J. (1984). Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students. In C. Rivera (Ed.). *Language proficiency and academic achievement* (pp. 2–19). Multilingual Matters. http://www.eric.ed.gov/contentdelivery/servlet /ERICServlet?accno=ED240882.

Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. Street & N. H. Hornberger (Eds.). *Encyclopedia of language and education,* 2nd ed., Volume 2: *Literacy* (pp. 71–83). Springer. 10.1007/978-0-387-30424-3_36

Cummins, J., & Swain, M. (1986). *Bilingualism in education.* Longman. https://doi.org/10.4324 /9781315835877.

De Wilde, V., Brysbaert, M., & Eyckmans, J. (2021). Young learners' L2 English after the onset of instruction: longitudinal development of L2 proficiency and the role of individual differences. *Bilingualism: Language and Cognition, 24*, 439–453. https://doi.org/10.1017/S1366728920000747.

Djiwandono, P. I. (2005). Teach my children English : Why parents want English teaching. *Lang. Teach., 1*, 62–72. https://doi.org/10.25170/ijelt.v1i1.1407.

Enever, J. (2012). Current policy issues in early foreign language learning. *CEPS Journal, 2*(3), 9–26. https://doi.org/10.26529/cepsj.345.

Ferrari, P. L. (2003). Abstraction in mathematics. *Philosophical Transactions: Biological Sciences*, *358*(1435), 1225–1230. 0.1098/rstb.2003.1316.

Fishbein, B., Foy, P., & Yin, L. (2021). TIMSS 2019 user guide for the international database (2nd ed.). https://www.iea.nl/publications/user-guides/user-guide-international-database-timss-2019.

Frisco, M. L., Muller C,, & Frank, K. (2007). Family structure change and adolescents' school performance: A propensity score approach. *Journal of Marriage and Family; 69*, 721–741.

GCC Education. (2020). *Sector report GCC education.* https://gfh.com/wp-content/uploads/2020/05 /GFH-Education-Sector-Report-2020.pdf.

Goldenberg, C. (2013). Unlocking the research on English learners: What we know—and don't yet know—about effective instruction. *The American Educator, 37*(2), 4–11. https://eric .ed.gov/?id=EJ1014021.

Hamidaddin, H. A. (2008). *Important factors to consider for bilingual education in the UAE* (MA thesis). American University of Sharjah. https://dspace.aus.edu/xmlui/bitstream/handle/11073/54/29 .232-2008.07%20Huda%20Hamidaddin.pdf?sequence=1&isAllowed=y.

Hattie, John (2009): *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge. https://doi.org/10.4324/9780203887332.

Herbert, W. M., Hua, T. K., & Kong, K. C. (2002). Multilevel causal ordering of academic self-concept and achievement: Influence of language of instruction (English compared with Chinese) for Hong Kong students. *American Educational Research Journal, 39*, 727–763. https:// doi.org/10.3102/00028312039003727.

Heyneman, S. P. (1997). The quality of education in the Middle East and North Africa (MENA). *International Journal of Educational Development, 17*, 449–466. https://www.academia.edu/595378/Quality_of_education_in_the_Middle_East.

Holla, A., Bendini, M., Dinarte, L., & Trako, I. (2021). Is investment in pre-primary education too low? lessons from (quasi) experimental evidence across countries. *Policy Research Working Paper*, No. 9723. World Bank. http://hdl.handle.net/10986/35894.

Karmani, S. (2005). Petro-linguistics: The emerging nexus between oil, English, and Islam. *Journal of Language, Identity and Education*, *4*(2), 87–102. https://doi.org/10.1207/s15327701jlie0402_2.

Martin, M. O., Mullis, I. V. S., Foy, P. & Hooper, M. (2016). TIMSS achievement methodology. In *Methods and procedures in TIMSS 2015* (pp. 12.1–12.9). https://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Martin, M. O., von Davier, M., & Mullis, I. V. S. (Eds.). (2020). *Methods and procedures: TIMSS 2019 Technical Report.* https://timssandpirls.bc.edu/timss2019/methods/.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457.

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 international results in mathematics and science*. https://timssandpirls.bc.edu/timss2019/.

Muñoz, C. (2014a). Contrasting effects of starting age and input on the oral performance of foreign language learners. *Applied Linguistics, 35*(4), 463–482. https://doi.org/10.1093/applin/amu024.

Muñoz, C. (2014b). Starting age and other influential factors: Insights from learner interviews. *Studies in Second Language Learning and Teaching, 4*, 465–484. https://doi.org/10.14746/ssllt.2014.4.3.5.

National Academies of Sciences, Engineering, and Medicine. (1997.) *Improving schooling for language-minority children: A research agenda.* The National Academies Press. https://doi.org/10.17226/5286.

Oliver, R., & Azkarai, A. (2017). Review of child second language acquisition (SLA): Examining theories and research. *Annual Review of Applied Linguistics*, *37*, 62–76. https://doi.org/10.1017/S0267190517000058.

Papanastasiou, C. (2000). Internal and external factors affecting achievement in mathematics: Some findings from TIMSS. *Studies in Educational Evaluation, 26*, 1–7. https://doi.org/10.1016/S0191-491X(00)00002-X.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. http://dx.doi.org/10.1002/9780470316696.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, *2*, 169–188. 10.1023/A:1020363010465.

Sabic-El-Rayess, A. (2020). Epistemological shifts in knowledge and education in Islam: A new perspective on the emergence of radicalization amongst Muslims. *International Journal of Educational Development, 73,* 1–10. https://doi.org/10.1016/j.ijedudev.2019.102148.

Salehi-Isfahani, D., Hassine, N. B., & Assaad, R. (2014). Equality of opportunity in educational achievement in the Middle East and North Africa. *The Journal of Economic Inequality, 12*(4), 489–515. https://doi.org/10.1007/s10888-013-9263-6.

