WHAT BIG DATA CAN TELL US ABOUT THE PSYCHOLOGY OF LEARNING AND TEACHING

EDITED BY: Ronnel B. King, Jiesi Guo and Ching Sing Chai
PUBLISHED IN: Frontiers in Psychology

frontiers Research Topics

frontiers

ISSN 1664-8714
ISBN 978-2-88974-632-3
DOI 10.3389/978-2-88974-632-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

WHAT BIG DATA CAN TELL US ABOUT THE PSYCHOLOGY OF LEARNING AND TEACHING

Topic Editors:
Ronnel B. King, University of Macau, China
Jiesi Guo, Australian Catholic University, Australia
Ching Sing Chai, The Chinese University of Hong Kong, China

Citation: King, R. B., Guo, J., Chai, C. S., eds. (2022). What Big Data Can Tell Us About the Psychology of Learning and Teaching. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88974-632-3

Table of Contents

frontiers in Psychology

ORIGINAL RESEARCH
published: 21 January 2020
doi: 10.3389/fpsyg.2019.03053

Cultural Variation in the Effectiveness of Feedback on Students' Mistakes

Kimmo Eriksson[1,2], Jannika Lindvall[1], Ola Helenius[3] and Andreas Ryve[1]*

[1] School of Education, Culture and Communication, Mälardalen University, Västerås, Sweden, [2] Centre for Cultural Evolution, Stockholm University, Stockholm, Sweden, [3] NCM, University of Gothenburg, Göteborg, Sweden

One of the many things teachers do is to give feedback on their students' work. Feedback pointing out mistakes may be a key to learning, but it may also backfire. We hypothesized that feedback based on students' mistakes may have more positive effects in cultures where teachers have greater authority over students, which we assume to be cultures that are high on power distance and religiosity. To test this hypothesis we analyzed data from 49 countries taking part in the 2015 wave of the TIMSS assessment, in which students in the 4th and 8th grades were asked whether their teachers in mathematics and science told them how to do better when they had made a mistake. For each country we could then estimate the association between the reported use of mistake-based feedback and student achievement. Consistent with our hypothesis, the estimated effect of mistake-based feedback was positive only in certain countries, and these countries tended to be high on power distance and religiosity. These results highlight the importance of cultural values in educational practice.

Keywords: negative feedback, power distance, religiosity, cultural values, effective instruction, mistakes

OPEN ACCESS

Edited by:
Ronnel B. King,
The Education University of Hong Kong, Hong Kong

Reviewed by:
Yulia Kuzmina,
Psychological Institute of Russian Academy of Education, Russia
Natalia Suárez,
University of Oviedo, Spain

****Correspondence:***
Kimmo Eriksson
kimmo.eriksson@mdh.se;
kimmoe@gmail.com

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *15 August 2019*
Accepted: *24 December 2019*
Published: *21 January 2020*

Citation:
Eriksson K, Lindvall J, Helenius O and Ryve A (2020) Cultural Variation in the Effectiveness of Feedback on Students' Mistakes. Front. Psychol. 10:3053. doi: 10.3389/fpsyg.2019.03053

INTRODUCTION

Some scholars argue that pedagogical methods and concepts are culturally embedded and that transplanting them from one culture to another is not always feasible (e.g., Hatano and Inagaki, 1998; Chen and Wong, 2015). In the present paper we focus on cultural differences in the effects of a specific teacher practice: to give feedback on students' mistakes. Students' mistakes have been argued to play a key role for learning (Mangels et al., 2006; Boaler, 2016) and reform initiatives in mathematics education (National Council of Teachers of Mathematics, 2000), research within the field of mathematics education (e.g., Kazemi and Stipek, 2001; Bray, 2011) as well as studies in psychology and neuroscience (see Boaler, 2016) emphasize that capitalizing on students' mistakes may be a particularly productive teaching practice. On the other hand, to go from mistakes to learning is not straightforward. Feedback on mistakes, also known as corrective feedback, may even be counter-productive, for example if students perceive they cannot understand the feedback, if it makes them focus on right and wrong answers instead of the solving process, or if it makes them dwell on their mistakes (Gagné et al., 1987; Fyfe and Rittle-Johnson, 2017). The timing and character of corrective feedback may therefore influence its efficacy. In general, feedback are thought to work through cognitive, motivational and meta-cognitive mechanisms that are affected by the relationship between the learning situation and the learner, and by the level of expertise and experience of the learner and the teacher (Hattie and Timperley, 2007; Harks et al., 2014). This means that social and cultural factors that influence the learning context may affect the efficacy of

feedback. The purpose of the present paper is to examine the cultural dependence of the efficacy of mistake-based feedback.

Cultural Variations in Mistake-Based Feedback

In classroom practice, teachers must choose how to react to students' mistakes. The reaction could be anything from simply ignoring mistakes to making them starting points for whole-class discussions (Bray, 2011). Teachers' feedback on how to do better when students have made a mistake is a particularly interesting teaching practice to study. Such corrective feedback can take many forms. For example, the teacher can criticize the student for making a mistake or praise the thought and emphasize the learning potential (Tulis, 2013), or the teacher can give feedback in the form of statements (e.g., giving the correct answer or providing an explanation) as well as questions (e.g., redirecting the question or asking for student explanation; Schleppenbach et al., 2007). Also how teachers' feedback on errors is perceived by students may differ. For example, students can feel embarrassed when their teacher point out errors or view them as opportunities to improve (Tulis, 2013).

The ways in which teachers handle mistakes and how mistakes are perceived by the students may be nationally embedded and have shown to differ between countries. For example, in analyzing teacher–student interactions surrounding students' mistakes in 60 videotaped 8th grade mathematics lessons, Santagata (2004) found differences between Italian and United States teachers. While United States teachers tended to mitigate student mistakes and students rarely blamed themselves for making them, Italian teachers more often aggravated student mistakes (e.g., by openly showing their disappointment) and students took responsibility for making them. Another example is Schleppenbach et al. (2007) who analyzed videotaped lessons from elementary mathematics lessons and found differences in how mistakes were treated in the United States and China. Their results indicate that United States and Chinese students made errors at similar frequencies, but that the teachers in the two countries responded to them differently. The United States teachers made more statements about errors than the Chinese teachers, who instead asked more follow-up questions about errors. Moreover, when questions were used, the United States teachers tended to question students by asking them to evaluate the answer, while the Chinese teachers questioned students by asking them to correct or explain the error. Cultural differences in how students mistakes are viewed have also been discussed by Stevenson and Stigler (1992) and Wang and Murphy (2004) who argue that, in the American context, mistakes are often seen as failures and something that makes students appear silly, while they in China and Japan are viewed as signs of what needs to be learned. Video analysis combined with student questionnaires confirm the existence of culturally dependent feedback effects in comparisons of Switzerland and Germany, where the Swiss students rate their opportunities to learn from errors higher (Dalehefte et al., 2012). Although all of these studies lay important groundwork for using a cross-cultural approach when looking at feedback on students' mistakes, we know of no previous cross-cultural work in education focusing on the *effects* of teachers' mistake-based feedback on student achievement at a larger-scale.

How Culture May Influence the Effectiveness of Mistake-Based Feedback

In an influential review of how culture may influence students' approaches to learning, Littlewood (1999) focused on aspects in which East Asian culture tends to differ from Western culture: high power distance (normalcy of inequality in power and authority) and high collectivism (emphasis on interdependence instead of individuality), including a belief in the adaptability of individuals through effort. These aspects may both have bearing on the effect of negative feedback. Psychological research has suggested that Japanese are more willing than North Americans to accept negative feedback and try to improve from it (Heine et al., 2001a,b). The theoretical rationale for this difference, according to Heine et al. (2001a, p. 435), is that Japan is a culture that emphasizes "hierarchy and role mastery" and that in this context, "the discovery of negative features about the self serves to highlight where one needs to make efforts so as to move toward the consensually shared standards." In other words, the authors simultaneously appealed to power distance (hierarchy) and collectivism (consensually shared standards to which individuals can adapt through effort). On the basis of this prior research we may hypothesize that mistake-based feedback from teachers leads to better student achievement specifically in cultures that are high on power distance and high on collectivism. We shall now elaborate on the rationale for these hypotheses.

Teacher Authority, Power Distance, and Religiosity

According to Hofstede (2001), societies vary in the extent to which inequality in power is accepted and regarded as normal. When applied to the teacher-student relationship, high power distance implies that teachers have great authority. Students respect the teacher, they appreciate that the teacher tells them what to do, they speak up only when invited, and they do not contradict the teacher. Students in societies with low power distance have less respect for teachers and are more likely to challenge teachers' authority and rely on their own experience instead (Hofstede, 1986; Woodrow and Sham, 2001; Joy and Kolb, 2009; Holtbrügge and Mohr, 2010). It would fit with this general picture of teacher authority that students would be more accepting of negative feedback from the teacher[1].

Although we will use power distance as a proxy for cultural differences in teacher authority, we acknowledge power and authority are not exactly synonymous. An important aspect of authority is being a "reliable guide as to how things are" (Raz, 1990, p. 2), thus connecting authority with religion. It stands to reason that more religious societies are more accepting of religious authority and, plausibly, of teacher authority in general. This notion does not seem to be well researched, but scholars

[1]The teacher's authoritative role in high power distance societies may at the same time provide a barrier for other teaching practices such as self-evaluation and peer-assessment (Brown et al., 2009; Carless and Lam, 2014; Thanh Pham and Renshaw, 2015).

have argued for a strong parallel between religious authority and teachers' authority (Smith, 2013). Another supporting piece of evidence is that country-level religiosity and power distance are strongly, but not perfectly, correlated (Parboteeah et al., 2008). For these reasons we shall use country-level religiosity as an alternative proxy for teacher authority to complement power distance.

Collectivism and the Interdependent and Adaptable Self

The cultural dimension of collectivism vs. individualism concerns the degree to which individuals are first and foremost regarded as parts of a collective and perceive themselves as interdependent (Markus and Kitayama, 1991; Hofstede, 2001). Compared to people in individualist countries, people in collectivist countries seem to be more likely to view ability and wisdom as adaptable through effort rather than fixed in an individual (Heine et al., 2001b). This may further be connected to the concept of fixed (believing intelligence is as fixed entity) versus growth (believing intelligence is malleable) mindset (Dweck, 2006), which have also been indicated to be associated with trait use and could help account for cultural differences (Dweck et al., 1995). In a recent meta-analytic review, for example, Costa and Faria (2018) found that educational studies conducted in Asia and Oceania reported a significant association between growth mindset and student achievement while, in Europe, a fixed mindset was modestly and positively associated with student achievement. The authors suggest that this may reflect cultural differences were more collectivist societies (such as many Asian countries) might encourage students to value the learning process over individual academic achievement, while in Europe there is a tendency toward a more academically competitive society where students may prioritize individual results over knowledge.

Moreover, it has been suggested that collectivism may facilitate the acceptance of negative feedback because it enables individuals to identify their weaknesses in order to improve and blend in Heine et al. (2001b) and Gelfand et al. (2002). As Heine et al. (2001a,b), a sense that ability is not innate but improves with effort may make negative feedback less threatening and thus presumably more effective in promoting learning. In fact, studies in neuroscience (Mangels et al., 2006; Moser et al., 2011) have shown that students react differently to negative feedback depending on differences in their mindsets. Compared to students endorsing a more fixed mindset, more growth minded individuals showed superior knowledge gains in that they demonstrated greater remediation of errors and were more likely to reflect awareness of and allocation of attention to mistakes.

The Present Study

The aim of the present study is to examine the relation between the above-mentioned cultural dimensions and the effectiveness of teachers giving mistake-based feedback to students. Every country then provides just a single data point. It is therefore imperative to obtain data from as many countries as possible. We use data on student achievement and teaching practices in 49 countries obtained from TIMSS, the Trends in International Mathematics and Science Study. TIMSS is conducted every four years by IEA (International Association for the Evaluation of Educational Achievement). Entire classes in grades 4 and 8 are sampled and participating students are linked to the teacher/classroom level. Students are given achievement tests in mathematics and science as well as a background questionnaire including some items on teachers' use of various instructional practices. Such data can be related to student outcomes to estimate the association between instructional practices and achievement (e.g., Blömeke et al., 2016; Eriksson et al., 2019). Specifically, the 2015 wave of TIMSS[2] included the item "My teacher tells me how to do better when I make a mistake." This allows us to estimate the association between the use of mistake-based feedback and student achievement within each country that participated in 2015 TIMSS.

We shall assume that these associations reflect the effects of mistake-based feedback on achievement (other possibilities are addressed in the discussion). Under this assumption our hypotheses can be tested by examination of how effects of mistake-based feedback correlate with available country-measures of collectivism, power distance, and religiosity.

MATERIALS AND METHODS

In brief, the method of our study consists of two steps. The first step is to use TIMSS data to obtain estimates per country of the effect of mistake-based feedback on achievement. The second step is to examine if these estimates are related to country measures of power distance, religiosity, and collectivism.

Out of 55 countries that participated in TIMSS 2015 (4th grade, 8th grade, or both)[3], we study 49 countries for which country measures of religiosity, power distance, and collectivism were available. The 49 countries are listed in **Table 1**, which also reports the size of the TIMSS student sample and the number of classes sampled in each country.

Countries, TIMSS Samples, and Country Measures From Other Sources

From the 2009 global Gallup we obtained country measures of *religiosity* in terms of the percentage of the sampled population who responded "yes" to the question: "Is religion important in your daily life?" (Crabtree, 2010). In our sample of countries, the percentage who judged religion as important ranged from 17 to 99 (M = 58, SD = 27).

Estimates of the *power distance* and *individualism* for each country in our sample, on a scale from 0 to 100, were taken from Hofstede's website[4] and are reported in **Table 1**. To obtain a collectivism measure we reverse coded the individualism measure (i.e., collectivism = 100 - individualism). In our sample of countries, power distance ranged from 13 to 100 (M = 59, SD = 22) and collectivism ranged from 9 to 86 (M = 51, SD = 24).

[2]Data and questionnaires are available at https://timssandpirls.bc.edu/timss2015/international-database/.

[3]TIMSS data from England and Northern Ireland were merged to represent the United Kingdom.

[4]https://www.hofstede-insights.com/country-comparison/

TABLE 1 | TIMSS Sample Sizes and Country Measures from Other Sources than TIMSS.

	Grade 4		Grade 8		Relig. imp.	Power dist.	Collect.	GNI/cap.
Country	**students**	**classes**	**students**	**classes**				
Australia	6057	498	10338	645	0.32	36	10	43
Belgium (Flemish)	5404	295			0.33	61	22	41
Bulgaria	4228	233			0.34	70	70	16
Canada	12283	696	8757	409	0.42	39	20	43
Chile	4756	179	4849	173	0.70	63	77	22
Croatia	3985	223			0.70	73	67	20
Czechia	5202	265			0.21	57	42	28
Denmark	3710	194			0.19	18	26	45
Egypt			7822	215	0.97	70	75	10
Finland	5015	300			0.28	33	37	39
France	4873	273			0.30	68	29	38
Germany	3948	213			0.40	35	33	45
Hong Kong	3600	145	4155	145	0.24	68	75	54
Hungary	5036	241	4893	241	0.39	46	20	23
Indonesia	4025	312			0.99	78	86	10
Iran	3823	291	6130	251	0.73	58	59	16
Ireland	4344	214	4704	204	0.54	28	30	44
Israel			5512	200	0.51	13	46	31
Italy	4373	257	4481	230	0.72	50	24	34
Japan	4383	148	4745	147	0.24	54	54	37
Jordan			7865	260	0.96	70	70	10
Korea	4669	188	5309	170	0.43	60	82	35
Kuwait	3593	294	4503	191	0.91	90	75	76
Lebanon			3873	185	0.87	75	60	13
Lithuania	4529	290	4347	252	0.42	42	40	26
Malaysia			9726	326	0.96	100	74	25
Malta			3817	223	0.86	56	41	29
Morocco	5068	374	13035	375	0.97	70	54	7
Netherlands	4515	223			0.33	38	20	46
New Zealand	6322	459	8142	377	0.33	22	21	33
Norway	4329	222	4697	216	0.21	31	31	68
Poland	4747	254			0.75	68	40	24
Portugal	4693	321			0.72	63	73	26
Qatar	5194	224	5403	238	0.95	93	75	130
Russia	4921	217	4780	221	0.34	93	61	23
Saudi Arabia	4337	189	3759	149	0.93	95	75	51
Serbia	4036	192			0.54	86	75	12
Singapore	6517	358	6116	334	0.70	74	80	78
Slovakia	5773	327			0.47	100	48	27
Slovenia	4445	255	4257	217	0.47	71	73	29
South Africa			12514	328	0.85	49	35	12
Spain	7764	379			0.49	57	49	33
Sweden	4142	211	4090	206	0.17	31	29	46
Taiwan	4291	177	5711	191	0.45	58	83	46
Thailand			6482	213	0.97	64	80	15
Turkey	6456	251	6079	220	0.82	66	63	19
United Arab Emirates	21177	891	18012	763	0.91	90	75	66
United Kingdom	7122	242	4814	213	0.27	35	11	38
United States	10029	497	10221	534	0.69	40	9	53
Total sample	227714	12012	223938	9262				

The last four columns are (1) the proportion of the population that thinks religion is important, (2) Hofstede's measure of power distance, (3) Hofstede's measure of collectivism, and (4) GNI per capita in thousands of international dollars.

Country scores of *gross national income (GNI) per capita* in 2015, measured in international dollars, were downloaded from the Human Development Report Office of the United Nations[5]. For Taiwan we used the measure for 2015 from their National Statistics agency[6]. In our sample of countries, GNI per capita ranged from 7,000 in Morocco to 130,000 in Qatar (M = 35,000, SD = 22,000).

Estimation of the Effect of Mistake-Based Feedback on Student Achievement

To estimate the effectiveness of mistake-based feedback we used data from TIMSS on student achievement and teachers' use of mistake-based feedback, as well as some control variables.

Student Achievement in Mathematics and Science

TIMSS uses an elaborate method to measure student achievement in mathematics and science (Martin et al., 2016). In brief, each student responds to only a subset of test questions and five "plausible values" for the total score of each student are generated through an imputation method. Plausible values are given on a scale that was calibrated so that the 1995 TIMSS results had a mean of 500 and a standard deviation of 100. We used the set of five plausible values of student achievement in math and science as measured in TIMSS 2015, standardized within each country to unit standard deviation.

Use of Mistake-Based Feedback

The grade 4 and grade 8 student questionnaires of TIMSS 2015 included one part about mathematics and one part about science. Both parts included a set of ten items about the teacher. For each item, students gave their response on a four-point scale: *Disagree a lot* (coded 1); *Disagree a little* (coded 2); *Agree a little* (coded 3); *Agree a lot* (coded 4). Our focus is on the item "My teacher tells me how to do better when I make a mistake," which we shall refer to as MBF (mistake-based feedback). On the MBF item, almost all responses were either *Agree a lot* or *Agree a little* (93% in grade 4, 85% in grade 8). This means that MBF was nearly a binary variable. (Indeed, if we recode it as binary by lumping the two *Disagree* options together with *Agree a little,* all the main results presented in this paper would remain virtually identical).

Following prior research we average responses from all students in a class to obtain a measure of the teacher's teaching style (Eriksson et al., 2019). (In the binary recoding, the class average would simply reflect how frequently students in a class responded by *Agree a lot* to the MBF item about a given teacher.) This is taken as a measure of how much the teacher uses mistake-based feedback. The class-average response to the MBF item for each of the two teachers yielded two class-level measures, which we refer to as MBF:Math and MBF:Science. For descriptive statistics of MBF:Math and MBF:Science in each grade in each country, see **Table 2**. There were eight countries in which science grade 8 was not taught by a single teacher but by several teachers specializing in different science disciplines. For these countries no MBF:Science measures in grade 8 were calculated (as they would be ambiguous).

In **Table 2**, note that the country-means of MBF:Math and MBF:Science are consistently between 3 and 4, reflecting that these were the dominant individual responses. However, there were specific classes where the MBF measures were much lower than 3, as illustrated in **Figure 1** showing the distribution of the MBF measure for mathematics across all participating classes in 8th grade. The corresponding distributions for 4th grade mathematics and for science look similar.

Control Variables

As described in detail below, we estimate the effect of mistake-based feedback both with and without including control variables. Ideally, results are robust to the model specification. The following control variables are used.

First, when estimating the effect of MBF of the mathematics, we control for the MBF of the science teacher, and vice versa.

In addition to the MBF item, the student questionnaire included nine other items (using the same response scale) about the teacher: "I know what my teacher expects me to do," "My teacher is easy to understand," "I am interested in what my teacher says," "My teacher gives me interesting things to do," "My teacher has clear answers to my questions," "My teacher is good at explaining mathematics," "My teacher lets me show what I have learned," "My teacher does a variety of things to help us learn," "My teacher listens to what I have to say." Note that all of these items are positive statements about the teacher. For each teacher subject (math and science) we calculated the student's mean response to these items (Cronbach's $\alpha > 0.86$ for each academic subject in each grade), and then averaged this measure over all students of the class. We refer to these class-level measures as Pos:Math and Pos:Science. These measures were typically between 3 and 4, meaning that students tended to agree at least a little with the nine positive statements about the teacher. We use the Pos measures as control variables to ascertain that estimated effects of MBF do not simply reflect effects of a generally positive view of the teacher.

When studying antecedents of student achievement it is common to control for socio-economic status and gender. Following some previous research on TIMSS data (Blömeke et al., 2016; Eriksson et al., 2019), we used the response to the item "About how many books are there in your home?" as a proxy for socio-economic status, henceforth referred to as *SES*. This item has a five-point response scale from *None or very few* (*0–10 books*) (coded 1) to *Enough to fill three or more bookcases* (*more than 200*) (coded 5). Student *gender* was coded 1 for girl and 2 for boy.

TIMSS also includes four teacher background variables that we used as controls: experience (years of teaching), age, gender, and level of formal education.

Missing Data

There were at most a few percent missing data on the items we use. Missing data were handled using the multiple imputation functionality of SPSS v. 24, generating five sets of imputed data,

[5]http://hdr.undp.org/

[6]http://eng.stat.gov.tw/

TABLE 2 | Descriptive Statistics of MBF Measures.