Samuelson, B. L., & Freedman, S. W. (2010). Language policy, multilingual education, and power in Rwanda. *Language Policy*, *9*(3), 191–215. 10.1007/S10993-010-9170-7.

Sayer, P. (2018). Does English really open doors? Social class and English teaching in public primary schools in Mexico. *System, 73,* 58–70. https://doi.org/10.1016/j.system.2017.11.006.

Shaaban, K., & Ghaith, G. (1999). Lebanon's language-in-education policies: From bilingualism to trilingualism. *Language Problems and Language Planning*, *23*(1), 1–16. 10.1075/lplp.23.1.01leb.

Song, J. (2018). English Just is not enough: Neoliberalism, class, and children's study abroad among Korean families. *System, 73,* 80–88 https://doi.org/10.1016/j.system.2017.10.007.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Allyn & Bacon. https://www.pearsonhighered.com/assets/preface/0/1/3/4/0134790545.pdf.

Theodoropoulou, I., & Tyler, J. (2014). Perceptual dialectology of the Arab world: A principal analysis. *Al-'Arabiyya, 47,* 21–39. : https://www.jstor.org/stable/24635371.

Winckler, O. (2010). Labor migration to the GCC states: Patterns, scale, and policies. Migration and the Gulf. *Middle East Institute Viewpoints*. https://www.mei.edu/publications/labor-migration-gcc-states-patterns-scale-and-policies.

Wößmann, L. (2003). European "education production functions": What makes a difference for student achievement in Europe? *Economic papers* no. 190. CESifo Munich. https://ec.europa.eu/economy_finance/publications/pages/publication862_en.pdf.

Wößmann, L. (2004). How equal are educational opportunities? Family background and student achievement in Europe and the United States. *IZA discussion paper* no. 1284. https://docs.iza.org/dp1284.pdf.

Zakharia, Z. (2009). Positioning Arabic in schools: Language policy, national identity, and development in contemporary Lebanon. In F. Vavrus & L. Bartlett (Eds.), *Critical approaches to comparative education: Vertical case studies from Africa, Europe, the Middle East, and the Americas* (pp. 215–231). Palgrave Macmillan. https://doi.org/10.1057/9780230101760_13.

Zakharia, Z. (2016). Bilingual education in the Middle East and North Africa. In O. García et al. (Eds.), *Bilingual and multilingual education, encyclopedia of language and education* (pp. 1–13). Encyclopedia of Language and Education. https://doi.org/10.1007/978-3-319-02324-3_21-1.

APPENDIX

Table A1. Students' Performance in TIMSS 2019 (English)

	Mathematics			**Science**		
International Average	500			500		
	Overall performance	Content	Cognitive	Overall performance	Content	Cognitive
TREATMENT Group						
The sample of the present study	460.886	460.034	476.69	441.873	440.16	441.045
Native Speakers						
England	559.151	559.129	560.610	539.912	539.285	539.150
USA	546.430	545.584	546.396	548.542	548.085	548.570

Table A2. Coverage and Exclusion Rates

Country	**Coverage**	**Overall exclusion**	**Schools**	**Students**
Bahrain	100%	0.8%	0.4%	0.4%
Oman	100%	2.2%	1.4%	0.8%
Qatar	100%	2.2%	1.2%	1%
Dubai	100%	5.6%	2.6%	3%

Source. Mullis et al. (2020).

Table A3. Dependent Variables List, Coding, and Meaning

List of Variables	Coding	Meaning
Overall Performance Math1-Math5	asmmat 01–05	The 1st to 5th plausible value of Overall Performance in Mathematics
Overall Performance Science1-Science5	asssci 01–05	The 1st to 5th plausible value of Overall Performance in Science
Content Performance Math1-Math5 (50%Number, 30% Measurement and Geometry and 20% Data)	0.5* asmnum 01–05 +0.3* asmgeo 01–05 +0.2*asmdat 01–05	The 1st to 5th plausible value of Content Performance in Mathematics
Cognitive Performance Math1-Math5 (40% Knowing, 40% Applying and 20% Reasoning)	0.4* asmkno 01–05 +0.4* asmapp 01–05 + 0.2* asmrea 01–05	The 1st to 5th plausible value of Cognitive Performance in Mathematics
Content Performance Science1-Science5 (45% Life Science, 35% Physical Science and 20% Earth Science)	0.45* asslif 01–05 +0.35* assphy 01–05 +0.2*assear 01–05	The 1st to 5th plausible value of Content Performance in Science
Cognitive Performance Science1-Science5 (40% Knowing, 40% Applying and 20% Reasoning)	0.4* asskno 01–05 +0.4* assapp 01–05 +0.2*assrea 01–05	The 1st to 5th plausible value of Cognitive Performance in Science

Note. For each subdomain, five plausible values are also derived. They are referred to in the table as 01–05. Coding of the sub domains are highlighted in column two. Source: Fishbein et al. (2021).

Table A4. Independent Variables' List, Coding, and Meaning

List of Variables	**Coding**	**Meaning**
Age	asdage	Quantitative variable which indicates student's age.
Sex	asbg01	Dummy variable which takes the value 1 for female and 0 for male.
Pre-Education	asbh04b	Quantitative variable which indicates the number of years in pre-primary education for each student.
Education	asdhedup	Categorical variable which reflects parents' education level as follows; some primary, upper secondary, post secondary and university or higher. The category some primary is considered as reference category.
Resources	asdghrl	A score calculated based on the number of books and other study materials in the students' homes, their parents' level of education, and their parents' employment. Scores were used to define students into three categories: students with Many Resources, students with Some Resources and students with Few Resources. The category few resources is considered as a reference category.

Source: Fishbein et al. (2021).