	Grade 4				Grade 8			
	Math		Science		Math		Science	
Country	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Australia	3.58	0.24	3.50	0.26	3.13	0.43	3.02	0.39
Belgium (Flemish)	3.65	0.20	3.52	0.25				
Bulgaria	3.85	0.17	3.84	0.17				
Canada	3.65	0.22	3.59	0.24	3.33	0.34	3.17	0.38
Chile	3.68	0.23	3.65	0.23	3.34	0.40	3.33	0.33
Croatia	3.65	0.20	3.67	0.21				
Czechia	3.42	0.32	3.39	0.32				
Denmark	3.57	0.25	3.36	0.37				
Egypt					3.52	0.26	3.54	0.24
Finland	3.66	0.22	3.60	0.25				
France	3.54	0.22	3.43	0.30				
Germany	3.60	0.26	3.55	0.27				
Hong Kong	3.37	0.26	3.37	0.26	3.12	0.35	3.12	0.32
Hungary	3.68	0.23	3.65	0.23	3.18	0.41		
Indonesia	3.77	0.28	3.72	0.27				
Iran	3.74	0.23	3.77	0.22	3.40	0.35	3.41	0.34
Ireland	3.73	0.20	3.63	0.25	3.28	0.27	3.12	0.36
Israel					3.24	0.36	3.08	0.49
Italy	3.63	0.21	3.59	0.24	3.22	0.36	3.10	0.38
Japan	3.29	0.29	3.17	0.30	2.92	0.29	2.79	0.30
Jordan					3.60	0.22	3.58	0.26
Korea	3.13	0.36	3.06	0.36	2.67	0.27	2.58	0.29
Kuwait	3.67	0.30	3.71	0.29	3.41	0.34	3.42	0.35
Lebanon					3.51	0.34		
Lithuania	3.75	0.16	3.74	0.17	3.37	0.31		
Malaysia					3.44	0.27	3.44	0.29
Malta					3.31	0.38		
Morocco	3.73	0.32	3.74	0.31	3.47	0.33		
Netherlands	3.51	0.24	3.47	0.23				
New Zealand	3.58	0.23	3.49	0.27	3.15	0.36	3.13	0.35
Norway	3.71	0.19	3.65	0.23	3.28	0.37	3.16	0.40
Poland	3.39	0.31	3.39	0.30				
Portugal	3.89	0.13	3.86	0.16				
Qatar	3.61	0.26	3.61	0.25	3.26	0.36	3.24	0.37
Russia	3.78	0.17	3.75	0.18	3.48	0.28		
Saudi Arabia	3.65	0.27	3.60	0.27	3.39	0.34	3.38	0.35
Serbia	3.81	0.15	3.80	0.15				
Singapore	3.51	0.25	3.43	0.28	3.23	0.30	3.20	0.27
Slovakia	3.65	0.26	3.59	0.29				
Slovenia	3.58	0.23	3.57	0.23	3.13	0.29		
South Africa					3.57	0.23	3.46	0.25
Spain	3.81	0.22	3.77	0.24				
Sweden	3.55	0.22	3.44	0.24	3.12	0.35		
Taiwan	3.51	0.28	3.44	0.29	3.20	0.33	3.03	0.32
Thailand					3.48	0.22	3.44	0.24
Turkey	3.82	0.17	3.83	0.17	3.63	0.27	3.61	0.28
United Arab Emirates	3.61	0.27	3.60	0.29	3.33	0.34	3.26	0.34
United Kingdom	3.73	0.17	3.61	0.23	3.28	0.39	3.13	0.29
United States	3.68	0.20	3.61	0.22	3.27	0.39	3.22	0.38

Each MBF measure is the class-mean response on whether the teacher (math or science) uses mistake-based feedback, on a scale from 1 (disagree a lot) to 4 (agree a lot).

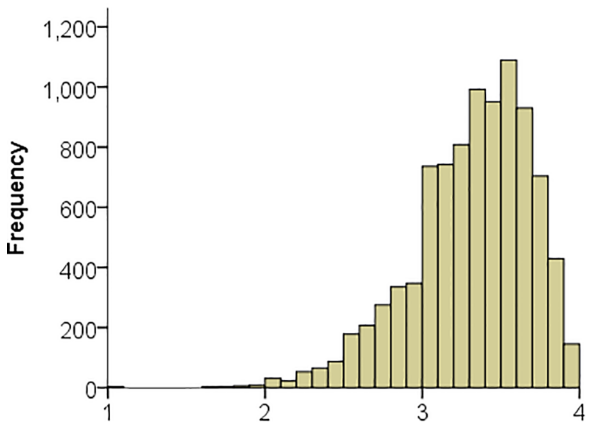

FIGURE 1 | Histogram over MBF:Math, the class-level measure of mistake-based feedback in mathematics, of all classes in 8th grade.

to each of which one of the five pairs of plausible values on mathematics and science achievement was assigned.

Estimation of Within-Country Effects of MBF on Achievement in Math and Science in Grades 4 and 8

To account for the multiple levels of data in each country (class and student) we include a random effect of class (O'Connell and McCoach, 2008). We estimate the effect of MBF *in a given country for a given subject in a given grade* using two different models: without control variables,

$$Y_{ij} = \gamma_{00} + \gamma_{01} MBF_j + u_j + r_{ij}$$

and with control variables,

$$Y_{ij} = \gamma_{00} + \gamma_{01} MBF_j + \gamma_{02} Pos_j + \gamma_{03} Ex_j + \gamma_{04} Ag_j + \gamma_{05} Ge_j + \gamma_{06} Ed_j + \gamma_{07} OMBF_j + \gamma_{10} SES_{ij} + \gamma_{20} Gen_{ij} + u_j + r_{ij}.$$

Here Y_{ij} denotes the achievement in the given subject for student i in class j; γ_{00} is the class-level intercept; MBF_j, Pos_j, Ex_j, Ag_j, Ge_j, and Ed_j are the MBF and Pos measures and the experience, age, gender and level of education for the teacher in the given subject in class j; $OMBF_j$ denotes the MBF measure of the teacher in the *other* subject in class j; SES_{ij} and Gen_{ij} denote the socio-economic status and gender of student i in class j; u_j is a random error term representing a unique effect associated with class j and r_{ij} is a random error term at the individual level. Error terms are assumed to have a normal distribution with a mean of zero and constant variance.

Analyses were conducted using restricted maximum likelihood (REML) estimation in the linear mixed model function of SPSS v. 24. Using the SPSS functionality for analysis of multiply imputed data, analyses were performed on each set of imputed data and then pooled to yield unbiased estimates of effects and standard errors.

By the above procedure the effect of MBF on achievement (i.e., the coefficient γ_{01}) was estimated up to eight times per country: two models (with controls and without controls) for each of two subjects (math and science) in each of two grades (4th and 8th). Estimates and standard errors are reported in **Tables 3**, **4**. To obtain approximate 95% confidence intervals, take the estimate plus/minus two standard errors.

An Alternative Approach

Our main approach has two steps. In the first step we estimate the effect of class-level MBF on student achievement separately in each country, using a two-level analysis (student and class). In the second step we examine how these estimates per country relate to country-level measures of power distance, religiosity, and collectivism.

An alternative approach is to include all countries from the beginning in a three-level analysis (student, class, and country) of student achievement to examine the interaction of class-level MBF and country-level measures. Without controls, the model for a given subject in a given grade would then be

$$Y_{ijk} = \gamma_{000} + \gamma_{001} CLM_k + \gamma_{011} MBF_{jk} + \gamma_{012} CLM_k{}^* MBF_{jk} + v_k + u_{jk} + r_{ijk},$$

where Y_{ijk} denotes the achievement in the given subject for student i in class j in country k; γ_{000} is the country-level intercept; CLM_k is a country-level measure (say, power distance) in country k; MBF_{jk} is the MBF measure for the teacher in the given subject in class j in country k; v_k and u_{jk} are random error terms representing unique effects associated with country k and class j, respectively, and r_{ijk} is a random error term at the student level. When adding control variables to this model, we also include their interactions with the country-level measure (e.g., we would include a Pos_{jk} term as well as the interaction term $CLM_k{}^* Pos_{jk}$).

The advantage of our main approach is that we explicitly obtain country estimates of the MBF effect, thereby allowing easy examination of their consistency across grades, subjects, and model specifications, as well as easy illustration of their relation to a country measure using a scatter plot. The advantage of the alternative approach is that it yields a more accurate estimate of the statistical significance of the latter relation, which in the above model is captured by the interaction term $CLM_k{}^* MBF_{jk}$. We use the alternative approach only to verify the statistical significance of the interaction. These analyses were performed in the lme4 package (Bates et al., 2014).

RESULTS

Descriptive Statistics of Estimated MBF Effects

In the estimation of MBF effects, all variables were standardized within each country. Therefore, estimated MBF effects are measured in the unit "within-country standard deviation in achievement per within-country standard deviation in MBF."

TABLE 3 | Estimates of the MBF effect on achievement in Grade 4.

	Math				Science			
	W/o controls		With controls		W/o controls		With controls	
Country	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
Australia	0.02	0.03	0.01	0.03	−0.03	0.02	−0.05	0.03
Belgium (Flemish)	−0.08*	0.03	−0.03	0.04	−0.13***	0.03	−0.10	0.05
Bulgaria	0.09	0.05	−0.08	0.08	0.13*	0.05	−0.01	0.07
Canada	0.00	0.02	0.02	0.04	−0.01	0.02	−0.13*	0.04
Chile	0.06	0.05	0.00	0.09	0.05	0.05	0.17	0.10
Croatia	0.02	0.03	0.03	0.05	0.03	0.03	0.07	0.05
Czechia	−0.05	0.03	−0.05	0.05	−0.02	0.03	0.04	0.05
Denmark	0.03	0.04	−0.04	0.05	−0.01	0.04	−0.15*	0.07
Finland	−0.02	0.03	−0.01	0.06	−0.01	0.05	0.00	0.09
France	−0.07*	0.03	−0.08	0.04	−0.08*	0.03	−0.05	0.05
Germany	0.05	0.04	0.06	0.05	0.00	0.04	−0.02	0.06
Hong Kong	0.01	0.06	−0.19*	0.09	0.08	0.05	−0.18	0.10
Hungary	−0.06	0.04	0.02	0.06	−0.10	0.04	−0.18**	0.06
Indonesia	0.18***	0.04	0.16*	0.06	0.26***	0.04	0.25***	0.06
Iran	0.05	0.04	0.02	0.06	0.04	0.04	−0.02	0.06
Ireland	−0.04	0.03	0.04	0.04	−0.03	0.03	0.01	0.06
Italy	0.01	0.03	−0.02	0.05	0.03	0.03	0.06	0.05
Japan	0.05	0.03	0.03	0.05	0.03	0.03	−0.03	0.07
Korea	0.02	0.03	−0.29*	0.07	0.04	0.04	−0.20*	0.07
Kuwait	0.09*	0.04	0.09	0.07	0.12***	0.03	0.05	0.06
Lithuania	0.07*	0.03	0.06	0.04	0.08	0.04	0.07	0.05
Morocco	0.14***	0.04	0.04	0.06	0.10**	0.03	0.00	0.05
Netherlands	0.00	0.03	0.09	0.05	−0.05	0.03	−0.17*	0.06
New Zealand	−0.09**	0.03	−0.01	0.04	−0.11*	0.04	−0.11	0.07
Norway	0.01	0.03	−0.08	0.04	0.01	0.03	−0.06	0.05
Poland	−0.06	0.03	−0.10	0.05	−0.05	0.03	−0.06	0.05
Portugal	0.02	0.03	−0.03	0.04	0.02	0.03	−0.03	0.04
Qatar	0.31***	0.04	0.25*	0.09	0.32***	0.04	0.19*	0.07
Russia	−0.03	0.05	−0.18*	0.07	0.00	0.05	0.11	0.06
Saudi Arabia	0.26***	0.05	0.21	0.11	0.26***	0.05	0.15	0.09
Serbia	−0.10**	0.03	−0.06	0.05	−0.09**	0.03	0.01	0.06
Singapore	0.00	0.04	−0.35***	0.06	−0.02	0.04	−0.21*	0.08
Slovakia	−0.12**	0.04	0.04	0.06	−0.11*	0.04	0.04	0.06
Slovenia	0.00	0.03	−0.09	0.05	0.05	0.03	0.05	0.05
Spain	0.02	0.03	0.08	0.05	−0.02	0.03	−0.12	0.07
Sweden	−0.10*	0.04	−0.20**	0.06	−0.07	0.04	−0.03	0.06
Taiwan	0.02	0.03	0.00	0.05	0.02	0.03	−0.09	0.05
Turkey	0.28***	0.04	0.06	0.05	0.26***	0.04	0.02	0.05
United Arab Emirates	0.31***	0.02	0.18**	0.06	0.33***	0.02	0.18**	0.06
United Kingdom	0.00	0.01	−0.08*	0.03	0.03**	0.01	−0.02	0.04
United States	0.01	0.03	−0.12*	0.05	0.00	0.03	−0.17***	0.04

Estimates are standardized coefficients for the effect of MBF on achievement. Stars indicate p-values with respect to the null hypothesis that the true MBF effect is zero ($p < 0.05$; **$p < 0.01$; ***$p < 0.001$).*

Starting with grade 4, the estimated MBF effects per country in **Table 3** can be summarized as follows: The mean MBF effect was close to zero in both subjects, regardless of model specification, but there was substantial variation between countries. To illustrate, consider MBF effects for math estimated with controls in grade 4: the mean effect was −0.01, $p = 0.47$, with a standard deviation of 0.12 and a range from −0.35 to 0.25. Thus, it seems that there are some countries where the MBF effect is positive and other countries where the MBF effect is negative.

Estimated MBF effects per country in grade 8 showed the same pattern, see **Table 4**. To illustrate, consider MBF effects for math

TABLE 4 | Estimates of the MBF effect on achievement in Grade 8.

	Math				Science			
	W/o controls		With controls		W/o controls		With controls	
Country	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
Australia	0.14***	0.03	0.00	0.06	0.08*	0.03	−0.19***	0.05
Canada	0.02	0.03	−0.16*	0.07	0.00	0.03	−0.29***	0.07
Chile	−0.03	0.06	−0.21	0.14	0.01	0.05	−0.38*	0.15
Egypt	0.03	0.04	0.23*	0.09	0.10*	0.04	0.05	0.09
Hong Kong	0.08	0.07	0.23	0.20	0.09	0.06	−0.45*	0.18
Hungary	−0.10	0.05						
Iran	−0.01	0.04	0.08	0.10	0.00	0.04	0.04	0.10
Ireland	0.00	0.04	−0.15*	0.07	0.13**	0.04	−0.07	0.07
Israel	−0.10	0.05	−0.32**	0.11		0.05	−0.37*	0.14
Italy	−0.03	0.03	−0.07	0.08	−0.07*	0.03	−0.12	0.08
Japan	0.07*	0.03	−0.03	0.07	0.07*	0.03	−0.14*	0.06
Jordan	0.08*	0.03	0.14	0.07	0.09**	0.03	0.08	0.07
Korea	−0.03	0.03	−0.16**	0.05	0.10***	0.02	−0.09	0.05
Kuwait	0.03	0.04	−0.05	0.08	0.13*	0.05	−0.01	0.12
Lebanon	0.12*	0.05						
Lithuania	0.02	0.04						
Malaysia	0.41***	0.05	0.43***	0.11	0.50***	0.04	0.46***	0.11
Malta	0.17**	0.06						
Morocco	0.05	0.03						
New Zealand	0.03	0.04	−0.06	0.08	0.07	0.04	−0.15	0.08
Norway	0.08*	0.03	−0.06	0.05	0.03	0.03	0.03	0.08
Qatar	0.21***	0.04	0.11	0.10	0.20***	0.04	−0.15	0.11
Russia	0.07	0.04						
Saudi Arabia	0.11*	0.05	0.03	0.13	0.15**	0.05	0.05	0.16
Singapore	0.08	0.05	−0.34**	0.10	−0.03	0.05	−0.43***	0.10
Slovenia	0.00	0.03						
South Africa	−0.08	0.04	0.20**	0.07	−0.15***	0.04	−0.23*	0.08
Sweden	0.01	0.04						
Taiwan	0.04	0.04	−0.17**	0.06	0.07*	0.03	−0.11	0.06
Thailand	−0.10	0.06	0.25*	0.12	0.05	0.06	0.23*	0.11
Turkey	0.06	0.04	0.08	0.07	0.09*	0.04	0.14	0.09
United Arab Emirates	0.23***	0.02	0.17*	0.07	0.18***	0.02	0.09	0.08
United Kingdom	0.04	0.07	−0.60***	0.14	0.12	0.06	0.00	0.11
United States	0.03	0.03	−0.27***	0.07	0.05	0.03	−0.33***	0.07

Estimates are standardized coefficients for the effect of MBF on achievement. Stars indicate p-values with respect to the null hypothesis that the true MBF effect is zero ($p < 0.05$; **$p < 0.01$; ***$p < 0.001$). For eight countries in which science is taught by several teachers, no MBF measure for science was calculated. Hence for these countries we could not estimate the MBF effect for science, nor could we estimate the MBF effect for math with controls (as the MBF for science is one of the controls in the model).*

estimated with controls in grade 8: the mean effect was −0.03, $p = 0.56$, with a standard deviation of 0.22 and a range from −0.60 to 0.43.

Relation Between the MBF Effect and Other Country Variables

Table 5 reports pairwise correlations, with bootstrapped confidence intervals, of the estimated MBF effects against religiosity, power distance, collectivism, and GNI per capita. The table reveals a consistent pattern. Regardless of the method and data used to estimate the MBF effect, it was always positively correlated with religiosity, power distance, and collectivism. The strength of the correlations varied across different estimates, but overall correlations tended to be stronger for religiosity (average correlation = 0.47) and power distance (average correlation = 0.44) than for collectivism (average correlation = 0.33). In **Table 5**, a few of the confidence intervals include zero, indicating a non-significant relation. However, when we conducted corresponding analyses using the alternative approach described in section 2.3, the interaction between MBF and these culture variables always came out as significantly positive. Thus, we conclude that there are robust positive associations between MBF effects on the

TABLE 5 | Correlations between MBF effect estimates and country measures from other sources.

MBF Effect Estimates			Religion important	Power distance	Collectivism	GNI per capita
Grade	**Subject**	**Controls**				
4th	Math	without	0.66 [0.43, 0.77]	0.38 [0.04, 0.65]	0.46 [0.26, 0.62]	0.31 [− 0.19, 0.63]
		with	0.46 [0.12, 0.71]	0.18 [− 0.17, 0.53]	0.05 [− 0.28, 0.43]	0.13 [− 0.44, 0.54]
	Science	without	0.65 [0.42, 0.77]	0.46 [0.14, 0.72]	0.57 [0.39, 0.71]	0.28 [− 0.25, 0.61]
		with	0.52 [0.24, 0.71]	0.51 [0.28, 0.69]	0.41 [0.07, 0.72]	−0.02 [− 0.53, 0.34]
8th	Math	without	0.26 [− 0.09, 0.51]	0.54 [0.24, 0.72]	0.23 [− 0.12, 0.48]	0.33 [− 0.02, 0.67]
		with	0.50 [0.18, 0.75]	0.54 [0.26, 0.75]	0.41 [0.03, 0.69]	−0.18 [− 0.56, 0.11]
	Science	without	0.24 [− 0.15, 0.50]	0.53 [0.15, 0.74]	0.31 [0.06, 0.53]	0.18 [− 0.24, 0.65]
		with	0.45 [0.10, 0.69]	0.41 [0.04, 0.67]	0.23 [− 0.12, 0.54]	−0.25 [− 0.53, 0.01]

Correlations are calculated based on 41 data points (countries) in grade 4, based on 34 data points in 8th grade math without controls, and based on 26 data points in the other three analyses of 8th grade data. Correlation coefficients are reported with 95% confidence intervals (BCa, 1000 bootstrap samples).

one hand and religiosity, power distance, and collectivism on the other hand.

To increase the set of countries and use both grades and both subjects, we calculated an aggregate estimate of the controlled MBF effect by taking the average of all available controlled estimates for a given country. This yielded an aggregate estimate of the controlled MBF effect for 47 different countries. This aggregate estimate correlated with religiosity at $r = 0.54$, bootstrapped 95% CI [0.30,0.73], with power distance at r = 0.52 [0.24,0.71], and with collectivism at $r = 0.37$ [0.08,0.63]. Using religiosity, power distance, and collectivism as simultaneous predictors in a multiple linear regression of the aggregate MBF effect, we found they together explained 37% of the total variance, with statistically significant independent effects of both religiosity, $\beta = 0.38$, $p = 0.015$, and power distance, $\beta = 0.39$, $p = 0.035$, but not of collectivism, $\beta = -0.11$, $p = 0.52$.

The relation between the aggregate MBF effect and power distance is illustrated by a scatter plot in **Figure 2**. Note that the regression line fits the data points fairly with two exceptions: Singapore and Malaysia are outliers in different directions. If the two outliers are excluded, the correlation between the MBF effect and power distance increases slightly to $r = 0.55$ [0.38,0.71], and similarly for the correlations with religiosity, $r = 0.58$ [0.32,0.78] and collectivism, $r = 0.46$ [0.19,0.72].

DISCUSSION

In this paper we have used data from an international assessment of mathematics and science achievement to examine the effect of teachers giving feedback on students' mistakes. This is a teaching practice that has both proponents and critics. Our data support both views. In some countries (such as the United Arab Emirates) we found a positive association between teachers' use of feedback on mistakes and their students' achievement relative other students in the same country. In some other countries (such as the United States), the association was negative, at least after controlling for some potential confounders.

Based on prior cross-cultural work on negative feedback in other contexts (e.g., Heine et al., 2001a,b), we hypothesized that culture would moderate the effectiveness of mistake-based feedback. Specifically, it should be more effective in cultures where teachers have more authority. In the absence of a direct measure we examined two other widely available cultural measures, power distance and religiosity, which other scholars have thought to be associated with teachers' authority (e.g., Hofstede, 1986; Smith, 2013). In line with our hypothesis, we found both measures to be positively correlated with the effectiveness of mistake-based feedback.

We also hypothesized that mistake-based feedback would be more effective in cultures where students are more motivated to adapt to consensually shared standards and are more likely to have a growth mind-set. In the absence of direct measures we examined another widely available cultural measure, collectivism, which other scholars have thought to be associated with these traits (e.g., Littlewood, 1999; Gelfand et al., 2002). Although collectivism was indeed found to be associated with the effect of mistake-based feedback, this association disappeared when we controlled for power distance and religiosity. For this reason we tentatively conclude that teachers' authority is the main moderator of the effectiveness of mistake-based feedback.

When drawing conclusions from our study, some important limitations must be acknowledged. First, to measure the use of mistake-based feedback we only had access to a single student questionnaire item with a simple four-step scale (on which the vast majority of students used only the third or fourth step). A more complex measure would have been preferable for two reasons. For one thing, mistake-based feedback is a complex phenomenon, the many nuances of which a single item is unable to capture. For another, a single-item measure will typically have poor reliability. A likely consequence of poor reliability of the MBF measure is that the size of MBF effects on achievement will tend to be underestimated. In other words, with a more reliable measure of MBF we should expect to observe larger effects on achievement.

A second limitation is that our results are purely correlational. Within countries, we have assumed that a certain relation between use of feedback on mistakes and student performance is evidence of the effectiveness of the feedback practice.

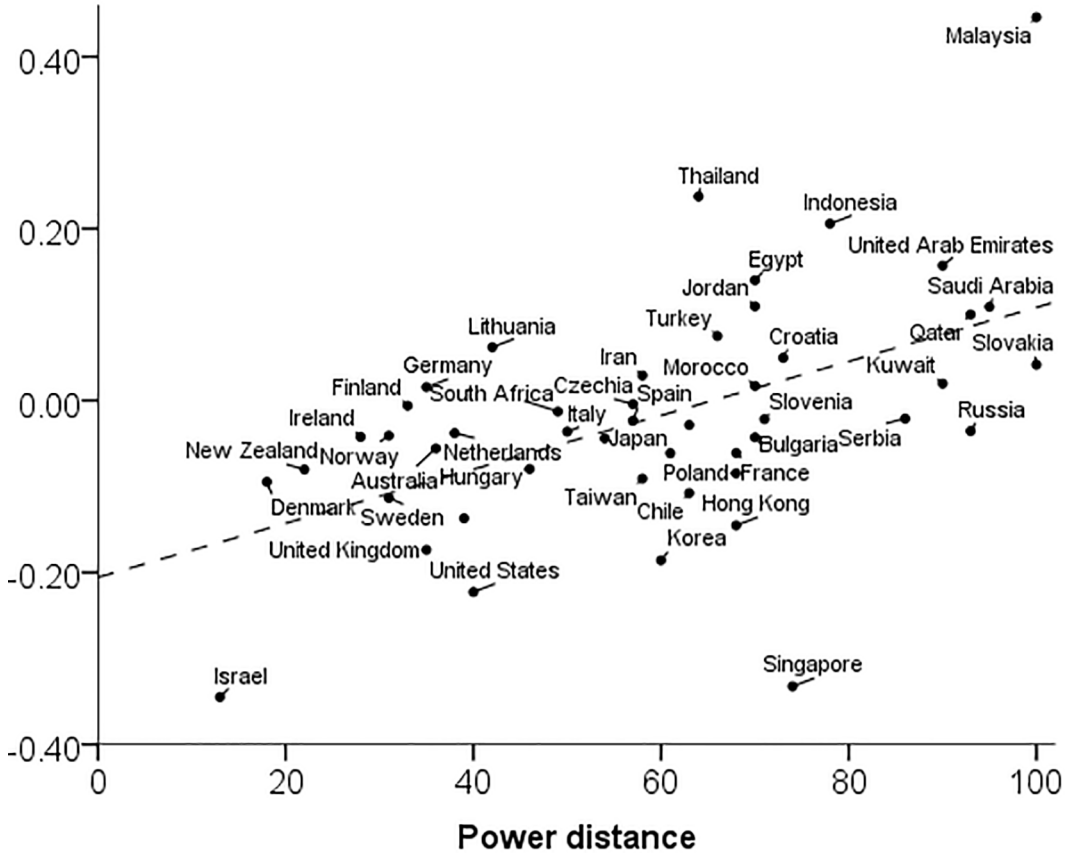

FIGURE 2 | Scatter plot of the aggregate controlled MBF effect against power distance in 47 countries.

An alternative possibility is that associations reflect teachers adapting their teaching practices to the performance level of the student group. Under this alternative interpretation, our between-countries finding would require that teachers respond to higher student performance levels by *increasing* the use of feedback on mistakes in high power distance countries, whereas teachers in low power distance countries would respond to high-performers by *decreasing* their use of such feedback. This interpretation, although equally interesting, seems less plausible to us.

As mentioned above, our simple measure do not allow us to distinguish between different ways of implementing mistake-based feedback. There are many ways of using errors as a springboard for further learning (Borasi, 1994; Boaler, 2016). Thus, it is an open question to what extent the difference in effectiveness between countries lies in teachers implementing mistake-based feedback differently and to what extent it lies in students responding differently to the same feedback. Our findings are consistent with the hypothesis that we gave in the introduction: in countries that are high on power distance and religiosity, young people are more accepting of teachers' authority and therefore more accepting of negative feedback. However, we acknowledge that in the absence of more direct evidence there may be other explanations of the associations we have found.

CONCLUSION

Cultural psychologists have long been interested in how negative feedback may work differently in different cultures. Here we have examined how teachers' feedback on mistakes in math and science class is associated with student achievement in 49 countries. This study differs from classic cross-cultural studies of feedback, both in context and methodology. Still, the finding that feedback on mistakes was associated with better achievement in countries where authority is expected to be more important (namely, countries that are high on power distance and religiosity) was as we expected from prior research. These results highlight the importance of cultural values in educational practice.

DATA AVAILABILITY STATEMENT

The data files used in the analysis can be accessed in the Open Science Framework data repository (https://osf.io/z3h5c/).

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

KE performed the statistical analysis and wrote the manuscript together with JL. All authors contributed to the conception of the study, read and approved the submitted version. OH and AR provided critical revision inputs.

FUNDING

This research was supported by the Swedish Research Council (Grant Number 2014–2008).

ACKNOWLEDGMENTS

KE was grateful to Pontus Strimling for their helpful discussions.

REFERENCES

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48.

Blömeke, S., Olsen, R. V., and Suhl, U. (2016). "Relation of student achievement to the quality of their teachers and instructional quality," in *Teacher Quality, Instructional Quality and Student Outcomes: Relationships Across Countries, Cohorts and Time*, eds T. Nilsen, and J.-E. Gustafsson (Cham: Springer International Publishing), 21–50. doi: 10.1007/978-3-319-41252-8_2

Boaler, J. (2016). *Mathematical Mindsets: Unleashing Students' Potential Through Creative Math, Inspiring Messages and Innovative Teaching*. San Francisco, CA: Jossey-Bass.

Borasi, R. (1994). Capitalizing on errors as "springboards for inquiry": a teaching experiment. *J. Res. Math. Educ.* 25, 166–208.

Bray, W. S. (2011). A collective case study of the influence of teachers' beliefs and knowledge on error-handling practices during class discussion of mathematics. *J. Res. Math. Educ.* 42, 2–38.

Brown, G. T., Kennedy, K. J., Fok, P. K., Chan, J. K. S., and Yu, W. M. (2009). Assessment for student improvement: understanding Hong Kong teachers' conceptions and practices of assessment. *Assess. Educ. Princ. Policy Pract.* 16, 347–363. doi: 10.1080/09695940903319737

Carless, D., and Lam, R. (2014). "Developing assessment for productive learning in Confucian-influenced settings," in *Designing Assessment for Quality Learning*, eds C. Wyatt-Smith, V. Klenowski, and P. Colbert (Dordrecht: Springer), 167–179. doi: 10.1007/978-94-007-5902-2_11

Chen, W. W., and Wong, Y. L. (2015). Chinese mindset: theories of intelligence, goal orientation and academic achievement in Hong Kong students. *Educ. Psychol.* 35, 714–725. doi: 10.1080/01443410.2014.893559

Costa, A., and Faria, L. (2018). Implicit theories of intelligence and academic achievement: a meta-analytic review. *Front. Psychol.* 9:829. doi: 10.3389/fpsyg.2018.00829

Crabtree, S. (2010). *Religiosity Highest in World's Poorest Nations*. Available at: http://news.gallup.com/poll/142727/religiosity-highest-world-poorest-nations.aspx (accessed August 31, 2010).

Dalehefte, I. M., Seidel, T., and Prenzel, M. (2012). "Reflecting on learning from errors in school instruction: findings and suggestions from a Swiss-German video study," in *Human Fallibility*, eds J. Bauer, and C. Harteis (Dordrecht: Springer), 197–213. doi: 10.1007/978-90-481-3941-5_12

Dweck, C. S. (2006). *Mindset: The New Psychology of Success*. New York, NY: Random House.

Dweck, C. S., Chiu, C. Y., and Hong, Y. Y. (1995). Implicit theories: elaboration and extension of the model. *Psychol. Inq.* 6, 322–333. doi: 10.1207/s15327965pli0604_12

Eriksson, K., Helenius, O., and Ryve, A. (2019). Using TIMSS items to evaluate the effectiveness of different instructional practices. *Instr. Sci.* 47, 1–18. doi: 10.1007/s11251-018-9473-1

Fyfe, E. R., and Rittle-Johnson, B. (2017). Mathematics practice without feedback: a desirable difficulty in a classroom setting. *Instr. Sci.* 45, 177–194. doi: 10.1007/s11251-016-9401-1

Gagné, E. D., Crutcher, R. J., Anzelc, J., Geisman, C., Hoffman, V. D., Schutz, P., et al. (1987). The role of student processing of feedback in classroom achievement. *Cogn. Instr.* 4, 167–186. doi: 10.1207/s1532690xci0403_2

Gelfand, M. J., Higgins, M., Nishii, L. H., Raver, J. L., Dominguez, A., Murakami, F., et al. (2002). Culture and egocentric perceptions of fairness in conflict and negotiation. *J. Appl. Psychol.* 87, 833–845. doi: 10.1037//0021-9010.87.5.833

Harks, B., Rakoczy, K., Hattie, J., Besser, M., and Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: the role of feedback's perceived usefulness. *Educ. Psychol.* 34, 269–290. doi: 10.1080/01443410.2013.785384

Hatano, G., and Inagaki, K. (1998). "Cultural contexts of schooling revisited: a review of the learning gap from a cultural psychology perspective," in *Global Prospects for Education, Development, Culture and Schooling*, eds S. G. Paris, and H. M. Wellman (Washington, DC: American Psychological Association), 79–104. doi: 10.1037/10294-003

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112.

Heine, S. J., Kitayama, S., and Lehman, D. R. (2001a). Cultural differences in self-evaluation: Japanese readily accept negative self-relevant information. *J. Cross Cult. Psychol.* 32, 434–443. doi: 10.1177/0022022101032004004

Heine, S. J., Kitayama, S., Lehman, D. R., Takata, T., Ide, E., Leung, C., et al. (2001b). Divergent consequences of success and failure in Japan and North America: an investigation of self-improving motivations and malleable selves. *J. Pers. Soc. Psychol.* 81, 599–615. doi: 10.1037/0022-3514.81.4.599

Hofstede, G. H. (1986). Cultural differences in teaching and learning. *Int. J. Intercult. Relat.* 10, 301–320.

Hofstede, G. H. (2001). *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*. Beverly Hills, CA: Sage.

Holtbrügge, D., and Mohr, A. T. (2010). Cultural determinants of learning style preferences. *Acad. Manag. Learn. Educ.* 9, 622–637. doi: 10.5465/amle.2010.56659880

Joy, S., and Kolb, D. A. (2009). Are there cultural differences in learning style? *Int. J. Intercult. Relat.* 33, 69–85.

Kazemi, E., and Stipek, D. (2001). Promoting conceptual thinking in four upper-elementary mathematics classrooms. *Elem. Sch. J.* 102, 59–80. doi: 10.1086/499693

Littlewood, W. (1999). Defining and developing autonomy in East Asian contexts. *Appl. Linguist.* 20, 71–94. doi: 10.1093/applin/20.1.71

Mangels, J. A., Butterfield, B., Lamb, J., Good, C. D., and Dweck, C. S. (2006). Why do beliefs about intelligence influence learning success? A social cognitive neuroscience model. *Soc. Cogn. Affect. Neurosci.* 1, 75–86. doi: 10.1093/scan/nsl013

Markus, H. R., and Kitayama, S. (1991). Culture and the self: implications for cognition, emotion, and motivation. *Psychol. Rev.* 98, 224–253. doi: 10.1037/0033-295x.98.2.224

Martin, M. O., Mullis, I. V. S., and Hooper, M. (eds) (2016). *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA: Boston College.

Moser, J. S., Schroder, H. S., Heeter, C., Moran, T. P., and Lee, Y. H. (2011). Mind your errors: evidence for a neural mechanism linking growth mind-set

to adaptive posterror adjustments. *Psychol. Sci.* 22, 1484–1489. doi: 10.1177/0956797611419520

National Council of Teachers of Mathematics, (2000). *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.

O'Connell, A. A., and McCoach, D. B. (eds) (2008). *Multilevel Modeling of Educational Data*. Greenwich, CT: Information Age.

Parboteeah, K. P., Hoegl, M., and Cullen, J. B. (2008). Managers' gender role attitudes: a country institutional profile approach. *J. Int. Bus. Stud.* 39, 795–813. doi: 10.1057/palgrave.jibs.8400384

Raz, J. (1990). *Authority*. New York, NY: New York University Press.

Santagata, R. (2004). "Are you joking or are you sleeping?" Cultural beliefs and practices in Italian and US teachers' mistake-handling strategies. *Linguist. Educ.* 15, 141–164. doi: 10.1016/j.linged.2004.12.002

Schleppenbach, M., Flevares, L. M., Sims, L., and Perry, M. (2007). Teacher responses to student mistakes in Chinese and U.S. mathematics classrooms. *Elem. Sch. J.* 108, 131–147. doi: 10.1086/525551

Smith, M. U. (2013). The role of authority in science and religion with implications for science teaching and learning. *Sci. Educ.* 22, 605–634. doi: 10.1007/s11191-012-9469-1

Stevenson, H. W., and Stigler, J. W. (1992). *The Learning Gap: Why Our Schools are Failing and What We Can Learn from Japanese and Chinese Education*. New York, NY: Touchstone.

Thanh Pham, T. H., and Renshaw, P. (2015). Formative assessment in Confucian heritage culture classrooms: activity theory analysis of tensions, contradictions and hybrid practices. *Assess. Eval. High. Educ.* 40, 45–59. doi: 10.1080/02602938.2014.886325

Tulis, M. (2013). Error management behavior in classrooms: teachers' responses to student mistakes. *Teach. Teach. Educ.* 33, 56–68. doi: 10.1016/j.tate.2013.02.003

Wang, T., and Murphy, J. (2004). "An examination of coherence in a Chinese mathematics classroom," in *How Chinese Learn Mathematics: Perspectives from Insiders*, eds L. Fan, N.-Y. Wong, J. Cai, and S. Li (Hackensack, NJ: World Scientific), 107–123. doi: 10.1142/9789812562241_0004

Woodrow, D., and Sham, S. (2001). Chinese pupils and their learning preferences. *Race Ethn. Educ.* 4, 377–394. doi: 10.1080/13613320120096661

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 20 February 2020
doi: 10.3389/fpsyg.2020.00236

The Relation Between Gender Egalitarian Values and Gender Differences in Academic Achievement

Kimmo Eriksson[1,2], Marie Björnstjerna[3] and Irina Vartanova[3]*

[1] School of Education, Culture and Communication, Mälardalen University, Västerås, Sweden, [2] Centre for Cultural Evolution, Stockholm University, Stockholm, Sweden, [3] Institute for Futures Studies, Stockholm, Sweden

Gender differences in achievement exhibit variation between domains and between countries. Much prior research has examined whether this variation could be due to variation in gender equality in opportunities, with mixed results. Here we focus instead on the role of a society's values about gender equality, which may have a more pervasive influence. We pooled all available country measures on adolescent boys' and girls' academic achievement between 2000 and 2015 from the Program for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS) assessments of math, science, and reading. We then analyzed the relation between gender differences and country levels of gender egalitarian values, controlling for country levels of living standards and indicators of gender equality in opportunities. Gender egalitarian values came out as the most important predictor. Specifically, more gender egalitarian values were associated with improved performance of boys relative to girls in the same countries. This pattern held in reading, where boys globally perform substantially worse than girls, as well as in math and science where gender differences in performance are small and may favor either boys or girls. Our findings suggest a previously underappreciated role of cultural values in moderating gender gaps in academic achievement.

Keywords: gender egalitarian values, gender equality, gender differences, academic achievement, mathematics education, literacy abilities

OPEN ACCESS

Edited by:
Jiesi Guo,
Australian Catholic University, Australia

Reviewed by:
Erin Michelle Fahle,
St. John's University, United States
Pey-Yan Liou,
National Central University, Taiwan

****Correspondence:***
Kimmo Eriksson
kimmo.eriksson@mdh.se;
kimmoe@gmail.com

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *05 August 2019*
Accepted: *31 January 2020*
Published: *20 February 2020*

Citation:
Eriksson K, Björnstjerna M and Vartanova I (2020) The Relation Between Gender Egalitarian Values and Gender Differences in Academic Achievement. Front. Psychol. 11:236. doi: 10.3389/fpsyg.2020.00236

INTRODUCTION

The rise in gender egalitarian values in industrialized and postindustrial societies has had wide-ranging effects (Inglehart and Norris, 2003). It is now a nearly universal phenomenon that girls outperform boys in school (Voyer and Voyer, 2014) and that gender differences in grade progression and school dropout favor girls—even in developing countries (Grant and Behrman, 2010). Many researchers are concerned about boys falling behind in school (Morris, 2012; DiPrete and Buchmann, 2013). However, gender differences in achievement vary with academic domains. This is especially well documented with respect to performance on standardized tests in the domains of reading and mathematics; the average girl always outperforms the average boy at reading, but in some countries, the opposite relation holds for math (Stoet and Geary, 2013). In other words, the gender difference in reading performance consistently favors girls, whereas the direction of the gender difference in math performance varies between countries (and may also

vary over time within a country). In this paper, we will always calculate gender differences by taking the achievement of boys minus the achievement of girls, and we shall refer to this difference as *the relative achievement of boys vs. girls*. A negative value of the relative achievement of boys vs. girls thus signifies a gender difference that favors girls.

Could it be that the relative achievement of boys vs. girls in a country depends on the level of gender equality? This idea has been around for several decades (Baker and Jones, 1993). It has been examined in a number of studies using cross-national datasets on achievement on standardized tests, such as Program for International Student Assessment (PISA). A ground-breaking study by Guiso et al. (2008) found that the relative math achievement of boys vs. girls in 2003 PISA was negatively associated with country-level indicators of gender equality in opportunities. However, this finding has not been replicated in other waves of PISA (Stoet and Geary, 2013). Results vary across waves due to varying country samples as well as sampling error within countries. From the total evidence, it is unclear whether there is any robust association between the relative math achievement of boys vs. girls and gender equality in opportunities.

The purpose of this paper is to shift attention from gender equality in opportunities to gender egalitarian values. Whereas the former refers to outcome measures such as the relative participation of women and men in the workforce or in politics, the latter refers to cultural attitudes regarding the value of the genders: Are women and men equally important? Are they equally competent? Should they have the same rights? We will argue that such cultural values are likely to have a direct influence on the academic efforts of adolescent girls and boys, thereby shaping gender differences in achievement.

The potential role of gender egalitarian values has largely been neglected in prior research on gender differences in achievement. In the above-mentioned analysis of 2003 PISA data, a measure of gender egalitarian values was included alongside a measure of gender equality in opportunities (Guiso et al., 2008). Both measures were found to be negatively associated with the relative math achievement of boys vs. girls, but the effect was not statistically significant for the measure of gender egalitarian values. Subsequent research has focused exclusively on the effect of gender equality in opportunities.

There are two reasons why we decided to revisit the role of gender egalitarian values. First, there are theoretical reasons to believe that gender egalitarian values may affect the schoolwork of girls and boys. Second, there are empirical reasons to doubt the robustness of the original findings. Subsequent research has shown that patterns of gender differences in achievement in the 2003 PISA dataset do not tend to replicate in other waves of PISA (Stoet and Geary, 2013).

Our Methodological Approach

An important question is why different waves of PISA and Trends in International Mathematics and Science Study (TIMSS) would yield differing patterns of gender differences in achievement. The time between consecutive waves of PISA and TIMSS is just 3 and 4 years, respectively. It seems unlikely that gender differences change much in such a short time. However, the *measures* of gender differences may still change due to sampling error. The size of student samples in these assessments is usually around 5000 per country, see the official reports of the methods used in TIMSS and PISA (Mullis et al., 2009; Schleicher et al., 2009). These are large samples and consequently the standard errors of mean scores are small—but not negligible. Descriptive statistics of data from TIMSS and PISA are easily available from the data explorer service provided by the National Center for Education Statistics[1]. On the scoring scale that is used (which has a global mean of 500), the standard error for country measures of gender differences is around four points. At the same time, the standard deviation of gender differences *across* countries is only around 10 points. Thus, although the measurement error at the country level is small, it is sufficiently large for an estimation of between-country variation in a single wave to be unreliable. Another issue is that the *sample of countries* choosing to participate in PISA or TIMSS varies between waves. These samples are of limited size (e.g., 40 countries in 2003 PISA) and not representative of all countries that potentially could participate. A correlation may come out quite differently when estimated in different non-representative samples. Both these sampling issues can be alleviated by pooling the data from many years. We then obtain a larger sample of countries as well as several measures of gender differences in each country. These can then be analyzed by multi-level methods, nesting country-years in countries.

Background

Gender Differences in Academic Achievement

There is an extensive empirical literature on gender differences in academic achievement. While girls have historically often been disadvantaged, they are now surpassing boys in rate of school enrollment and grade completion even in many developing countries (Grant and Behrman, 2010). In modern times, the big picture is that girls tend to do better than boys in school (Voyer and Voyer, 2014), with differences tending to be more pronounced in minority groups, in urban areas, and among students from families with low socioeconomic status (Morris, 2012). To explain these observations, several theories about the impact of individual and contextual factors have been put forward. Although men and women seem to be similar on most psychological variables (Hyde, 2005), it has been argued that girls on average have superior performance on some behavioral skills that are of importance for academic success, such as self-discipline (Duckworth and Seligman, 2006). Boys, on the other hand, more often express aggressive behaviors and display more developmental difficulties and negative attitudes toward learning (Zill and West, 2001; Lansford et al., 2012). It could therefore be argued that girls, in general, more easily adjust to the school environment. Gender norms have also been suggested to play a role; specifically, aspects of expressed masculinity might negatively affect boys' achievement (Morris, 2012).

Gender differences in academic performance vary somewhat from kindergarten and up to high school (Robinson and Lubienski, 2011). In the present paper, we focus on gender

[1] nces.ed.gov

differences among adolescents, where the best data is available. Importantly, gender differences also depend on whether performance is measured in terms of grades awarded in school or in terms of scores on standardized tests. In brief, boys are at a clear disadvantage when it comes to grades awarded, but perform relatively better at achievement tests (Duckworth and Seligman, 2006). If we look at achievement tests in mathematics, boys even perform somewhat better than girls in many countries; at reading tests, however, girls seem to outperform boys everywhere (Stoet and Geary, 2013). Our focus in this paper is on standardized tests in reading, math, and science.

How Gender Differences in Achievement Vary Between Countries

Most studies of the above-mentioned research on gender differences in school performance have taken place in western countries, especially in the United States. However, there is also an extensive literature on how gender differences vary between countries. For instance, a comprehensive cross-national meta-analysis on gender differences in school grades found that the size of the gender gap depended on whether studies covered North America, or Scandinavia, or the rest of the world (Voyer and Voyer, 2014). Studies where gender differences are measured using the same method in many countries are particularly informative. The foremost example is PISA, a large-scale assessment that has been conducted with a new wave every three years since 2000. PISA tests mathematics, science, and reading literacy among representative samples of 15-year-olds, mainly in OECD countries.

The first two major studies of gender differences in reading and mathematics based on PISA data came out the same year but used data from different waves, either the 2000 wave (Marks, 2008) or the 2003 wave (Guiso et al., 2008). Both studies found that the levels of relative achievement of boys vs. girls in mathematics and reading were very strongly correlated (i.e., in those countries where boys did particularly well in math relative to girls, they also did particularly well in reading). However, the studies reached somewhat different conclusions with respect to the role of gender equality. The first study examined the proportion of women in the workforce and found that it was "not associated with the gender gaps in mathematics" (Marks, 2008, p. 89). The second study examined the same measure of female participation in the workforce, as well as a measure of female political empowerment and a measure of gender egalitarian values, and concluded that "the gender gap in math scores disappears in countries with a more gender-equal culture" (Guiso et al., 2008, p. 1164). The latter conclusion was upheld also in a later analysis of the same 2003 PISA dataset (Else-Quest et al., 2010). However, it was not supported in subsequent analyses of the 2009 PISA data (Kane and Mertz, 2012; Reilly, 2012). Later detailed studies of all four waves of PISA from 2000 to 2009 confirmed that different waves yield results that point in different directions with respect to the role of gender equality (Stoet and Geary, 2013, 2015).

Another source of cross-national data on achievement in math and science is TIMSS. TIMSS tests eighth graders on what is covered by curricula, with a new wave every fourth year since 1995. Also research on TIMSS data has given a mixed picture of the relation between gender equality and gender differences in achievement. Kane and Mertz (2012) found a measure of gender equality to correlate either positively or negatively with the relative math achievement of boys vs. girls, depending on whether the 2003 or 2007 TIMSS data were used. Other researchers using the 2011 TIMSS data found a null correlation (Reilly et al., 2019).

In sum, prior research suggests two important conclusions on how gender differences in math achievement vary between countries. First, gender differences in math achievement seem to be robustly linked to gender differences in reading achievement. Second, there does *not* seem to be a robust link between gender differences in academic achievement and country levels of gender equality.

Theories and Research Questions

Theories on How Gender Equality Could Shape Gender Differences in Achievement

There are several theories on how the level of gender equality in a society could influence the gender difference in academic achievement. One idea is that gender segregation, specifically with respect to job opportunities, influence students' motivation. This idea, known as the gender stratification hypothesis, has mainly been applied to achievement in mathematics: female students may do less well in math if there are less opportunities to jobs that require math skills for women than for men (Baker and Jones, 1993; Else-Quest et al., 2010). An alternative idea is that gender equality and prosperity are conditions that give individuals more freedom to pursue their intrinsic interests. To the extent that boys have a greater intrinsic interest in mathematics than girls do, boys would then perform relatively well at math in countries with greater gender equality and prosperity (Stoet et al., 2016). Note that these ideas lead to opposite predictions about the relation between gender equality and gender differences in math achievement, but none of them seem to account for the robust association between gender differences in reading and gender differences in mathematics achievement. A third idea is that gender differences in academic achievement (both in reading and math) may vary due to policy differences with respect to gender and education (Marks, 2008). However, gender differences in enrollment do not consistently predict gender differences in achievement (e.g., Else-Quest et al., 2010). In sum, none of the three ideas has received consistent support across the published analyses of PISA and TIMSS data.

The Case for Examining Gender Egalitarian Values

The above-mentioned theories all focus on gender equality in opportunities. A society's underlying values with respect to gender equality may be an alternative driver of boys' and girls' relative academic achievement (Nollenberger et al., 2016). Here we elaborate on why, taking as our starting point Bronfenbrenner's ecological perspective on human development (Bronfenbrenner, 1979, 2005). In brief, Bronfenbrenner's model views children's development as shaped through a dynamic interaction between the child and its environment. The environment has multiple layers, from the closest layer consisting of the child's family, friends, and teachers to the outermost

layer consisting of the society and culture in which the child lives. The surrounding culture may influence a child's academic achievement through its influence on how the child interacts with family, friends, and teachers. We shall now outline various pathways by which gender egalitarian values could influence the relative achievement of boys vs. girls. We distinguish mechanisms that would yield a positive influence from those that would yield a negative influence. Our data will not allow us to test any of these mechanisms; the point of this section is demonstrating that there are plausible ways in which academic achievement could be influenced by cultural values rather than by expectations of future opportunities.

Attitudes toward gender could play a major role in what *parents* allow their children to do. This may be important for academic achievement, because students may put less effort into their studies if they have extra-curricular activities such as jobs, sports, and dating (Stevenson et al., 1993). Parental constraints on out-of-home activities for adolescents may be stricter for girls than for boys, in particular in cultural groups with less gender egalitarian values (Carrington et al., 1987). As values become more gender egalitarian we expect parental constraints on girls to become more relaxed. As the range of extra-curricular activities that girls are allowed to take part in increases, they will have less time and attention to spare for schoolwork. This would be a mechanism whereby more gender egalitarian values would have a negative effect on girls' academic achievement, and hence a positive effect on the relative achievement of boys vs. girls.

We now turn to the influence of *peers.* An often mentioned factor behind boys' underachievement in school is a social norm that putting effort into studying is regarded as "uncool" (Morris, 2012). This attitude is most common among boys, but it is also found among girls (Jackson and Nyström, 2015). An effect of gender egalitarian values could be that attitudes to studying become more similar between the genders, so that studying would be more "uncool" among girls in more gender egalitarian countries. As such negative attitudes to studying are likely to negatively affect achievement, this would amount to a positive effect of gender egalitarian values on the relative achievement of boys vs. girls.

Moreover, students' achievement tends to be positively influenced by high expectations from their *teachers* (Jussim and Eccles, 1992). Teachers may hold various forms of gender bias in their expectations of students and in the various ways they interact with students (Frawley, 2005; Berekashvili, 2012). To the extent that teachers hold stereotypes that boys are rowdy and that girls are neat and disciplined, this could influence student achievement by lowering the relative achievement of boys vs. girls. More gender egalitarian values could reduce such gender stereotypes and thereby have a positive effect on the relative achievement of boys vs. girls.

Gender egalitarian values could also have a negative effect on the relative achievement of boys vs. girls. Specifically, parents with non-egalitarian attitudes about gender could have a lower interest in the future career of their girls than of their boys. Similarly, teachers with non-egalitarian attitudes about gender could think that girls do not need the highest levels of academic skills. If so, an increase in gender egalitarian values may lead to an increase in the encouragement that girls receive to excel in school (and hence a negative effect on the relative achievement of boys vs. girls).

Questions Addressed in the Present Research

The PISA and TIMSS assessments do not include measures of the gender egalitarian values of students' family, friends, and teachers. We therefore cannot examine the specific pathways discussed above. Instead, we test the broader implication that country differences in gender egalitarian values may create country differences in the relative achievement of boys vs. girls. We do this by examining whether these variables are correlated at the country level.

In light of the lack of robustness of earlier findings on moderators of country differences in the relative achievement of boys vs. girls, we are especially interested in examining whether the moderating effect of gender egalitarian values is robust. We study robustness along the following dimensions.

Effects on gender differences vs. gender differences in effects

Our main question is whether gender egalitarian values are associated with gender differences in achievement. However, we are also interested in how gender egalitarian values are associated with boys' and girls' absolute levels of achievement. For instance, a positive effect on the relative achievement of boys vs. girls could arise either from a positive effect on boys' achievement or from a negative effect on girls' achievement.

Different sources of achievement scores

To explain that their analyses of 2003 PISA data and 2003 TIMSS data yielded inconsistent findings, Else-Quest et al. (2010) speculated that this might be due to the fact that PISA and TIMSS differ somewhat in the aim of the tests, in which case we should expect results from PISA and TIMSS data to be robustly different from each other. An alternative possibility is that the observed inconsistency was spurious, in which case we should expect findings from the two data sources to be generally consistent.

Mean vs. 90th percentile achievement scores

Although much research on gender differences in achievement focuses on the mean achievement levels of boys and girls, it is well-known that gender differences vary over the performance continuum (Robinson and Lubienski, 2011). Globally, the largest gender differences in mathematics performance tend to be found at the high end; by contrast, the high end in reading performance exhibits the smallest gender differences (Stoet and Geary, 2013). The high end of the performance continuum is of interest also because high achievers are particularly likely to enroll in higher education (Lubinski and Benbow, 2006). For these reasons, we will examine gender differences in achievement both at mean levels and at the 90th percentile. Measures of the mean achievement and 90th percentile achievement of boys and girls are provided by both PISA and TIMSS.

Different sources of gender egalitarian values scores

The concept of gender egalitarian values has been measured in various ways. As detailed in Section "Materials and Methods," we shall use two different sources of data on how gender egalitarian

values vary across countries: the World Values Survey (WVS) and the GLOBE project.

Controlling for other country-level variables

Whereas prior work has focused on gender equality in opportunities, the premise of the current work is that gender egalitarian values may exert a stronger influence on gender differences in achievement. Of course, gender egalitarian values and gender equality in opportunities are not independent of each other (Brandt, 2011). When analyzing the effect of the former variable, it is important to check that the results are robust to controlling for the latter variable. Further, both these aspects of gender equality are correlated with the standard of living in the country, which is in itself an important predictor of student achievement (Stoet and Geary, 2013). We therefore also control for measures of standard of living.

MATERIALS AND METHODS

Countries were included in this study if they satisfied three criteria: the country had participated at least once in PISA or TIMSS (excluding benchmark and off-grade participants), so that gender gaps in achievement could be calculated; gender equality measures for the country were available from the World Economic Forum, see below; data on gender egalitarian values in the country were available either from the WVS or the GLOBE project, see below. **Supplementary Table S1** lists the 74 countries included in the study and indicates for which countries data were available from PISA, TIMSS, WVS, and GLOBE. The dataset can be accessed in the Open Science Framework data repository[2].

Girls' and Boys' Achievement Levels on PISA and TIMSS Tests Since 2000

PISA is an international assessment of 15-year-old students' achievement in math, reading, and science (Schleicher et al., 2009). It is conducted by the Organization for Economic Co-operation and Development (OECD). PISA uses a representative sample of students from each participating country. Sample sizes are usually around 5000 per country but sometimes considerably larger. Data are available from six data collections: 2000, 2003, 2006, 2009, 2012, and 2015. Test scores are normalized, with a mean of 500 and a standard deviation of 100. Due to details of the design, comparability between different waves depends on the subject: whereas reading scores are comparable for all waves since 2000, math scores are only comparable from 2003 and onward, and science scores are only comparable from 2006 and onward. For these waves with comparable scores (six for reading, five for math, and four for science), the mean score and the score at the 90th percentile, calculated separately for boys and girls in each country, were downloaded from the National Center for Education Statistics[3]. Scores were obtained for 63 countries in our study. From the same source, we also downloaded the percentage of boys among participants in each country.

TIMSS is a similar international assessment of math and science achievement of students in the eighth grade (as well as the fourth grade, which is not used here), with most participants being about 14 years old. The assessment is conducted by the International Association for the Evaluation of Educational Achievement (IEA). Details on the design and execution of the TIMSS assessment are provided in reports from the IEA (e.g., Mullis et al., 2009). Although TIMSS differs from PISA in several ways, important similarities include that both use representative samples of similar sizes and that test scores are normalized in the same way. Thus, absolute levels and gender differences in country mean scores are roughly comparable between PISA and TIMSS. Since 2000, there have been four waves of TIMSS: 2003, 2007, 2011, and 2015. For these waves, the mean score and the score at the 90th percentile, calculated separately for girls and boys, were downloaded from the National Center for Education Statistics[3]. Data on scores and percentage of boys were obtained for 51 countries in our study.

Gender Egalitarian Values

The WVS is a survey of human beliefs and values that has been conducted in six waves since 1981 by a global network of social scientists. Every wave is conducted over a period of 5 years and the sample of participating countries changes for every wave. Waves 3–6, conducted during 1994–2014, all included an index for gender egalitarian values called Equality (Welzel, 2013), which is based on three items:

> *Jobs:* When jobs are scarce, men should have more right to a job than women.
> *Politics:* On the whole, men make better political leaders than women do.
> *University:* University is more important for a boy than for a girl.

The equality index is coded such that higher values of the index represent more egalitarian responses. We downloaded the full set of data from the WVS website[4]. We pooled waves 3–6 and calculated the country means of the equality index. WVS measures of gender egalitarian values were available for 67 countries in our study.

An alternative to the WVS is provided by the GLOBE project, which measured cultural practices and cultural values in societies across the world in the mid-1990s (House et al., 2004). For each participating country, the GLOBE project reported cultural values on nine dimensions, one of which is gender egalitarianism. The country index of gender egalitarian cultural values is based on survey responses to several attitude items on how society should be with respect to gender equality in education and leadership (e.g., "I believe that boys should be encouraged to attain a higher education more than girls," "I believe that opportunities for leadership positions should be more available for men than for women"). The value of the composite index for each country was downloaded from the project website[5]. GLOBE

[2] https://osf.io/v7bqt/

[3] nces.ed.gov

[4] worldvaluessurvey.org/wvs.jsp

[5] globeproject.com

measures of gender egalitarian values were available for 47 countries in our study. Data on separate items are not available.

Control Variables

Gender Equality Indicators From the World Economic Forum

The Global Gender Gap Index (GGI) is a composite measure of gender equality published by the World Economic Forum every year since 2006. It has been used in many studies of the link between gender equality and math achievement (Guiso et al., 2008; Hyde and Mertz, 2009; Else-Quest et al., 2010; Stoet and Geary, 2013, 2015). The GGI is based on four component scores: "Economic participation and opportunity," "Educational attainment," "Health and survival," and "Political empowerment." All scores have a theoretical range from 0 to 1. Details on the GGI and its component scores are provided by the World Economic Forum (Hausmann et al., 2009). GGI scores for all available years were downloaded from https://tcdata360.worldbank.org.

The Human Development Index From the United Nations

The Human Development Index (HDI) is a composite measure of the standard of living in a country that is known to be a strong predictor of academic achievement levels (Stoet and Geary, 2013, 2015). The HDI is based on measures of people's life expectancy at birth, education (expected and mean years of schooling), and income (GNI per capita). For details on the construction of the HDI and its component measures, see the report by the United Nations Development Program (UNDP, 2015). From their website[6], we downloaded country scores of the HDI and its components for all years from 2000 and onward.

Ethical Approval

No institutional ethical approval was necessary for carrying out this secondary data analysis of publicly available and fully anonymized datasets.

Analysis

Our main analytic approach is to nest country-years in countries, using a two-level analysis ("mixed model") on the form:

$$Y_{tc} = \beta_0 + \beta_1 GV_c + \beta_2 HDI_{tc} + \beta_3 GGI_{tc} + \beta_4 PercB_{tc} + \beta_5 Yr_t + u_c + e_{tc}$$

In the first set of analyses, the dependent variable denoted by Y_{tc} is the absolute achievement levels of boys and girls in country c in year t; in the second set of analyses, the dependent variable is instead the difference between these achievement levels (i.e., the relative achievement of boys vs. girls). Every predictor is centered at the global mean and scaled by standard deviation. The dependent variable is not standardized. We report unstandardized regression coefficients, which tell us effects in terms of the number of achievement score points by which the dependent variable increases when the predictor increases by one standard deviation.

[6]hdr.undp.org

The predictors are: the country's gender egalitarian values (either WVS or GLOBE), denoted by GV_c; the country's prosperity in the given year[7], denoted by HDI_{tc}; the country's level of gender equality in opportunities in the given year (or the closest year available[8]), denoted by GGI_{tc}; the gender composition of students taking the test in the country in the given year (in terms of the percentage of boys in the sample), denoted by $PercB_{tc}$; and Yr_t is the year of measurement. The country random intercept u_c and the error term e_{tc} are normally distributed with mean 0 and standard deviation σ_u^2 and σ_e^2.

Mixed model analyses were conducted in the lme4 package (Bates et al., 2014) in R, using restricted maximum likelihood (REML) estimation. Because the dependent variables are estimates of country levels of achievement based on random samples, we use an alternative model weighted by the inverse of the standard errors, normalized so that the sum of the weights is equal to the sample size. For each model, we estimate marginal pseudo R squared, which shows how much variance is explained by the fixed effects (Nakagawa et al., 2017). As some measures were not normally distributed, we estimate 95% confidence intervals based on 2.5 and 97.5 percentiles from 1000 bootstrap samples.

RESULTS

Descriptive statistics for all variables are reported in **Table 1**.

Using Gender Egalitarian Values to Predict the Achievement Levels of Boys and Girls

Our first set of analyses predicted the mean and 90th percentile achievement scores of boys and girls from the gender egalitarian values in the country. Five separate cases of dependent variables were analyzed: achievement on the three PISA tests in reading, math, and science, as well as achievement on the two TIMSS tests in math and science. Moreover, each analysis was conducted two times using gender egalitarian values either from WVS or from GLOBE. Overall, the fixed effects explain around 55% of the observed variation in achievement levels. Most of the unexplained variation was on the country level (average SDc = 36.8) rather than on the country-year level (average SDres = 14.7).

The estimated effect of gender egalitarian values in each analysis is presented graphically in **Figure 1**; the exact numbers can be found in **Supplementary Table S2**. The first thing to note is that, with a single exception, none of the estimated effects of gender egalitarian values was significantly different from zero. Nonetheless, note the consistency in the difference between the estimated effects on boys' and girls' achievement. The gender

[7]Lebanon is missing information for the education subindex of the HDI for the years 2000 and 2003; we used the value from 2005 to reestimate the HDI index for those years.

[8]GGI are generally available from 2006, so for 2000 and 2003, it is the 2006 data that are used. For some countries, however, GGI data start later than 2006; for earlier years, we then use data from the earliest year available, which was 2007 for Azerbaijan, Armenia, Vietnam, and Qatar, 2010 for Lebanon, and 2012 for Serbia.

TABLE 1 | Descriptive statistics.

Assessment	Domain	Measure	Nobs	Nc	Nobs/Nc	Mean	*SD*
PISA	Math	Boys, averages	244	63	3.87	469.9	54.98
		Boys, percentiles	244	63	3.87	588.96	58.06
		Relative achievement, averages	244	63	3.87	7.83	8.97
		Relative achievement, percentiles	244	63	3.87	16.13	9.31
		Girls, averages	244	63	3.87	462.07	53.01
		Girls, percentiles	244	63	3.87	572.83	55.85
	Reading	Boys, averages	279	63	4.43	447.42	52.13
		Boys, percentiles	279	63	4.43	569.43	51.97
		Relative achievement, averages	279	63	4.43	−36.46	12.67
		Relative achievement, percentiles	279	63	4.43	−25.53	10.21
		Girls, averages	279	63	4.43	483.88	51.14
		Girls, percentiles	279	63	4.43	594.96	50.66
	Science	Boys, averages	212	63	3.37	468.88	53.86
		Boys, percentiles	212	63	3.37	589.66	58.98
		Relative achievement, averages	212	63	3.37	−2.19	11.01
		Relative achievement, percentiles	212	63	3.37	7.11	10.49
		Girls, averages	212	63	3.37	471.06	50.48
		Girls, percentiles	212	63	3.37	582.54	55.19
TIMSS	Math	Boys, averages	134	51	2.63	464.83	71.18
		Boys, percentiles	134	51	2.63	576.15	67.48
		Relative achievement, averages	134	51	2.63	−1.76	11.13
		Relative achievement, percentiles	134	51	2.63	5.02	9.55
		Girls, averages	134	51	2.63	466.59	70.4
		Girls, percentiles	134	51	2.63	571.13	67.18
	Science	Boys, averages	134	51	2.63	474.25	68.03
		Boys, percentiles	134	51	2.63	585.19	57.32
		Relative achievement, averages	134	51	2.63	−1.65	17.34
		Relative achievement, percentiles	134	51	2.63	6.2	12.93
		Girls, averages	134	51	2.63	475.89	64
		Girls, percentiles	134	51	2.63	578.98	54.72
PISA		% Boys	280	63	4.44	0	1
TIMSS		% Boys	134	51	2.63	0	1
WVS			338	67	5.04	0	1
GLOBE			265	47	5.64	0	1
GGI			378	74	5.11	0	1
HDI			378	74	5.11	0	1
Year			378	74	5.11	0	1

Nobs refers to the total number of observations (one observation for every year in which a country has participated in the assessment). Nc refers to the number of countries on which there are observations. Nobs/Nc is the average number of observations per country. Mean and SD refer to the mean and standard deviation of the observed scores.

difference in the estimated effect of gender egalitarian values consistently favored boys' achievement. Averaging over all 20 analyses (using the weighted model), the estimated effect of gender egalitarian values on boys' and girls' achievement was 2.7 and −2.6, respectively, yielding a difference of 5.2 points. We briefly summarize the effects of the covariates by similarly calculating their average effects across the twenty analyses: HDI had a large positive effect on achievement, almost identical for boys and girls (41.0 vs. 41.1). The effects of the other covariates were smaller but still almost identical for boys and girls (GGI: 6.5 vs. 6.7; % boys in sample: 1.9 vs. 1.7; year: −8.3.1 vs. −7.9).

In sum, these analyses suggest two things. First, more gender egalitarian values may on the whole be less beneficial for girls' achievement than for boys' achievement. Second, although prosperity (HDI) and gender equality in opportunities (GGI) came out as more important determinants of countries' achievement levels, on the whole these variables seem to be equally beneficial for girls' and boys' achievement. Our second set of analyses will provide further illumination of these patterns.

Using Gender Egalitarian Values to Predict the Relative Achievement of Boys vs. Girls

The above analyses of absolute achievement levels of boys and girls suggested that higher levels of gender egalitarian values

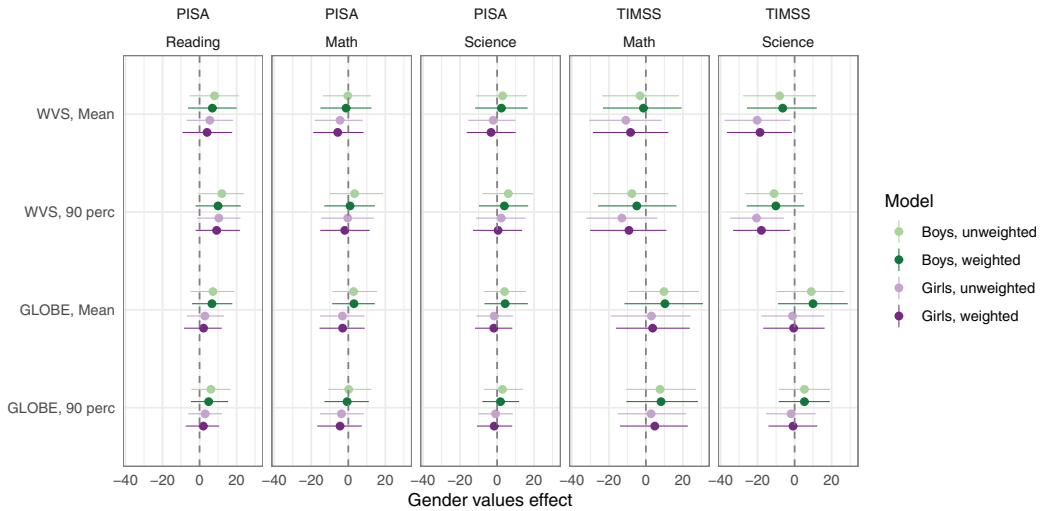

FIGURE 1 | Estimated effects of gender egalitarian values on the mean and 90th percentile achievement of boys and girls. *Note*. The figure shows fixed effects of gender egalitarian values on the mean and 90th percentile of boys' (green) and girls' (violet) achievement, with bootstrapped 95% CI for weighted (dark) and unweighted (light) models. At the country level, all models included random intercepts. Entries estimate the number of points by which achievement scores increase when the gender egalitarian values increase by one standard deviation, holding all covariates constant. Different analyses are based on data from different samples of countries: 56 countries had both PISA and WVS data (total number of observations: n = 242 for reading, n = 213 for math, n = 186 for science), 40 countries had both PISA and GLOBE data (n = 203 for reading, n = 175 for math, n = 147 for science), 49 countries had both TIMSS and WVS data (n = 129), and 32 countries had both TIMSS and GLOBE data (n = 90). Standard deviations for the country random intercept ranged between 23.0 and 52.7 (mean = 36.8), and for residuals between 10.1 and 22.3 (mean = 14.7). The average marginal R squared was 0.55 (SD = 0.09, min = 0.38, max = 0.73).

may be associated with higher relative achievement of boys vs. girls. In the second set of analyses, we examined this relation directly, by using the relative achievement of boys vs. girls as the dependent variable. All fixed effect estimates are presented graphically in **Figure 2**; the exact numbers can be found in **Supplementary Table S3**.

To understand what **Figure 2** is saying, first consider the intercepts. Because predictors are centered at their means, intercepts represent the estimated relative achievement of boys vs. girls when all predictors take their mean values. The figure shows substantial negative intercepts for reading, reflecting the global tendency for reading achievement to be lower among boys than among girls.

Our focus is the estimated effects of gender egalitarian values. Note that these were all positive and statistically significant. Averaged across all 20 analyses, an increase in the level of gender egalitarian values by one standard deviation was associated with an estimated increase of 6.0 points in the relative achievement of boys vs. girls. This finding dovetails nicely with our first set of analyses, which indicated that one standard deviation higher gender egalitarian values was on average 5.2 points more beneficial for boys than for girls.

Now consider the effects of the other predictors. Inspection of **Figure 2** shows that, in every analysis, the effect of gender egalitarian values was larger than the effect of any other predictor. Moreover, the directions of the estimated effects of prosperity (HDI) and gender equality in opportunities (GGI) were inconsistent, sometimes positive and sometimes negative. The gender distribution of the sample was never a significant predictor. The estimated time trends had different signs in TIMSS and PISA for the same academic domains, suggesting that they lack reliability; we return to this issue in Section "Lack of Robustness of Estimates of How Gender Differences in Achievement Change Over Time."

The fixed effects explain more substantial proportions of the variation in relative achievement in the domains of science (32% on average) and math (22% on average) than in the domain of reading (10% on average). As reported in **Figure 2**, most of the explained variance could be attributed to the effect of gender egalitarian values alone. The unexplained variation was roughly equally distributed between the country level (average SDc = 7.5) and the country-year level (average SDres = 6.1).

Using Domain-Specific Covariates Instead of Indexes

In the above analyses of relative achievement, we used index measures of prosperity (HDI), and gender equality in opportunities (GGI) as covariates. These indexes are based on component measures from different domains. It is possible that domains matter, so that the use of indexes miss important aspects of what is actually going on. We therefore reran the analyses with the HDI and GGI indexes replaced by their domain-specific components (seven in total). The results were similar to the previous analyses. Across all 20 analyses, an increase in the level of gender egalitarian values by one standard deviation was associated with an estimated increase of ranging from 1.7 to 8.7 points (mean 4.4) in the relative mean achievement of boys vs. girls, whereas the covariates showed no robust effects. Thus, the effect of gender egalitarian values was not accounted for

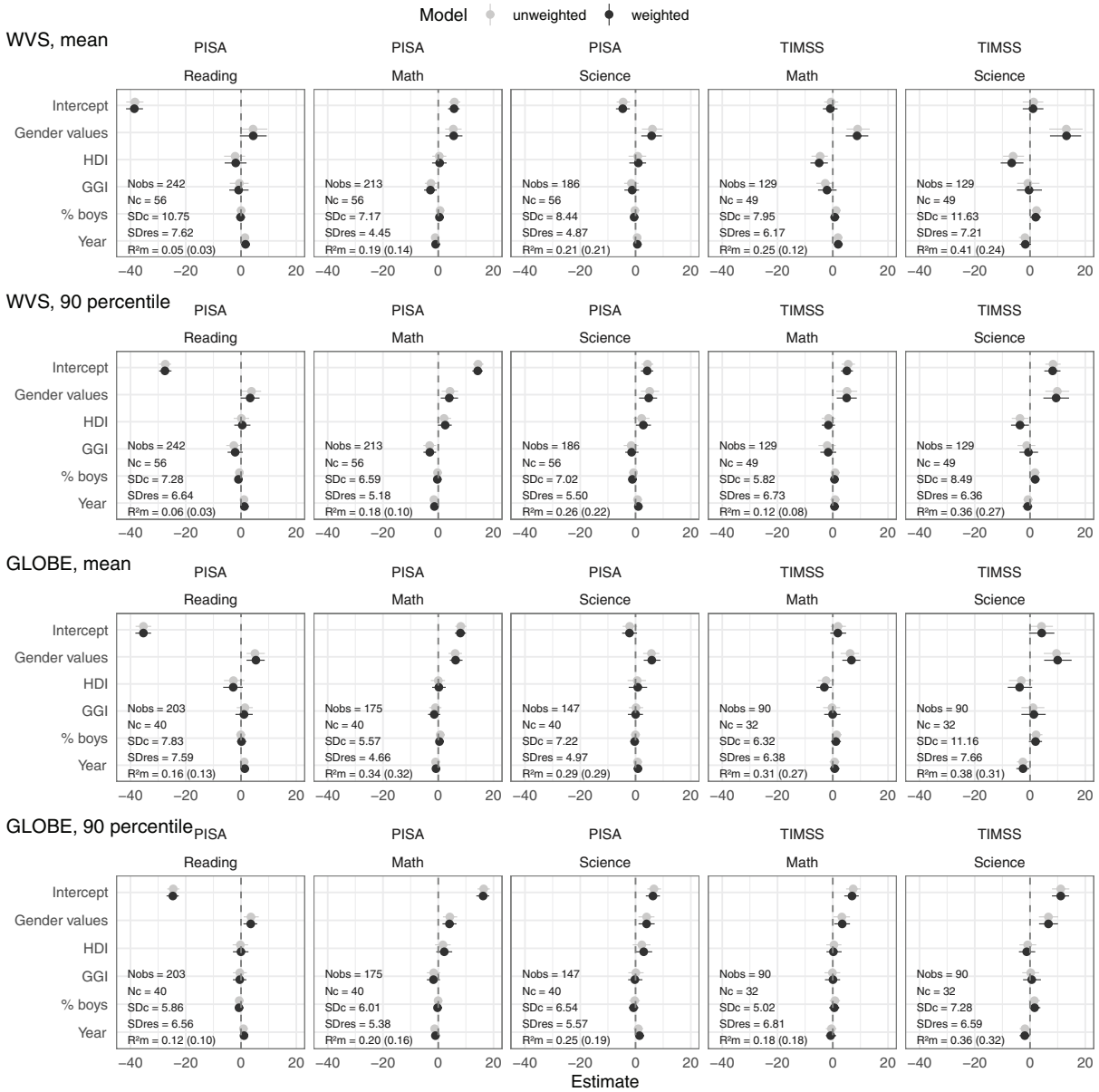

FIGURE 2 | Mixed effects models of gender egalitarian values and covariates on the mean and 90th percentile relative achievement of boys vs. girls. *Note*. The figure shows fixed effects with 95% bootstrap CI for weighted (dark) and unweighted (light) models. For the weighted models, we display the standard deviations of the random country intercept (SDc) and of residuals (SDres), as well as marginal R-squared (R2m). Marginal R-squared for gender values alone is included in parentheses. All models included random intercepts at country level. All independent variables are centered on the mean and standardized to have unit standard deviation. Entries estimate by how many points the relative achievement of boys vs. girls tends to increase when the corresponding independent variable increases by one standard deviation and the other variables are held constant. Collinearity was not at problematic levels; all variance inflation factors were less than or equal to 3.3.

by any domain of prosperity or any domain of gender equality in opportunities.

Which Are the Countries at Different Ends of Gender Egalitarian Values?

The scatter plots in **Figure 3** illustrate the relation between gender egalitarian values and the relative mean achievement of boys vs. girls without any controls. On the *y*-axis is simply the average of all available relative mean achievement scores for a country (per academic domain and assessment organization). On the *x*-axis is the country's gender egalitarian values as measured by WVS (left) or GLOBE (right). Countries are identified by their three-letter country codes (ISO 3166-1 alpha-3, e.g., KWT for Kuwait). The scatter plots reveal culturally based clusters of countries. In the lower left corner, characterized by low levels of gender

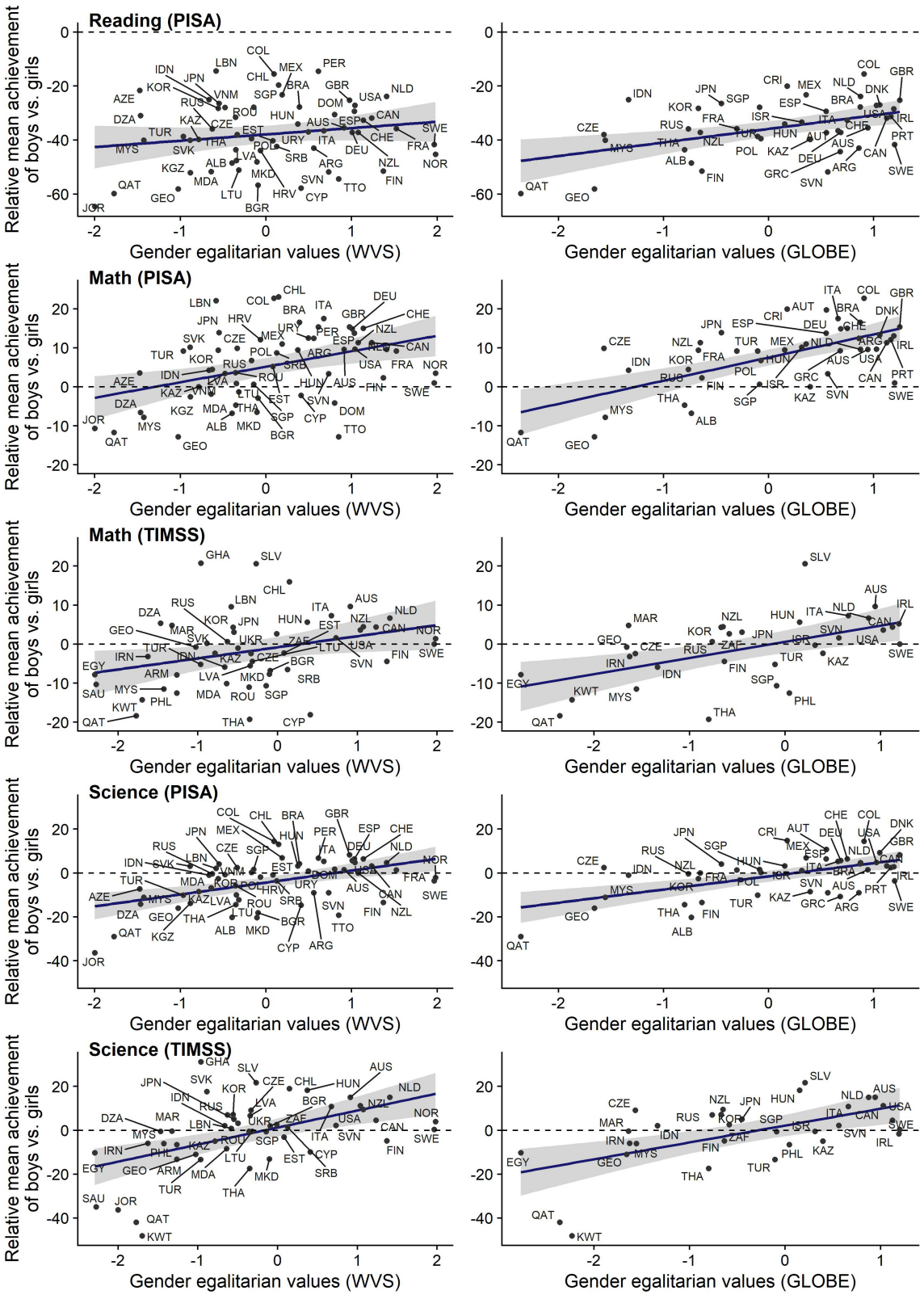

FIGURE 3 | Scatterplots of the relative achievement of boys vs. girls per academic subject (reading, math, science) and assessment (PISA, TIMSS), averaged over all available waves and plotted against gender egalitarian values measured by WVS **(left)** or GLOBE **(right)**. The solid lines are the best fitting regression lines. The dashed lines indicate equal achievement of boys and girls; above the dashed line, the gender gap favors boys, below it, the gender gap favors girls.

egalitarian values and low relative achievement of boys vs. girls, we tend to find countries on the Arabian Peninsula (Egypt, Qatar, Kuwait, and in the top panel also Jordan and Saudi Arabia). Whereas their economic conditions vary considerably, these neighboring countries are culturally similar. In the top right corner, characterized by high levels of gender egalitarian values and high relative achievement of boys vs. girls, we tend to find countries in Latin America, North America, Western and Middle Europe, Australia, and New Zealand. In between these groups, we tend to find countries from South-East Asia, Central Asia, and Eastern Europe.

The scatter plots for the domain of reading stand out from those of the other domains in three ways. First, there are fewer countries plotted. This is because reading is not included in TIMSS, so any country that has only participated in TIMSS will not have data on reading achievement. Second, all dots lie below the dashed reference line at zero in the scatter plots for reading, but not in the other scatter plots. Thus, we replicate the finding that girls consistently exhibit stronger achievement than boys in the domain of reading. Third, the points in the scatter plots for reading achievement are more widely scattered around the regression line than in the other scatter plots. This is consistent with our earlier observation that gender egalitarian values explain a smaller proportion of the total variation in the relative achievement of boys vs. girls in the reading domain than in the domains of mathematics and science.

Variation in Results Across Different Waves of PISA

A previous study found that the relative achievement of boys vs. girls on PISA tests was somewhat lower in countries with more gender egalitarian values, as measured by WVS (Guiso et al., 2008). It is remarkable that we have here obtained a robust effect in the opposite direction. The explanation lies in that Guiso et al. only analyzed data from the 2003 wave of PISA. As discussed in Section "Introduction," findings on gender differences in that dataset have failed to replicate in other waves of PISA (Stoet and Geary, 2013). To illustrate this phenomenon in the context of our study, we calculated the raw correlation between gender egalitarian values (WVS) and relative mean math achievement of boys vs. girls (PISA) separately for each wave (see **Table 2**). In each of the last four waves of PISA, we see a statistically significant positive raw correlation (ranging between 0.31 and 0.42), consistent with the results from our previous analysis of the pooled data. Moreover, consistent with the finding of Guiso et al. (mentioned above), the 2003 PISA data stands out by yielding a negative correlation (−0.17). Thus, the finding for 2003 does not depend on the details of how the data analysis was conducted, which differed somewhat between our study and the study of Guiso et al.[9] Instead, the crucial difference between the studies is that we pooled data from numerous waves to obtain a more representative dataset. The rest of **Table 2** shows that correlations vary substantially across waves also for the PISA reading test and the TIMSS math test.

TABLE 2 | Correlations between gender egalitarian values (WVS) and relative mean achievement of boys vs. girls for each wave of PISA and TIMSS.

PISA					TIMSS		
Domain	**Year**	***r***	**95% CI**	***n***	***r***	**95% CI**	***n***
Reading	2000	−0.06	[−0.37, 0.29]	30			
Reading	2003	−0.40	[−0.70, −0.01]	27			
Reading	2006	0.20	[−0.17, 0.53]	41			
Reading	2009	0.20	[−0.06, 0.46]	47			
Reading	2012	0.19	[−0.19, 0.49]	47			
Reading	2015	0.27	[−0.04, 0.54]	50			
Math	2003	−0.17	[−0.56, 0.23]	27	0.23	[−0.14, 0.56]	35
Math	2006	0.42	[0.10, 0.65]	42			
Math	2007				0.36	[0.08, 0.59]	35
Math	2009	0.41	[0.16, 0.64]	47			
Math	2011				0.40	[0.08, 0.68]	31
Math	2012	0.40	[0.05, 0.66]	47			
Math	2015	0.31	[0.02, 0.58]	50	0.59	[0.40, 0.78]	28
Science	2003				0.43	[0.11, 0.68]	35
Science	2006	0.45	[0.07, 0.68]	42			
Science	2007				0.55	[0.30, 0.73]	35
Science	2009	0.53	[0.29, 0.72]	47			
Science	2011				0.58	[0.28, 0.82]	31
Science	2012	0.45	[0.05, 0.68]	47			
Science	2015	0.43	[0.12, 0.65]	50	0.68	[0.54, 0.81]	28

The first two columns for each assessment (PISA, TIMSS) report Pearson correlations with bootstrapped 95% confidence intervals. The last column reports the sample size, that is, the number of countries in our dataset that participated in the assessment in the given year.

Lack of Robustness of Estimates of How Gender Differences in Achievement Change Over Time

As a supplementary analysis, we examined whether we can reliably estimate for each country how gender differences in achievement change over time. To examine this question, we calculated separate change estimates from TIMSS data and PISA data for math and science achievement. To obtain change estimates for a given country, we used the relative achievement of boys vs. girls in each wave as data points and performed a linear regression on the year of the wave; the resulting regression coefficient is an estimate of the direction and rate by which the relative achievement of boys vs. girls has changed in that country. In this way, separate change estimates for TIMSS and PISA were obtained for all 23 countries in which data has to be available from at least two waves of each assessment). These change estimates reflect both genuine change and noise from sampling errors. If the signal from genuine change dominates, we would expect change estimates in the same domain from the two assessments to be strongly positively correlated. However, they were not; correlations were close to zero both in the mathematics domain, $r = -0.09$, $p = 0.66$, and in the science domain, $r = 0.18$, $p = 0.42$. We conclude that the data are not sufficient to yield

[9]For instance, out of concern about possible differential drop-out rates across genders, Guiso et al. excluded all students in the lower half of the distribution of socio-economic status where drop-out rates were assumed to be higher. By comparison, we used all students and instead included the gender composition as a covariate.

reliable estimates of how the relative achievement of boys vs. girls changes over time in different countries.

DISCUSSION

The present research examined the relative achievement of boys vs. girls on standardized tests and how it varies with societies' attitudes toward gender equality. A clear pattern emerged. Across all assessed domains (reading, mathematics, science), the relative achievement of boys vs. girls was higher in countries with high levels of gender egalitarian values. These findings are based on data from several waves of the PISA and TIMSS assessments, together covering 74 countries. The same findings were obtained regardless of whether gender differences in achievement were measured using PISA or TIMSS and whether measured at the mean or 90th percentile, and regardless of whether gender egalitarian values were measured by WVS or GLOBE. Thus, the findings appear robust.

It is noteworthy that the findings of a previous study (Guiso et al., 2008), which only used data from the 2003 wave of PISA, did not replicate in our larger dataset. As discussed earlier, results from a single wave may be unreliable due to sampling errors and limited sets of participating countries.

Whereas the same relationship between gender egalitarian values and the relative achievement of boys vs. girls was found across different academic domains, it is important to note that the baseline level of the relative achievement of boys vs. girls varies greatly between reading on the one hand and mathematics and science on the other hand. With respect to reading, boys underachieve relative to girls in every country in our sample. This contrasts with the results in math and science, where underachievement of boys is found mainly in countries with low levels of gender egalitarian values. In countries with high levels of gender egalitarian values, boys tend to overachieve relative to girls on these math and science tests. Note, however, that even in countries where boys achieve better than girls on math and science tests, they do not necessarily do better in school on these subjects, because school grades also reflect other aspects, such as self-discipline, which seem to favor girls (Duckworth and Seligman, 2006).

What Explains the Association Between Gender Egalitarian Values and Gender Differences in Achievement?

Future research can test possible explanations of the observed association between gender egalitarian values and gender differences in academic achievement. It would be ideal if future PISA and TIMSS questionnaires were to include items bearing on the gender egalitarian values of the friends, teachers, and parents of individual students, to allow examination of the mechanisms suggested in Section "Introduction." The common thread of those hypothetical mechanisms is that gender egalitarian values could have the (unintended) consequence that girls become less engaged with school. Note that this is consistent with the results in Section "Using Gender Egalitarian Values to Predict the Achievement Levels of Boys and Girls," which (although not statistically significant) indicated that the independent effect of more gender egalitarian values on girls' absolute achievement levels is slightly negative. In this context, it is worth considering how our results can be reconciled with results from studies of the relative achievement of boys vs. girls among second-generation immigrants, in which higher levels of gender equality in the country of ancestry were associated with *lower* relative achievement of boys vs. girls (Nollenberger et al., 2016; Rodríguez-Planas and Nollenberger, 2018). Assuming that the gender egalitarian values of immigrants tend to reflect the level of gender equality in the country they come from, this finding seems to contrast with our finding that more gender egalitarian values is associated with *higher* relative achievement of boys vs. girls. We suggest that the difference may be explained by the specific situation of second-generation immigrants, for whom the values of their parents may conflict with the values of teachers and friends in the new country, which could lead to less compliance with parental values (Choi et al., 2008).

Causality

We now consider the question of causal direction. Could it be that the causal direction is the reverse to what we have assumed above, so that gender differences in achievement would affect cultural values? We find this direction implausible. In order for gender differences in achievement to be able to influence cultural values, a minimum requirement should be that people can readily observe these gender differences at a sufficient level of accuracy to distinguish between different countries. But we know that gender differences in achievement are much too small for that; indeed, we needed this statistical analysis of millions of students to be able to quantify them with sufficient accuracy.

Yet another possibility to consider is that the observed association could follow from some third variable driving both cultural values and achievement levels. We have tried to account for this in our analyses by controlling for various indicators of prosperity and gender equality in opportunities, which prior research would suggest to be the most likely candidates. We cannot exclude that there may be other important unobserved country-level factors. However, in the absence of any theory about such factors, we tentatively conclude that more gender egalitarian values in a country may in fact be a cause of higher relative achievement of boys vs. girls.

To establish causality, it would have been ideal to examine whether changes over time in values predict changes over time in gender differences in achievement. Unfortunately, we found that such an examination is not meaningful due to the data being insufficient to reliably measure change over time.

Effect Sizes of Gender Differences

Here we have examined gender differences in achievement in terms of differences in test scores. Much previous research has instead examined gender differences in terms of Cohen's *d*, which is the gender difference in test scores divided by the standard deviation in test scores in the country. These measures are extremely closely correlated (typically $r > 0.99$). Approximate quantitative results for gender differences in terms of Cohen's *d*

are easily obtained by a simple rule of thumb: divide results for raw score differences by 100, the typical standard deviation. For instance, our analyses indicated that on average an increase of gender egalitarian values by one standard deviation corresponds to an increase by about 6 points in the relative achievement of boys vs. girls, which means a change by 0.06 in Cohen's *d*. A change in gender egalitarian values from the very low level on the Arabian Peninsula to the very high level in Western Europe amounts to a change in Cohen's *d* on the order of 0.3. Thus, the variation between countries in gender differences in achievement is not negligible, but gender differences are on the whole quite small compared to the variation in achievement between students in the same country.

Unexplained Variation

Finally, we emphasize that there is still a lot of unexplained variation in gender differences in achievement. **Figure 1** shows that the relative math achievement of boys vs. girls in Ghana was much higher than expected from gender egalitarian values alone, whereas Sweden, Norway, Cyprus, Trinidad and Tobago, and Dominican Republic are outliers in the opposite direction. Future research could address these outliers. For instance, it would be interesting to understand why the latter two Caribbean countries differ so much from neighboring Latin American countries with respect to gender differences in achievement, given that they have similar levels of gender egalitarian values.

CONCLUSION

The main conclusion of our study of 74 countries is that gender egalitarian values seem to play a role in shaping gender differences in academic achievement that has not been documented in previous research. A prior study of 2003 PISA data found that the relative mathematics achievement of boys vs. girls tended to be lower in countries with more gender egalitarian values. In stark contrast to this finding, our analysis of a much larger dataset found that, regardless of academic domain, the relative achievement of boys vs. girls tended to be *higher* in countries with more gender egalitarian values. By comparison, measures of gender equality in opportunities had no clear independent effect on gender differences in academic achievement. Cultural values are pervasive and could influence almost every aspect of the academic environment of boys and girls: family, friends, and teachers. The exact pathway by which gender egalitarian values influence the academic achievement of boys and girls is still an open question, but plausible candidates include their freedom to engage in extracurricular activities and expectations on their academic efforts.

DATA AVAILABILITY STATEMENT

The dataset analyzed for this study can be found in the OSF https://osf.io/v7bqt/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

KE conceived of the study, performed the statistical analysis, and wrote the manuscript. MB contributed to the literature review and provided critical input on the manuscript. All authors read and approved the submitted version.

FUNDING

This research was supported by the Swedish Research Council (Grant No. 2014-2008), the Knut and Alice Wallenberg Foundation (Grant No. 2015.0005), and the Swedish Foundation for Humanities and Social Sciences (Grant No. P17-0030:1).

ACKNOWLEDGMENTS

KE is grateful to Pontus Strimling for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00236/full#supplementary-material

REFERENCES

Baker, D. P., and Jones, D. P. (1993). Creating gender equality: cross-national gender stratification and mathematical performance. *Sociol. Educ.* 66, 91–103. doi: 10.2307/2112795

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv* [preprint].

Berekashvili, N. (2012). The role of gender-biased perceptions in teacher-student interaction. *Psychol. Lang. Commun.* 16, 39–51. doi: 10.2478/v10057-012-0004-x

Brandt, M. J. (2011). Sexism and gender inequality across 57 societies. *Psychol. Sci.* 22, 1413–1418. doi: 10.1177/0956797611420445

Bronfenbrenner, U. (1979). *The Ecology of Human Development*. Cambridge, MA: Harvard University Press.

Bronfenbrenner, U. (2005). *Making Human Beings Human: Bioecological Perspectives on Human Development*. Thousand Oaks: SAGE.

Carrington, B., Chivers, T., and Williams, T. (1987). Gender, leisure and sport: a case-study of young people of South Asian descent. *Leisure Stud.* 6, 265–279. doi: 10.1080/02614368700390211

Choi, Y., He, M., and Harachi, T. W. (2008). Intergenerational cultural dissonance, parent–child conflict and bonding, and youth problem behaviors among vietnamese and cambodian immigrant families. *J. Youth Adolesc.* 37, 85–96. doi: 10.1007/s10964-007-9217-z

DiPrete, T. A., and Buchmann, C. (2013). *The Rise of Women: The Growing Gender Gap in Education and What it Means for American Schools*. New York, NY: Russell Sage Foundation.

Duckworth, A. L., and Seligman, M. E. (2006). Self-discipline gives girls the edge: gender in self-discipline, grades, and achievement test scores. *J. Educ. Psychol.* 98, 198–208. doi: 10.1037/0022-0663.98.1.198

Else-Quest, N. M., Hyde, J. S., and Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychol. Bull.* 136, 103–127. doi: 10.1037/a0018053

Frawley, T. (2005). Gender bias in the classroom: current controversies and implications for teachers. *Child. Educ.* 81, 221–228.

Grant, M. J., and Behrman, J. R. (2010). Gender gaps in educational attainment in less developed Countries. *Popul. Dev. Rev.* 36, 71–89. doi: 10.1111/j.1728-4457.2010.00318.x

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. *Science* 320, 1164–1165. doi: 10.1126/science.1154094

Hausmann, R., Tyson, L. D., and Zahidi, S. (2009). *The Global Gender Gap Report 2009*. Geneva: World Economic Forum.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., and Gupta, V. (2004). *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. Thousand Oaks: SAGE Publications.

Hyde, J. S. (2005). The gender similarities hypothesis. *Am. Psychol.* 60, 581–592. doi: 10.1037/0003-066x.60.6.581

Hyde, J. S., and Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proc. Natl. Acad. Sci. U.S.A.* 106, 8801–8807. doi: 10.1073/pnas.0901265106

Inglehart, R., and Norris, P. (2003). *Rising Tide: Gender Equality and Cultural Change Around the World*. Cambridge, MA: Cambridge University Press.

Jackson, C., and Nyström, A.-S. (2015). 'Smart students get perfect scores in tests without studying much': why is an effortless achiever identity attractive, and for whom is it possible? *Res. Pap. Educ.* 30, 393–410. doi: 10.1080/02671522.2014.970226

Jussim, L., and Eccles, J. S. (1992). Teacher expectations: II. Construction and reflection of student achievement. *J. Pers. Soc. Psychol.* 63, 947–961. doi: 10.1037/0022-3514.63.6.947

Kane, J. M., and Mertz, J. E. (2012). Debunking myths about gender and mathematics performance. *Not. Am. Math. Soc.* 59:10. doi: 10.1090/noti790

Lansford, J. E., Skinner, A. T., Sorbring, E., Giunta, L. D., Deater-Deckard, K., Dodge, K. A., et al. (2012). Boys' and Girls' relational and physical aggression in nine Countries. *Aggress. Behav.* 38, 298–308. doi: 10.1002/ab.21433

Lubinski, D., and Benbow, C. P. (2006). Study of mathematically precocious youth after 35 years: uncovering antecedents for the development of math-science expertise. *Perspect. Psychol. Sci.* 1, 316–345. doi: 10.1111/j.1745-6916.2006.00019.x

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: evidence from 31 countries. *Oxf. Rev. Educ.* 34, 89–109. doi: 10.1080/03054980701565279

Morris, E. W. (2012). *Learning the Hard Way: Masculinity, Place, and the Gender Gap in Education*. New Brunswick, NJ: Rutgers University Press.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., and Preuschoff, C. (2009). *TIMSS 2011 Assessment Frameworks*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Nakagawa, S., Johnson, P. C., and Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *J. R. Soc. Interf.* 14:20170213. doi: 10.1098/rsif.2017.0213

Nollenberger, N., Rodríguez-Planas, N., and Sevilla, A. (2016). The math gender gap: the role of culture. *Am. Econ. Rev.* 106, 257–261. doi: 10.1257/aer.p20161121

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PLoS One* 7:e39904. doi: 10.1371/journal.pone.0039904

Reilly, D., Neumann, D. L., and Andrews, G. (2019). Investigating gender differences in mathematics and science: results from the 2011 trends in mathematics and science survey. *Res. Sci. Educ.* 49, 25–50. doi: 10.1007/s11165-017-9630-6

Robinson, J. P., and Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: examining direct cognitive assessments and teacher ratings. *Am. Educ. Res.earch J.* 48, 268–302. doi: 10.3102/0002831210372249

Rodríguez-Planas, N., and Nollenberger, N. (2018). Let the girls learn! It is not only about math…it's about gender social norms. *Econ. Educ. Rev.* 62, 230–253. doi: 10.1016/j.econedurev.2017.11.006

Schleicher, A., Zimmer, K., Evans, J., and Clements, N. (2009). *PISA 2009 Assessment Framework: Key Competencies in Reading, Mathematics and Science*. Paris: OECD Publishing.

Stevenson, H. W., Chen, C., and Lee, S.-Y. (1993). Mathematics achievement of Chinese, Japanese, and American children: ten years later. *Science* 259, 53–58. doi: 10.1126/science.8418494

Stoet, G., Bailey, D. H., Moore, A. M., and Geary, D. C. (2016). Countries with higher levels of gender equality show larger national sex differences in mathematics anxiety and relatively lower parental mathematics valuation for girls. *PLoS One* 11:e0153857. doi: 10.1371/journal.pone.0153857

Stoet, G., and Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: within-and across-nation assessment of 10 years of PISA data. *PLoS One* 8:e57988. doi: 10.1371/journal.pone.0057988

Stoet, G., and Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence* 48, 137–151. doi: 10.1016/j.intell.2014.11.006

UNDP (2015). *Human Development Report 2015: Work for Human Development*. New York, NY: United Nations Development Programme.

Voyer, D., and Voyer, S. D. (2014). Gender differences in scholastic achievement: a meta-analysis. *Psychol. Bull.* 140, 1174–1204. doi: 10.1037/a0036620

Welzel, C. (2013). *Freedom Rising*. Cambridge, MA: Cambridge University Press.

Zill, N., and West, J. (2001). *Entering Kindergarten: A Portrait of American Children When They Begin School. Findings from the Condition of Education, 2000*. Washington, DC: National Center for Education Statistics.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 08 April 2020
doi: 10.3389/fpsyg.2020.00626

A Multilevel Person-Centered Examination of Teachers' Workplace Demands and Resources: Links With Work-Related Well-Being

Rebecca J. Collie[1], Lars-Erik Malmberg[2], Andrew J. Martin[1], Pamela Sammons[2] and Alexandre J. S. Morin[3]*

[1] School of Education, University of New South Wales, Sydney, NSW, Australia, [2] Department of Education, University of Oxford, Oxford, United Kingdom, [3] Substantive Methodological Synergy Research Laboratory, Department of Psychology, Concordia University, Montreal, QC, Canada

OPEN ACCESS

Edited by:
Ronnel B. King, The Education University of Hong Kong, Hong Kong

Reviewed by:
Yuyang Cai, Shanghai University of International Business and Economics, China
Ma. Jenina N. Nalipay, The Education University of Hong Kong, Hong Kong

***Correspondence:**
Rebecca J. Collie
rebecca.collie@unsw.edu.au

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *13 December 2019*
Accepted: *16 March 2020*
Published: *08 April 2020*

Citation:
Collie RJ, Malmberg L-E, Martin AJ, Sammons P and Morin AJS (2020) A Multilevel Person-Centered Examination of Teachers' Workplace Demands and Resources: Links With Work-Related Well-Being. Front. Psychol. 11:626. doi: 10.3389/fpsyg.2020.00626

Teachers' healthy and effective functioning at work is impacted by the demands they face and the resources they can access. In this study, person-centered analysis was adopted to identify distinct teacher profiles of demands and resources. We investigated teachers' experiences of two job demands (barriers to professional development and disruptive student behavior), two job resources (teacher collaboration and input in decision-making), and one personal resource (self-efficacy for teaching). Using data from the Teaching and Learning International Survey (TALIS) 2013, the study involved 6,411 teachers from 369 schools in Australia and 2,400 teachers from 154 schools in England. In phase one, latent profile analysis revealed five teacher profiles that were similar across the two countries: the Low-Demand-Flourisher (12%), Mixed-Demand-Flourisher (17%), Job-Resourced-Average (34%), Balanced-Average (15%), and Struggler (21%). The profiles were differently associated with two background characteristics (teacher gender and teaching experience) and two work-related well-being outcomes (job satisfaction and occupational commitment). In phase two, we extended our analysis to the school-level to identify school profiles based on the relative prevalence of the five teacher profiles within a school. Indeed, a yield of large scale datasets such as TALIS is that there are sufficient units at the school-level to enable institutional insights, beyond insights garnered at the individual teacher-level. Two school profiles that were similar in both countries were revealed: the Unsupportive school profile (58%) and the Supportive school profile (42%). The Supportive school profile was associated with higher school-average teacher job satisfaction and occupational commitment than the Unsupportive school profile. Taken together, the findings yield knowledge about salient teacher and school profiles, and provide guidance for possible interventions at the teacher- and school level.

Keywords: job demands-resources theory, teacher well-being, latent profile analysis, multilevel, big data

INTRODUCTION

Teaching work is complex. The extent to which teachers thrive at work involves a delicate balance between the demands placed upon them and the resources they can access to support them in their work (Hakanen et al., 2006). A growing body of research has examined the role of job demands (e.g., disruptive student behavior), job resources (e.g., social support), and personal resources (e.g., adaptability) in predicting teachers' well-being at work (e.g., Desrumaux et al., 2015; Collie and Martin, 2017; Dicke et al., 2018; Skaalvik and Skaalvik, 2018). This prior work has tended to use variable-centered approaches (e.g., multiple/multivariate regression models within the structural equation modeling framework) that describe how the factors are interrelated (e.g., the association between job resources and well-being; Collie et al., 2018). Resting on the assumption of population-homogeneity, variable-centered approaches thus provide important information about relations at a sample-wide level and about the particular variables that could be targeted in broad intervention efforts. However, such research is less able to ascertain the extent to which there are different subpopulations of teachers identifiable based on commonly shared experiences of demands and resources reflecting population-heterogeneity, and whether there are particular combinations better aligned with well-being outcomes. To examine this, person-centered approaches, such as latent profile analyses, are ideally suited. Person-centered approaches identify distinct subpopulations (or profiles) of individuals who fare similarly on several factors. Person-centered approaches thus reveal knowledge of how intervention efforts can be tailored to the needs of each of these profiles.

A small, but growing body of research has conducted person-centered examinations of teachers' experiences at work (e.g., Klusmann et al., 2008; Watt and Richardson, 2008; Simbula et al., 2012; Collie et al., 2015; Morin et al., 2015, 2017; Collie and Martin, 2017; Meyer et al., 2019; Perera et al., 2019). However, researchers have yet to consider job demands, job resources, and personal resources simultaneously. Moreover, the extent to which schools can be identified based on the prevalence of different demand-resource profiles among teachers remains largely unexamined. We suggest that these two gaps are important to address in order to ascertain the major categories of teachers who work in schools, the different combinations of demands and resources that characterize these subpopulations of teachers, and to inform policies and practice on how best to target intervention relevant to each distinct profile within and across schools. Such understanding is essential for promoting healthy and effective teachers and schools (e.g., Arens and Morin, 2016).

The aims of the current study, therefore, were to identify profiles of demands and resources experienced by teachers and then to explore the extent to which distinct profiles are predicted by teachers' background characteristics and associated with meaningful differences in workplace well-being outcomes. We also investigated school-level profiles by identifying the proportion of the teacher-level profiles evident within different subpopulations of schools (i.e., school profiles), along with links to school-average well-being outcomes. **Figure 1** demonstrates the models under examination. We harnessed job demands-resources theory (Bakker and Demerouti, 2017), and examined two job demands (barriers to professional development and disruptive student behavior), two job resources (teacher collaboration and input in decision-making), and one personal resource (self-efficacy for teaching). We examined these factors because together they reflect demands and resources that help or hinder teachers' ability to undertake their work effectively (e.g., OECD, 2014; Skaalvik and Skaalvik, 2018). Of note, these factors have been shown to be implicated in teachers' well-being in variable-centered analyses (e.g., Skaalvik and Skaalvik, 2018).

Data from the Teaching and Learning International Survey (TALIS) 2013 were used in our study. Key strengths of using large scale datasets like TALIS is that sampling is considered to be nationally representative, and sufficient numbers of schools are sampled to allow insights at the institutional level, beyond insights garnered at the individual teacher-level. Our study was conducted with teachers from Australia and England in order to provide evidence of the generalizability of these profiles, and to help guide national and international research, practice, and policy. Taken together, the findings have important implications for understanding demand-resource profiles among teachers and whether particular combinations are better aligned with well-being outcomes. The findings also have the potential to yield knowledge about salient school profiles and provide guidance for the development of appropriate intervention at the teacher- and school-level.

Conceptual Framework

We rely on the job demands-resources (JD-R; Bakker and Demerouti, 2017) theory as our conceptual framework. This theoretical model highlights the idea that every job comprises varying types and levels of demands that impede employee functioning at work, as well as varying resources that support employee functioning at work (Bakker and Demerouti, 2017). Job demands (e.g., high workload) and job resources (e.g., strong social support) can be psychological, physical, social, or organizational in nature (Bakker and Demerouti, 2017). Although job resources are beneficial for employees' motivation and well-being, job demands have the reverse association and are linked with a variety of undesirable outcomes, including burnout (Bakker and Demerouti, 2017). JD-R theory also stipulates the important role of personal resources, which are personal capacities that reflect employees' potential to have influence on their working environment (e.g., self-efficacy; Bakker and Demerouti, 2017). Much like job resources, personal resources are beneficial for employees' motivation and well-being at work (Bakker and Demerouti, 2017). In the current study, we examined job demands, job resources, and personal resources simultaneously.

Demands and Resources Associated With Teachers' Ability to Undertake Their Work

There are many job demands (e.g., time pressure), job resources (e.g., leadership support) and personal resources

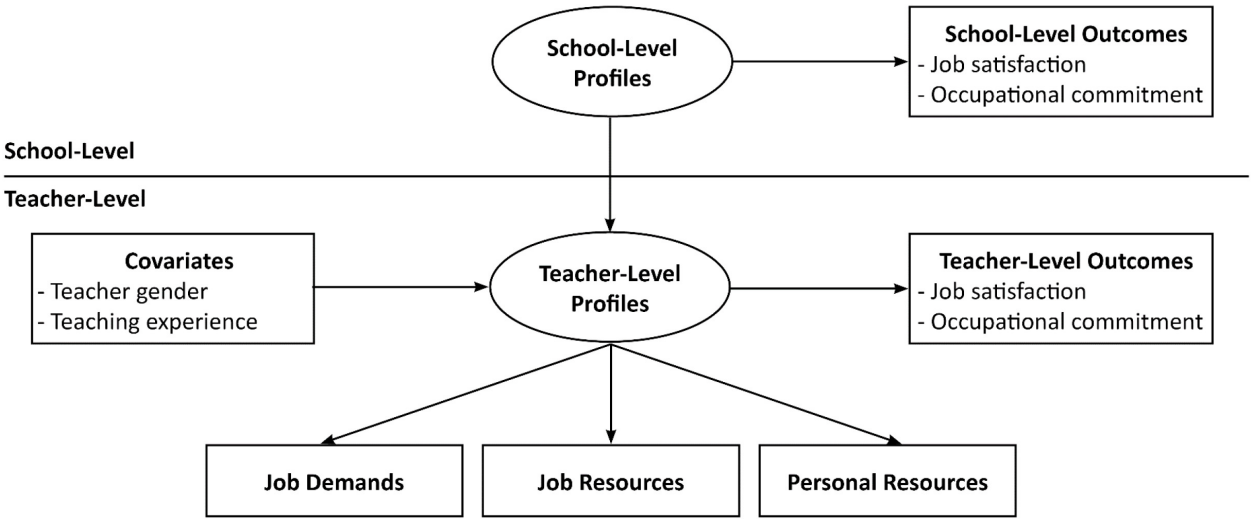

FIGURE 1 | Hypothesized models tested in the study at the teacher- and school-level. In phase one of analysis, teacher-level profiles were identified based on the demands and resources. Then, tests of predictor associations with the profiles were conducted, followed by examination of mean differences between profiles on the teacher-level outcome. In phase two of analysis, school-level profiles were identified and then mean differences in the aggregated outcome were tested across school-level profiles. Not shown here are the tests of profile similarity that were conducted to compare the teacher-level results across Australia and England (see section "Materials and Methods)."

(e.g., adaptability) that impact teachers' work (Collie et al., 2018; Skaalvik and Skaalvik, 2018). In the current study, we focused on several that are implicated in the extent to which teachers are able to effectively undertake their work (e.g., Collie et al., 2012), and that are also important for their well-being (e.g., Vangrieken et al., 2015).

The first job demand that we examined was *barriers to professional development*, which reflects teachers' experience of factors preventing them from accessing the training necessary for their ongoing learning and development as a teacher. Barriers to professional development may include financial constraints, lack of appropriate opportunities, limited support from leadership, or limited time to complete such activities (Kwakman, 2003; OECD, 2009; Broadley, 2010). The second job demand was *disruptive student behavior*, which reflects behavior that makes it difficult for effective instruction to occur (e.g., students' calling out or refusing to listen). Disruptive behavior has consistently been identified in research as challenging for teachers (e.g., Skaalvik and Skaalvik, 2018). Teachers who are unable to access professional development or who experience high levels of disruptive behavior in the classroom report reduced well-being (e.g., Collie et al., 2012; Skaalvik and Skaalvik, 2018). To our knowledge, researchers have yet to examine how these two job demands are interrelated. However, a positive association is assumed given that barriers to professional development may impinge on teachers' ability to develop their capacities to effectively engage students and manage the classroom.

Turning attention to job resources, *teacher collaboration* involves the extent to which teachers work with their colleagues to plan, develop, teach, and/or assess student learning (OECD, 2014. Teacher collaboration helps teachers to undertake their work effectively because it can save time and introduce teachers to new ideas and resources (Reeves et al., 2017). *Teacher input in decision-making* refers to the extent to which the school provides teachers with opportunities to participate in and share responsibility for school-level decisions (OECD, 2014). Having input in decisions helps teachers to feel supported in their role (Collie et al., 2016) and it helps to ensure teachers' needs are met (De Neve et al., 2015). Feeling supported and having needs met are important for helping teachers to undertake their work effectively (e.g., Taylor et al., 2008) and are also linked with teachers' well-being at work (e.g., Vangrieken et al., 2015). Moreover, prior research suggests that these two job resources are moderately and positively correlated (e.g., Collie et al., 2012), which likely occurs because both collaboration and input reflect a school climate that is more collegial in relation to teaching and learning. Collegiality and positive relationships are known to be important for well-being (e.g., Ryan and Deci, 2017).

We also examined one personal resource—*self-efficacy for teaching*—which reflects teachers' belief that they can bring about effective learning among students (Tschannen-Moran and Woolfolk Hoy, 2001). A large body of research has highlighted the important role of self-efficacy in many different countries (e.g., Fackler and Malmberg, 2016)—with both a focus on specific types of self-efficacy (for instructional practices, classroom management, and student engagement) and self-efficacy as a global construct (e.g., Collie et al., 2012; Granziera and Perera, 2019; Perera et al., 2019). The current study provided the opportunity to examine global levels of self-efficacy for teaching alongside other job demands and job resources to ascertain its role in teachers' well-being.

Taken together, we thus focus on five factors that are salient in helping or hindering teachers to undertake their work effectively (e.g., Leithwood et al., 2008; Vangrieken et al., 2015). More

precisely, these five factors encompass teachers' interactions with students, other teachers, and school leaders, as well as teachers' own professional growth and confidence. These different aspects reflect central components of teaching work (e.g., Collie et al., 2016) and are linked with teachers' psychological functioning (e.g., Klassen et al., 2012). Teachers who have positive interactions with other school members, and who feel confident at what they do, are more likely to be satisfied with their work and committed to their profession (e.g., Collie et al., 2012). The opposite is true for teachers who have challenging interactions with students or who do not experience agency in relation to their professional growth (e.g., Skaalvik and Skaalvik, 2018). Notably, the bulk of prior research on these factors has been variable-centered in nature, revealing important knowledge about how these factors are associated with one another and important outcomes. However, variable-centered approaches are less able to speak to teacher profiles which might each reflect a different mix of demands and resources (i.e., high, medium, or low on different demands and resources)—a useful insight for targeted intervention and teacher support. This is where person-centered research is particularly informative.

Identifying Teacher Profiles Through Person-Centered Research

Although variable-centered research harnessing JD-R theory is abundant (e.g., Simbula et al., 2012; Skaalvik and Skaalvik, 2018), person-centered research is only just emerging. We suggest that person-centered research has the potential to provide a theoretical contribution to JD-R theory because it allows simultaneous examination of the interplay between multiple factors. This interplay is relevant to a distinct process established in JD-R theory, the boosting process, which suggests that demands boost the impact of resources on workplace wellbeing (Bakker and Demerouti, 2017). More precisely, job resources play a more substantial role in promoting well-being when job demands are high. The boosting process is tested by examining interactions between demands and resources, and their joint association with well-being outcomes. Yet, it quickly becomes challenging, even impossible, to meaningfully interpret the results from interaction effects involving more than two or three interacting variables. Person-centered analysis provides another way of simultaneously considering the combined impact of multiple demands and resources, and allows a more nuanced overview of this combined role by revealing the specific interplay of factors that best characterize distinct segments (or profiles) of a population. For example, there may be a subpopulation of teachers for whom high levels of teacher collaboration and self-efficacy help to offset the detrimental role of disruptive student behaviors—resulting in levels of well-being that are similar or higher than in other groups who experience the same resources, but not the disruptive behavior (as per the boosting process in JD-R theory; Bakker and Demerouti, 2017).

Research is now investigating demands and resources using person-centered approaches, showing that various combinations of these factors exist among employees (e.g., Van den Broeck et al., 2012; Mäkikangas et al., 2018; Moeller et al., 2018). For example, Van den Broeck et al. (2012) examined several job demands (e.g., workload, emotional and cognitive demands) and job resources (e.g., social support, autonomy) among general employees. Their results revealed four demand-resource profiles: high demand-low resource, high demand/resource, low-demand/resource, and low demand-high resource. Employees in profiles with higher demands and lower resources reported greater burnout and lower engagement. To date, there appears to be very little research conducted among teachers. However, in one relevant teacher-focused study, Simbula et al. (2012) examined two job resources (professional development and collegial support) and two job demands (role ambiguity and over investment in work). The researchers identified three profiles: high demand/resource, high demand-low resource, and low resource-high demand. More generally, there is a growing number of teacher-focused studies that have involved examinations of well-being or motivation profiles, again revealing varying combinations that are associated with other important workplace outcomes (Klusmann et al., 2008; Watt and Richardson, 2008; Collie and Martin, 2017; Morin et al., 2017; Meyer et al., 2019).

In sum, there is emerging research undertaking person-centered examinations of demands and resources. However, to date, it appears that researchers have yet to simultaneously consider teachers' perceptions of job demands, job resources, and personal resources. The aim of the current study was thus to extend the literature by considering these three facets simultaneously among teachers, and by focusing on factors that are central to teachers' work and well-being. We suggest that this focus is important given that teaching work is uniquely distinct from other professional groups (e.g., different relationships with "clients"; Klassen et al., 2012) and given major concerns about teacher well-being worldwide (e.g., Skaalvik and Skaalvik, 2018).

School-Level Phenomena and Their Association With Teachers' Workplace Experiences

When factors play a significant role at the school-level, this indicates differences between schools that can then be a target for intervention. In research on teachers, the bulk of work has been conducted at the teacher-level. However, a growing body of work demonstrates the salience of considering factors at the school-level. For example, leadership style and school-average teacher collaboration have been associated with school-average teachers' levels of organizational commitment (Duyar et al., 2013) and self-efficacy (Fackler and Malmberg, 2016). Although the variance explained by school-level factors can often be relatively modest, the school-level is nonetheless important to consider from a measurement perspective (Bliese et al., 2018) and given prior work demonstrating that school-level phenomena have a role to play in teachers' outcomes (e.g., Duyar et al., 2013). Moving forward, it is important to extend this variable-centered knowledge to a person-centered understanding. Large secondary data sets like TALIS can play an important role in facilitating school-level research because they contain a

sufficient number of schools to allow robust multilevel modeling. Multilevel person-centered approaches are able to ascertain the extent to which organizations can be identified based on the prevalence of different profiles among members (Mäkikangas et al., 2018). For example, Urick (2016) examined school-level profiles of leadership that were based on teacher-level profiles of leadership perceptions. Multilevel person-centered research, therefore, reveals the nature of different schools that can then be used to guide intervention efforts that target the specific needs of each school. For example, if there are schools that largely comprise teacher profiles reflecting low resources, then school-wide intervention may focus on increasing resources within those schools (in addition to interventions focused at the teacher-level).

Do Teacher and School Characteristics Predict Profile Membership?

Background characteristics can provide additional understanding about the nature of profiles by revealing the extent to which profile membership is predicted by different teacher or school characteristics. Researchers have shown mixed findings on whether teacher gender predicts membership in demand-resource profiles (Simbula et al., 2012; Collie et al., 2015). Researchers have also shown that teaching experience is unrelated to membership in demand-resource profiles (Collie et al., 2015). This limited research provides preliminary understanding about the role of these background characteristics. However, more research is needed to see if these findings hold within additional samples and contexts. As such, we examined the role of teacher gender and teaching experience in predicting teacher-level profile membership.

Are Profiles Linked With Teacher Work-Related Well-Being?

Teacher well-being is a multidimensional concept that reflects positive and healthy functioning in the workplace (Collie et al., 2016; Ryan and Deci, 2017). In the current study, we focused on two workplace outcomes that reflect experiences of work-related well-being: job satisfaction and occupational commitment. Job satisfaction involves employees' feelings of contentment in relation to their work (Schleicher et al., 2011). Occupational commitment reflects employees' attachment to their profession (Meyer et al., 1993). Variable-centered research has demonstrated the salience of a variety of job demands (e.g., time pressure), job resources (e.g., leadership support), and personal resources (e.g., adaptability) in predicting these two outcomes (e.g., Lee and Nie, 2014; Malinen and Savolainen, 2016; Skaalvik and Skaalvik, 2017; Collie et al., 2018). Emerging work is beginning to show that demand-resource profiles are differently linked with various workplace outcomes. For example, the Simbula et al. (2012) study introduced above identified three profiles of job demands and job resources, and demonstrated that these were linked with significantly different levels of workplace outcomes such as engagement and job satisfaction (the profiles with higher resources tended to display higher levels of the outcomes). In the current study, we extend prior research by examining novel job demands, job resources, and personal resources simultaneously—and in relation to both job satisfaction and occupational commitment. In addition, we also examined the extent to which the school-level profiles were associated with differences in school-average job satisfaction and occupational commitment.

The Importance of Cross-National Research

There is a global recognition of the significance of demands and resources in impacting teachers' well-being at work (e.g., Reeves et al., 2017; Dicke et al., 2018; Skaalvik and Skaalvik, 2018). An important and growing area of research, therefore, investigates teachers' experiences across different countries (e.g., Watt et al., 2012; Fackler and Malmberg, 2016), and large scale datasets can provide important insights into this. Cross-national research can reveal similarities or differences in how teachers are faring across borders, and thus provide guidance for practice and policy at the national- and international-level (Watt et al., 2012). Despite growing awareness of the global relevance of teacher well-being, prior studies have typically been conducted within a single country (however, see Watt et al., 2012; Fackler and Malmberg, 2016).

The current research extends research in this area by examining the experiences of teachers in two English-speaking countries: Australia and England. There are some key similarities between the two countries. For example, Australia and England have similar schooling systems, ethnically diverse student populations, high inbound migration, and similar social stratification (e.g., Bulle, 2011; Scherer et al., 2016). In addition, major policy developments in the two countries over the past decade are relevant to teacher well-being. For example, both Australia and England have implemented national standardized testing for students, and increased scope of school evaluations (OECD, 2019). High-stakes tests and evaluations are known to be stressful for teachers (e.g., von der Embse et al., 2016). In addition, Australia and England have higher levels of disruptive behavior than the OECD average (OECD, 2019). Importantly, in both countries there is increasing unease about the demanding nature of teaching work and its impact on teachers' well-being. Indeed, concerns about teacher attrition and burnout have been formally raised by governments in both countries over the past 12–18 months (e.g., UK Department for Education, 2018; Parliament of Australia, 2019). Similar policy-focused attention does not appear to be mirrored at a government-level in other English-speaking countries (e.g., Canada, the United States). Nonetheless, teacher well-being is an issue relevant to many other education contexts across the globe (e.g., United States, Hong Kong; Gallup, 2014; McInerney et al., 2018), and there are many commonalities in the demands and resources experienced by teachers worldwide (e.g., Fackler and Malmberg, 2016; Dicke et al., 2018; Skaalvik and Skaalvik, 2018). Thus, the results of the current study have the potential to provide knowledge for practice and research for teachers, schools, and educational systems internationally.

STUDY OVERVIEW

In the current study, we identified demand-resource profiles among teachers and schools. We examined two job demands (barriers to professional development and disruptive behavior), two job resources (teacher collaboration and input in decision-making), and one personal resource (self-efficacy for teaching). In phase one of analysis, we identified demand-resource profiles among teachers in Australia and England. We then compared the solutions across the two countries to ascertain the generalizability of our results, as well as generalizability of links between the profiles and teachers' background characteristics (gender, teaching experience) and two well-being outcomes (job satisfaction and occupational commitment). In phase two of analysis, we extended our examination to the school-level where we sought to identify school profiles based on the relative frequency of teacher profiles. We then compared these profiles across the two countries to verify the generalizability of these results, and tested whether the school profiles displayed different levels of school-average teacher job satisfaction and occupational commitment. **Figure 1** illustrates the models under examination.

MATERIALS AND METHODS

Sample and Procedure

The sample comprised 6,411 teachers from 369 schools in Australia and 2,400 teachers from 154 schools in England who participated in the Teaching and Learning International Survey (TALIS) 2013. TALIS is a survey run by the OECD every 5 years and was chosen for this study as it yields comprehensive and nationally representative data on demands and resources relevant to teachers' work. Sample selection for TALIS 2013 involved a two-stage probability sampling design to ensure a representative sample of schools and of teachers within those schools (for full details see OECD, 2014).

Starting with the Australian sample, participating teachers were 57% female, had an average age of 43 (*SD* = 12) years, and had an average teaching experience of 16 (*SD* = 11) years. Most teachers (84%) were working full-time, and almost all (99%) had attained ISCED Level 5A (bachelor's degree) or higher. The Australian teachers taught at ISCED Level 2 (lower secondary) and/or ISCED Level 3 (upper secondary). Just over half (55%) of the participating schools were publicly managed, and most (74%) had fewer than one-third students from low-SES backgrounds. Most schools (62%) had a male principal with an average age of 55 (*SD* = 7) years and an average experience as a principal of 9 (*SD* = 7) years. The schools were located in hamlets/villages (5%; <3,000 people), small towns (7%; 3,000–15,000 people), towns (16%; 15,000–100,000 people), cities (26%; 100,000–1 million people), and large cities (46%; >1 million people). There were on average 18 (*SD* = 5) teachers per school.

Participating teachers from England were 63% female, had an average age of 39 (*SD* = 10) years, and had an average teaching experiences of 12 (*SD* = 9) years. Most teachers (87%) were working full-time, and almost all (97%) had attained ISCED Level 5A (bachelor's degree) or higher. The entire English sample taught at ISCED Level 2 (lower secondary). Just over half (55%) of the participating schools were publicly managed, and most (76%) had fewer than one-third of their students from low-SES backgrounds. Most schools (64%) had a male principal with an average age of 50 (*SD* = 7) years and an average experience as a principal of 7 (*SD* = 6) years. The schools were located in hamlets/villages (4%), small towns (18%), towns (41%), cities (20%), and large cities (16%). There were on average 16 (*SD* = 4) teachers per school.

Measures

Measures were drawn from the 2013 TALIS Teacher Questionnaire (OECD, 2014; see **Supplementary Material** for items). All variables were modeled at the teacher-level. The teacher well-being outcomes (job satisfaction and occupational commitment) were also modeled at the school-level.

Job Demands

Barriers to professional development was assessed with items from the TALIS "Barriers to Professional Development" scale (6 items; e.g., "Professional development is too expensive/unaffordable"). Items were scored on a scale from 1 (Strongly disagree) to 4 (Strongly agree). Reliability was assessed with coefficient omega[1] and was adequate for the Australian (ω = 0.74) and English (ω = 0.74) samples. The scale displayed 7% variance (intraclass correlation [ICC] = 0.07) at the school-level. Although this is somewhat modest, it does warrant multilevel analyses (Bliese et al., 2018).

Disruptive student behavior was assessed with items from the TALIS "Your Teaching" scale (3 items; "When the lesson begins, I have to wait quite a long time for students to quiet down," "I lose quite a lot of time because of students interrupting the lesson," and "There is much disruptive noise in this classroom"). Items were scored on a scale from 1 (Strongly disagree) to 4 (Strongly agree). Reliability was satisfactory (ω_{Aus} = 0.89; ω_{Eng} = 0.90) and the scale demonstrated adequate variance at the school-level (ICC = 0.14).

Job Resources

Teacher collaboration was assessed with items from the TALIS "Teaching in General" scale (3 items; "On average, how often do you do the following in this school? 'Exchange teaching materials with colleagues,' 'Engage in discussions about the learning development of specific students,' and 'Work with other teachers in my school to ensure common standards in evaluations for assessing student progress'"). Items were scored from 1 (Never) to 6 (Once a week or more). Reliability estimates were adequate (ω_{Aus} = 0.75; ω_{Eng} = 0.73) and the scale demonstrated modest, but adequate, variance at the school-level (ICC = 0.05; Bliese et al., 2018).

Teacher input in decision-making was assessed with items from the TALIS "School Climate" scale (3 items; "This school provides staff with opportunities to actively participate in school decisions," "This school has a culture of shared responsibility for

[1]Coefficient omega was calculated using factor loadings from congeneric single-level or multilevel CFAs calculated separately for the two countries.

school issues," and "There is a collaborative school culture which is characterized by mutual support"). Items were scored on a scale from 1 (Strongly disagree) to 4 (Strongly agree). Reliability was satisfactory ($\omega_{Aus} = 0.86$; $\omega_{Eng} = 0.87$) and the scale demonstrated adequate variance at the school-level (ICC = 0.14).

Personal Resources

Teacher self-efficacy was assessed with items from the TALIS "Teaching in General" scale that encompasses three types of self-efficacy: self-efficacy for classroom management (4 items; "In your teaching, to what extent can you do the following? 'Control disruptive behavior in the classroom'), self-efficacy for instruction (4 items; "In your teaching, to what extent can you do the following? 'Craft good questions for my students'), and self-efficacy for student engagement (4 items; "In your teaching, to what extent can you do the following? 'Help my students value learning'). Items were scored on a scale from 1 (Not at all) to 4 (A lot). Reliability was satisfactory ($\omega_{Aus} = 0.86$; $\omega_{Eng} = 0.86$) and there was modest, but adequate variance at the school-level (ICC = 0.05; Bliese et al., 2018). Preliminary analyses indicated that the three different self-efficacy factors co-occurred at similar levels across profiles. For reasons of parsimony and because the self-efficacy factors were quite strongly intercorrelated (r's = 0.64–0.72), self-efficacy was modeled as a higher-order factor.

Outcomes

Job satisfaction and *occupational commitment* were assessed with items from the TALIS "About Your Job" scale. For job satisfaction, 3 items were used ("I enjoy working at this school," "I would recommend my school as a good place to work," and "All in all, I am satisfied with my job"). For occupational commitment, 4 items were used ("The advantages of being a teacher clearly outweigh the disadvantages," "If I could decide again, I would still choose to work as a teacher," "I regret that I decided to become a teacher" [reverse coded], "I wonder whether it would have been better to choose another profession" [reverse coded]). Items for both scales were scored on a scale from 1 (Strongly disagree) to 4 (Strongly agree). Because the outcomes were modeled at both the teacher- and school-level, we assessed reliability at both levels. Reliability was satisfactory at the teacher ($\omega_{Aus} = 0.83$ and $\omega_{Eng} = 0.86$ for job satisfaction; $\omega_{Aus} = 0.85$ and $\omega_{Eng} = 0.87$ for occupational commitment) and school ($\omega_{Aus} = 0.96$ and $\omega_{Eng} = 0.97$ for job satisfaction; $\omega_{Aus} = 0.99$ and $\omega_{Eng} = 0.98$ for occupational commitment) levels. The two scales also demonstrated adequate variance at the school-level (ICC = 0.14 for job satisfaction; ICC = 0.04 for occupational commitment; Bliese et al., 2018).

Teacher Characteristics

Teacher gender was coded 0 for female, 1 for male. *Teaching experience* was a continuous variable measured in years.

DATA ANALYSIS

All analyses were conducted using M*plus* 8.4 (Muthén and Muthén, 2017). In our analyses, teacher (TCHWGT) and school weights (SCHWGT) were applied to adjust teacher and school scores to account for the probabilities of selection and participation at the different stages of sampling (see OECD, 2014 for full details about the weighting procedure). In addition, the clustering of teachers within schools was accounted for in single-level modeling by using the cluster command in M*plus*. The robust maximum likelihood (MLR) estimator was used in all models. Missing data were 5–8% for all variables (except disruptive student behavior, which was 17%). Missing data were handled with full information maximum likelihood procedures.

Preliminary Analyses

Confirmatory factor analyses (CFA) were run for each country separately to obtain estimates of correlations among the two background characteristics, the five demands and resources, and the two outcomes (see the **Supplementary Material** for further details). We also ran measurement invariance tests using multigroup CFA to ensure that the ratings obtained in the Australian and English samples could be considered to be comparable. These models involved the latent factors for the five demands and resources (self-efficacy was modeled as a higher-order factor defined from three first-order factors), which were directly estimated from their items. We examined four models that were progressively more restrictive: configural (allowing all parameters to be freely estimated across the two countries), metric (loadings fixed to equality across countries), scalar (loadings and intercepts fixed across countries), and latent variance-covariance (loadings, intercepts, variances, and covariances fixed across countries) invariance models. We looked for changes in RMSEA across the models of 0.015 or less and for changes in CFI and TLI of 0.01 or less to establish invariance (Cheung and Rensvold, 2002; Chen, 2007). Factor scores were saved from the most constrained measurement model that was found to be invariant (with background characteristics and mean scores of the outcomes included as auxiliary variables). These factors scores were used in the latent profile analyses (LPA). More precisely, we used factor scores for the two job demands (barriers to professional development and disruptive behavior), the two job resources (teacher collaboration and input in decision-making), and the personal resource (self-efficacy for teaching) as profile indicator variables in the single-level and multilevel LPAs described below. Using these factor scores, we then ran a multigroup (across countries) measurement model in order to standardize the L1 and L2 sampling weights separately for each country. This step was necessary because of the way the weights were prepared in the original data (i.e., separately by country; syntax for this step is available in the **Supplementary Material**). The within-country standardized weights were saved ("savedata") and used in all analyses as outlined below.

Single-Level LPA

For the single-level LPA conducted at the teacher level, we tested a range of solutions involving 1 through 8 profiles separately for Australia and England. Profile indicator variables were standardized ($M = 0$, $SD = 1$) for each country. All analyses relied on an assumption of conditional independence, meaning that any covariance between indicators is assumed to be entirely explained by the latent profile variable, given that we did not

have any *a priori* theoretical or empirical reason for relaxing this assumption (e.g., Meyer and Morin, 2016). Means and variances were allowed to differ across indicator variables and profiles.

Each model was estimated using at least 6,000 random start values, each allowed 100 iterations, and 100 final stage optimizations. We also verified that the best log-likelihood value was properly replicated for all models. Several indices were used to assess the fit of the different models. For the Akaike Information Criteria (AIC), Consistent Akaike Information Criteria (CAIC), Bayesian Information Criteria (BIC), and sample-size-adjusted Bayesian Information Criteria (SSA-BIC) smaller values reflect better fit. The *p*-value of the adjusted Lo–Mendell–Rubin Likelihood Ratio Test (*p*LMR) allows comparison of a *k*-profile model with a $k-1$ profile model to see if the former model results in an improvement in fit relative to the latter. Finally, we created elbow plots of the AIC, CAIC, BIC, and SSA-BIC indices. In these plots, the profile at which point the slope noticeably flattens is another indicator of an appropriate solution (Morin et al., 2016). We also report entropy where values closer to 1 reflect greater profile separation. Alongside fit indices, we used parsimony, conceptual relevance, and statistical adequacy to help determine the optimal solution for each country.

After determining the optimal solution for each country, we next undertook tests of profile similarity to determine the extent to which the profile solutions could be considered to be comparable across Australia and England (Morin et al., 2016). These tests were conducted in the following sequence (Morin et al., 2016): configural (testing that the appropriate number of profiles was the same in the two countries), structural (constraining the means of the profile indicators to be the same across the two countries), dispersion (constraining the variance of the profile indicators to be the same across the two countries), and distributional (constraining the relative size of the profiles to be equal across countries). As recommended by Morin et al. (2016) we considered that profile similarity was supported when two indicators out of the CAIC, BIC, and SSA-BIC were lower (or equal) for the more constrained models relative to the previous model in the sequence. More precisely, tests of profile similarity seek to assess whether observed variations in person-centered results represent meaningful cross-country differences or whether they can simply be assumed to reflect random sampling variations. In other words, person-centered interpretations should be based on examination of the most similar solution rather than on a less accurate check of solutions separately estimated across countries (Morin et al., 2016).

Following these initial profile similarity tests, two additional tests of similarity were also conducted in order to examine the equivalence of the associations between the predictors (i.e., background characteristics) and likelihood of profile membership (predictive similarity), as well as the equivalence of the associations between profile membership and the outcomes (explanatory similarity) across countries (Morin et al., 2016). For these tests, predictors and outcomes were added to the most similar model from the previous sequence. We first ran an unconstrained model in which associations between the profiles, and the predictors or outcomes were allowed to vary across country. Then, a second model constrained these associations to be equal across country. The precise role of predictors (i.e., gender, teaching experience) was further examined using a multinomial logistic regression, using one latent profile as a reference group (Vermunt, 2010). Unstandardized beta coefficients, standard errors, and odds ratios (ORs) are presented from this analysis. ORs with a value greater than one indicate the increased likelihood of membership in a profile (compared with a reference profile) for every unit of increase in the predictor variable. The reverse is true for ORs $<$ 1. Outcome levels were compared across profiles using the M*plus* MODEL CONSTRAINT option, which relies on the multivariate delta method for tests of statistical significance (e.g., Muthén and Muthén, 2017).

Multilevel LPA

In phase two, our aim was to extend the teacher-level (level 1, L1) findings to consider the extent to which school-level (level 2; L2) profiles could be identified. More precisely, these analyses sought to identify school profiles characterized by distinct proportions of teacher profiles. Thus, rather than estimating profiles (as in our single level analyses) based on the means and variance of profile indicators, this second set of analyses identified school profiles based on the relative frequency of the various categories of the L1 latent profiles (thus mathematically corresponding to a L2 latent class analysis; e.g., Morin and Litalien, 2017). To maintain the stability, and cross-country equivalence, of the previously identified teacher-level profiles, we relied on the manual 3-step approach described by Litalien et al. (2019; also see Morin and Litalien, 2017). This approach was necessary given the way the L2 analyses were conducted, allowing the L1 profiles to be "predicted" by the L2 profiles, thus making it impossible to implement any direct constraint on the relative frequency of occurrence of the L1 profiles across countries (i.e., distributional similarity; see Morin and Litalien, 2017). Additional details on the implementation of this approach are provided in the **Supplementary Material**.

Multilevel LPA solutions including 1 to 8 school-level profiles were first estimated separately in both countries. Each model was estimated using at least 6,000 random start values, each allowed 100 iterations, and 100 final stage optimizations. We also verified that the best log-likelihood value was properly replicated for all models. Model selection relied on the same criteria as used for the single level LPA, with the exception of the LMR, which is not available for multilevel LPA. After determining the optimal solution for each country, we ran L2 profile similarity tests to determine the extent to which the L2 profiles could be considered to be comparable across Australia and England. For this, we extrapolated upon the tests developed by Morin et al. (2016) for single level LPA as well as those developed by Eid et al. (2003) for single level latent class analyses. These tests were conducted in the following sequence: configural (testing that the appropriate number of L2 profiles was the same in the two countries), structural (constraining the relative frequency of the L1 profiles defining the L2 profiles to be the same across the two countries), and distributional (constraining the relative size of the

L2 profiles to be equal across countries). Finally, we tested for L2-explanatory similarity by adding school-average outcomes to the most similar model determined in the L2 profile similarity tests. Annotated M*plus* input files for the estimation of these models are provided in the **Supplementary Material**.

RESULTS

Preliminary Analyses

Table 1 displays reliability coefficients and descriptive statistics for each sample at L1 and L2. These data indicate appropriate reliability. Within-country CFAs provided correlations between all variables examined in the study (the resulting correlation matrix is available in the **Supplementary Material**). Tests of measurement invariance supported the equivalence of the factor loadings, item intercepts, latent variances, and latent covariances across countries with all ΔRMSEA $\leq$ 0.015 (Chen, 2007), and ΔCFI and ΔTLI $\leq$ 0.01 (Cheung and Rensvold, 2002; i.e., configural invariance RMSEA = 0.04, CFI = 0.93, and TLI = 0.92; metric invariance RMSEA = 0.04, CFI = 0.93, and TLI = 0.93; scalar invariance RMSEA = 0.04, CFI = 0.93, and TLI = 0.92; latent variance-covariance invariance RMSEA = 0.04, CFI = 0.93, and TLI = 0.92). Factor scores were thus obtained from the most constrained model (latent variance-covariance invariance) to use in the LPA and sampling weights were standardized within each country.

Single-Level LPA

The fit statistics associated with the solutions including 1 to 8 profiles estimated separately in Australia and England are reported in **Table 2**. For both countries, the AIC, CAIC, BIC, SSA-BIC decreased as additional profiles were added. For Australia, the *p*LMR supported the 6-profile solution. For England, the *p*LMR failed to support any specific solution. Elbow plots were also consulted for both countries (see the **Supplementary Material**) and showed a slight flattening of the slope around 5-profiles in both countries. Thus, the fit statistics themselves failed to pinpoint any specific solution in both countries, but suggest that the optimal solution might be close to five profiles in both countries. Thus, to support the selection of the optimal solution, we considered the conceptual relevance, parsimony, and meaningfulness of the 5-profile solution, together with that of the adjacent 4- and 6-profile solutions. A first noteworthy observation was that examination of these solutions already revealed a high level of similarity across country, thus providing early support for configural similarity. When we compared the 4-profile solution with the 5-profile solution, this examination revealed that the additional profile was meaningful in its own right in both countries, presenting a well-differentiated shape relative to the other profiles. However, adding a sixth profile did not appear to contribute additional information to the solution, simply resulting in the arbitrary division of one of the profiles into to smaller profiles presenting a similar shape. The 5-profile solution was thus retained for both countries, and submitted to more systematic tests of profile similarity.

The results from the tests of profile similarity conducted across countries are reported in **Table 3**. These results revealed that, each step of the sequence of similarity tests resulted in a decrease in the value of the CAIC, BIC, and SSA-BIC, thus supporting the complete (configural, structural, dispersion, and distributional) similarity of the solution across countries. A graphical representation of this final 5-profile solution of distributional similarity is presented in **Figure 2**, and detailed results are reported in the **Supplementary Material**.

Teachers corresponding to profile 1 (12% of the sample) reported low barriers to professional development, very low disruptive behavior, high teacher collaboration, high teacher input, and high self-efficacy. This profile was thus labeled *Low-Demand-Flourisher* to reflect this adaptive blend of low job demands and high job and personal resources. Teachers corresponding to profile 2 (17% of the sample) reported low barriers to professional development, average disruptive behavior, high teacher collaboration, high teacher input, and

TABLE 1 | Reliabilities and descriptive statistics for Australia and England.

	Australia			England		
	ω_{Aus}	*M*	*SD*	ω_{Eng}	*M*	*SD*
Teacher-level						
Gender	—	1.42	0.49	—	1.36	0.48
Teacher experience	—	16.02	11.06	—	12.71	9.42
Barriers to professional development	0.74	2.24	0.58	0.74	2.23	0.58
Disruptive student behavior	0.89	1.98	0.74	0.90	1.95	0.75
Teacher collaboration	0.75	4.91	1.04	0.73	4.72	1.06
Teacher input	0.86	2.66	0.67	0.87	2.64	0.69
Teacher self-efficacy	0.86	3.27	0.48	0.86	3.37	0.45
Job satisfaction	0.83	3.23	0.58	0.86	3.06	0.63
Occ. commitment	0.85	3.18	0.63	0.87	3.13	0.66
School-level						
School-average job satisfaction	0.96	3.22	0.26	0.97	3.04	0.26
School-average occ. commitment	0.99	3.18	0.23	0.98	3.12	0.20

ω = *Coefficient omega. Occ. commitment = Occupational commitment.*

TABLE 2 | Fit statistics and entropy for Australia and England.

	Log-likelihood	Free parameters	AIC	CAIC	BIC	SSA-BIC	Entropy	*p*LMR
Australia – Single-level								
(1) Profile	−45484.08	10	90988.15	91065.81	91055.81	91024.03	—	—
(2) profiles	−42500.10	21	85042.20	85205.28	85184.28	85117.55	0.84	<0.001
(3) Profiles	−40642.59	32	81349.19	81597.69	81565.69	81464.00	0.88	<0.001
(4) Profiles	−39425.46	43	78936.92	79270.85	79227.85	79091.21	0.79	<0.001
(5) Profiles	−38387.83	54	76883.66	77303.01	77249.01	77077.41	0.80	<0.001
(6) Profiles	−37853.14	65	75836.28	76341.06	76276.06	76069.50	0.81	<0.001
(7) Profiles	−37347.07	76	74846.15	75436.35	75360.35	75118.84	0.80	*ns*
(8) Profiles	−36786.70	87	73747.40	74423.02	74336.02	74059.56	0.81	<0.001
Australia – Multilevel								
(1) Profile	−9778.15	4	19564.30	19595.36	19591.36	19578.65	0.68	—
(2) Profiles	−9703.77	9	19425.54	19495.43	19486.43	19457.83	0.64	—
(3) Profiles	−9684.58	14	19397.16	19505.88	19491.88	19447.39	0.61	—
(4) Profiles	−9672.83	19	19383.66	19531.21	19512.21	19451.83	0.61	—
(5) Profiles	−9665.04	24	19378.07	19564.45	19540.45	19464.19	0.63	—
(6) Profiles	−9660.65	29	19379.31	19604.51	19575.51	19483.36	0.62	—
(7) Profiles	−9657.87	34	19383.75	19647.79	19613.79	19505.74	0.61	—
(8) Profiles	−9655.10	39	19388.20	19691.07	19652.07	19528.14	0.60	—
England – Single-level								
(1) Profile	−17027.26	10	34074.53	34142.36	34132.36	34100.59	—	—
(2) Profiles	−15929.48	21	31900.96	32043.40	32022.40	31955.68	0.87	<0.001
(3) Profiles	−15191.69	32	30447.38	30664.44	30632.44	30530.77	0.90	<0.001
(4) Profiles	−14717.76	43	29521.52	29813.20	29770.20	29633.57	0.80	0.01
(5) Profiles	−14321.04	54	28750.09	29116.38	29062.38	28890.81	0.81	<0.001
(6) Profiles	−14059.53	65	28249.05	28689.96	28624.96	28418.44	0.83	<0.001
(7) Profiles	−13889.55	76	27931.09	28446.62	28370.62	28129.15	0.83	0.003
(8) Profiles	−13730.97	87	27635.95	28226.09	28139.09	27862.67	0.84	0.003
England – Multilevel								
(1) Profile	−3746.68	4	7501.37	7528.50	7524.50	7511.79	0.67	—
(2) Profiles	−3711.05	9	7440.09	7501.14	7492.14	7463.55	0.69	—
(3) Profiles	−3703.07	14	7434.14	7529.10	7515.10	7470.62	0.70	—
(4) Profiles	−3697.53	19	7433.06	7561.94	7542.94	7482.57	0.65	—
(5) Profiles	−3695.35	24	7438.69	7601.49	7577.49	7501.24	0.67	—
(6) Profiles	−3694.20	29	7446.40	7643.12	7614.12	7521.98	0.69	—
(7) Profiles	−3693.21	34	7454.42	7685.05	7651.05	7543.02	0.69	—
(8) Profiles	−3692.64	39	7463.27	7727.82	7688.82	7564.91	0.70	—

AIC = Akaike Information Criteria. CAIC = Consistent Akaike Information Criteria. BIC = Bayesian Information Criteria. SSA-BIC = sample-size-adjusted Bayesian Information Criteria. pLMR = Lo–Mendell–Rubin Likelihood Ratio Test. ns = non-significant.

high self-efficacy. This profile was thus labeled *Mixed-Demand-Flourisher* to reflect the mixed blend of low to average job demands, coupled with high job and personal resources. Teachers corresponding to profile 3 (21% of the sample) reported slightly below average barriers to professional development, average disruptive behavior, high teacher collaboration, high teacher input, and average self-efficacy. We labeled this profile *Job-Resourced-Average* to reflect the above average job resources, and average job demands and self-efficacy. Teachers corresponding to profile 4 (15% of the sample) reported average barriers to professional development, average disruptive behavior, average teacher collaboration, average teacher input, and average self-efficacy. We labeled this profile *Balanced-Average* to reflect the matching average levels observed across all demands and resources. Teachers corresponding to profile 5 (34% of the sample) reported high barriers to professional development, high disruptive behavior, low teacher collaboration, low teacher input, and low self-efficacy. We labeled this profile *Struggler* to reflect this blend of high job demands, and low job and personal resources.

We next tested the predictive and explanatory similarity of this solution by including predictors (i.e., gender, teaching experience) and outcomes to this final model of distributional similarity. In terms of predictive similarity, the CAIC, BIC and SSA-BIC decreased when equality constraints across countries were included for the predictive paths (see **Table 3**), thus supporting the equivalence of these predictions across countries. The results from the multinomial regression paths estimated

TABLE 3 | Tests of profile similarity across Australia and England.

	Log-Likelihood	Free parameters	AIC	CAIC	BIC	SSA-BIC	Entropy
Single-level LPA							
Configural	−58082.70	109	116383.39	117264.52	117155.52	116809.14	0.87
Structural (means)	−57975.96	84	116119.92	116798.96	116714.96	116448.02	0.84
Dispersion (means and variances)	−58004.77	59	116127.55	116604.49	116545.49	116358.00	0.86
Distributional (means, variances, probabilities)	−58009.74	55	116129.49	116574.09	116519.09	116344.31	0.86
Predictive similarity							
- Unconstrained across country	−56650.91	21	113343.82	113513.13	113492.13	113425.40	0.86
- Constrained across country	−56660.69	13	113347.38	113452.20	113439.20	113397.89	0.86
Explanatory similarity							
- Unconstrained across country	−72421.46	27	144896.92	145115.18	145088.18	145002.38	0.88
- Constrained across country	−72548.36	17	145130.72	145268.15	145251.15	145197.12	0.88
Multilevel LPA							
Configural	−13739.54	19	27517.07	27670.67	27651.67	27591.29	0.74
Structural (proportion of L1 profiles)	−13750.02	11	27522.03	27610.96	27599.96	27565.00	0.72
Distributional (proportion of L2 profiles)	−13753.72	10	27527.43	27608.27	27598.27	27564.49	0.72
Explanatory similarity							
- Unconstrained across country	−13616.73	20	27273.45	27435.13	27415.13	27351.57	0.76
- Constrained across country	−13665.34	16	27362.68	27492.02	27476.02	27425.18	0.76

AIC = Akaike Information Criteria. CAIC = Consistent Akaike Information Criteria. BIC = Bayesian Information Criteria. SSA-BIC = sample-size-adjusted Bayesian Information Criteria.

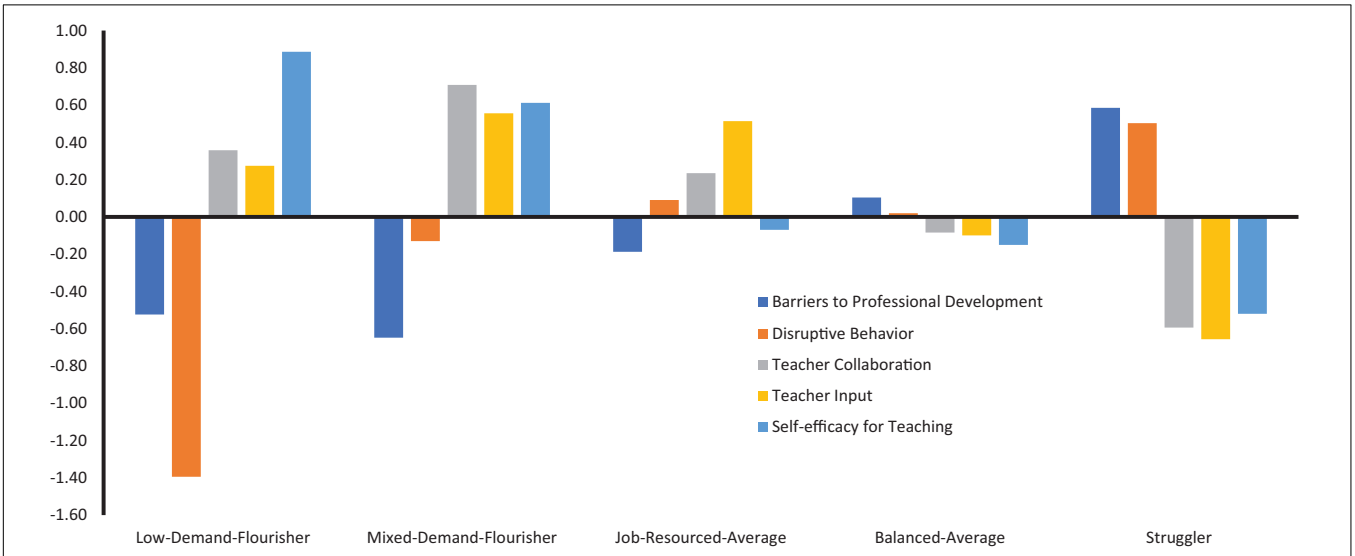

FIGURE 2 | Single-level LPA results of distributional similarity showing teacher profiles for both countries.

as part of this model are reported in **Table 4**. These results first show that Male teachers were less likely to correspond to the Low-Demand-Flourisher and Mixed-Demand-Flourisher profiles than to the Job-Resourced-Average, Balanced-Average, or Struggler profiles. Teachers with more extensive teaching experience were more likely to correspond to the Low-Demand-Flourisher profile than to the Mixed-Demand-Flourisher, Job-Resourced-Average, Balanced-Average, and Struggler profiles. Teachers with more extensive teaching experience were also more likely to correspond to the Mixed-Demand-Flourisher and Job-Resourced-Average profiles than to the Struggler profile. Taken together, male teachers and less experienced teachers were more likely to correspond to the apparently less desirable profiles.

In terms of explanatory similarity, the CAIC, BIC and SSA-BIC increased when profile-specific outcome levels were constrained to be equal across countries (see **Table 3**), suggesting that these outcomes associations were not equivalent across countries. To investigate the differences, we compared the means of the outcomes within and across the two countries. For Australia, there were significant differences in means across all profiles for both outcomes ($p < 0.05$). The

TABLE 4 | The role of teacher covariates in predicting profile membership in both countries from the single-level LPA.

	b	*SE*	*OR*	*b*	*SE*	*OR*
	Low-demand-flourisher vs. Mixed-demand-flourisher			**Low-demand-flourisher vs. Job-resourced-average**		
Gender (F/M)	0.17	0.10	1.18	−0.33**	0.09	0.72
Teaching experience	0.03**	0.01	1.03	0.04**	0.01	1.04
	Low-Demand-Flourisher vs. Balanced-Average			**Low-Demand-Flourisher vs. Struggler**		
Gender (F/M)	−0.41**	0.10	0.66	−0.46**	0.09	0.63
Teaching experience	0.04**	0.01	1.04	0.05**	0.01	1.05
	Mixed-Demand-Flourisher vs. Job-Resourced-Average			**Mixed-Demand-Flourisher vs. Balanced-Average**		
Gender (F/M)	−0.50**	0.09	0.61	−0.58**	0.10	0.56
Teaching experience	0.01	0.01	1.01	0.01	0.01	1.01
	Mixed-Demand-Flourisher vs. Struggler			**Job-Resourced-Average vs. Balanced-Average**		
Gender (F/M)	−0.62**	0.09	0.54	−0.08	0.09	0.92
Teaching experience	0.02**	0.01	1.02	0.01	0.01	1.01
	Job-Resourced-Average vs. Struggler			**Balanced-Average vs. Struggler**		
Gender (F/M)	−0.13	0.07	0.88	−0.05	0.08	0.95
Teaching experience	0.01**	0.01	1.01	0.01	0.01	1.01

*$p \leq 0.05$; ** $p \leq 0.01$; b = multinomial logistic regression coefficient; SE = standard error of the coefficient; OR = odds ratio; For gender, females were coded 0 and males were coded 1.*

Mixed-Demand-Flourisher profile displayed the highest levels of job satisfaction (M = 3.64) and commitment (M = 3.54), followed by the Low-Demand-Flourisher profile (M = 3.53 for job satisfaction and M = 3.46 for commitment), then by the Job-Resourced-Average profile (M = 3.36 and M = 3.23), followed by the Balanced-Average profile (M = 3.13 and M = 3.09), and finally by the Struggler profile (M = 2.88 and M = 2.89).

For England, all mean comparisons were also statistically significant ($p \leq 0.05$), with one exception. Starting with job satisfaction, the Mixed-Demand-Flourisher (M = 3.53) displayed the highest levels, followed by the Low-Demand-Flourisher profile (M = 3.41), then by the Job-Resourced-Average profile (M = 3.23), followed by the Balanced-Average profile (M = 3.06), and finally by the Struggler profile (M = 2.59). For occupational commitment, the Mixed-Demand-Flourisher displayed the highest levels (M = 3.56), followed by the Low-Demand-Flourisher profile (M = 3.44) and then equally by the Job-Resourced-Average (M = 3.18) and Balanced-Average (M = 3.13) profiles, which did not differ from one another, and finally by the Struggler profile (M = 2.75).

Finally, we also tested mean differences within matching profiles across the two countries. All profiles except one from the Australian sample displayed higher levels of job satisfaction when compared with the matching profiles from the English sample ($p < 0.01$). The exception was related to the Balanced-Average profile, which displayed similar levels of job satisfaction in both countries. For occupational commitment, there were no significant differences between countries, with the sole exception of the Struggler profile for which levels of occupational commitment were higher in Australia than in England ($p < 0.01$).

Multilevel LPA

The fit statistics associated with the multilevel solutions including 1 to 8 profiles estimated separately in Australia and England are reported in **Table 2** (corresponding elbow plots are reported in the **Supplementary Material**). In both countries, the solution including two school-level (L2) profiles resulted in the lowest value for the CAIC and BIC. The SSA-BIC was also lowest for the 2-profile solution in England, and although it kept in decreasing up until the 3-profile solution in Australia, the elbow plot displayed a clear flattening after the 2 profile solution. Finally, although the AIC kept on decreasing until the 5-profile solution in Australia and the 4-profile solution in England, this decrease also showed a marked flattening after 2 profiles in both countries. Taken together, these statistical results thus strongly support the 2-profile solution in both countries. Examination of this solution, together with an examination of the adjacent solutions, supported the theoretical value of considering two profiles, but not that of adding a third profile, which did not seem to bring any new information. Accordingly, a solution with 2 school-level profiles was selected as the final solution for both countries.

The results from the L2 tests of profile similarity conducted across countries are reported in **Table 3**. These results revealed that, each step of the sequence of similarity tests resulted in a decrease in the value of the CAIC, BIC, and SSA-BIC, thus supporting the complete (configural, structural, and distributional) similarity of the solution across countries. A graphical representation of this final 2-profile solution of distributional similarity is presented

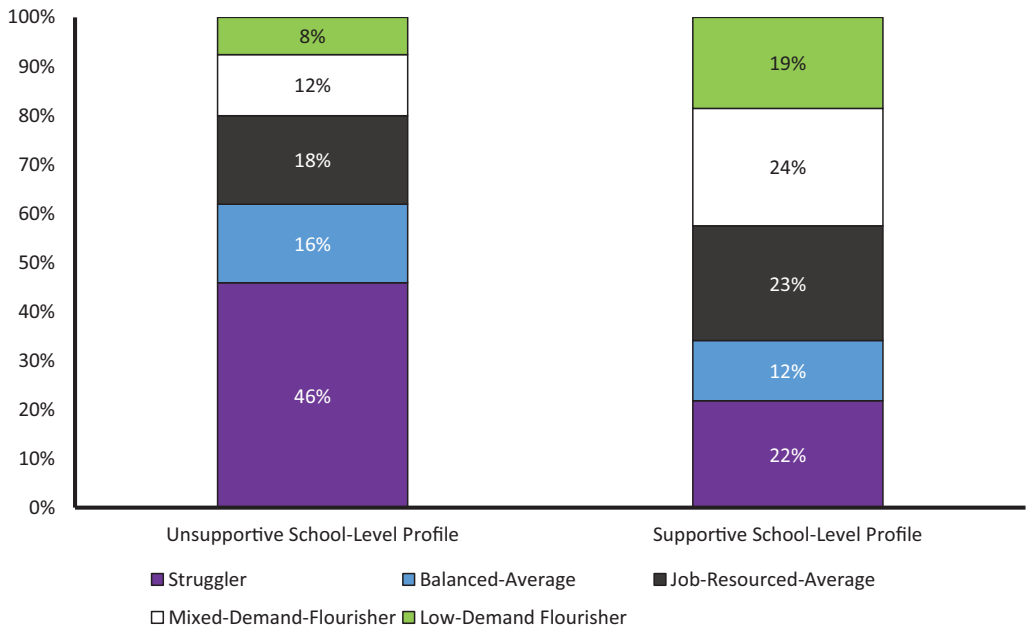

FIGURE 3 | Multilevel LPA results (with L1 and L2 distributional constraints) showing the school-level profiles for both countries.

in **Figure 3**. Examination of this solution suggested the presence of one Unsupportive school profile (58% of the schools) and one Supportive school profile (42% of the schools). The Unsupportive school profile included a high proportion of members from the Struggler (46%) profile, followed by the Job-Resourced-Average (18%), Balanced-Average (16%), Mixed-Demand-Flourisher (12%), and Low-Demand-Flourisher (8%) profiles. In contrast, the Supportive school profile included a higher proportion of members from the Mixed-Demand-Flourisher (24%) profile, followed by the Job-Resourced-Average (23%), Struggler (22%), Low-Demand-Flourisher (19%), and Balanced-Average (12%) profiles.

We next tested the explanatory similarity of this solution by including outcomes to this final model of distributional similarity. As for the single level models, the CAIC, BIC and SSA-BIC increased when profile-specific outcome levels were constrained to be equal across countries (see **Table 3**), suggesting that these outcomes associations were not equivalent across countries. To investigate the differences, we compared the school-level means of the outcomes within and across the two countries. For Australia, the Unsupportive school profile displayed significantly lower ($p < 0.01$) school-average job satisfaction ($M = 3.01$) and occupational commitment ($M = 3.04$) than the Supportive school profile ($M = 3.37$ and $M = 3.27$, respectively). The same was true for England, where the Unsupportive school profile also displayed significantly lower ($p \leq 0.01$) school-average job satisfaction ($M = 2.79$) and occupational commitment ($M = 2.92$) than the Supportive school profile ($M = 3.16$ and $M = 3.22$, respectively). Finally, the two L2-profiles from the Australian sample displayed higher levels of job satisfaction than the matching profiles estimated in the English sample ($p < 0.01$). In addition, the Unsupportive school profile from the Australian sample displayed higher levels of occupational commitment than the matching profile from the English sample ($p \leq 0.01$), but no differences in occupational commitment were observed for the Supportive school profile across country.

DISCUSSION

We used person-centered analyses to identify demand-resource profiles among teachers and schools across representative samples from two different countries. In phase one, five teacher-level demand-resources profiles were identified in both the Australian and English samples: the Low-Demand-Flourisher, Mixed-Demand-Flourisher, Job-Resourced-Average, Balanced-Average, and Struggler. Results showed that male teachers and less experienced teachers were more likely to be members of less adaptive profiles in both countries (e.g., the Struggler profile). The profiles were also associated with different levels on both well-being outcomes (highest for the Mixed-Demand-Flourisher profile, lowest for the Struggler profile) in each country. More precise cross-country comparisons showed that profile-specific levels of job satisfaction were higher in the Australian sample than levels observed in the matching English profiles, whereas few differences in occupational commitment were evident.

In phase two, we extended our analyses to the school-level. In both countries, we found evidence of two profiles of schools: a Supportive school profile comprising relatively similar levels of the Mixed-Demand-Flourisher, Job-Resourced-Average, Struggler, and Low-Demand-Flourisher profiles, and an

Unsupportive school profile comprising much higher levels of the Struggler profile. For both countries, the Supportive school profile was associated with significantly higher levels of school-average job satisfaction and occupational commitment. Cross-country differences in outcomes levels were also apparent, and are discussed below.

Findings of Note From the Teacher-Level Results

As noted, five demand-resource profiles were evident at the teacher-level for both Australia and England. Two of these profiles were named flourishers, the Low-Demand-Flourisher and Mixed-Demand-Flourisher, and in combination made up almost one-third of the sample. Given the nature of these two profiles (low or average job demands, high job and personal resources), this is a positive finding. The third profile, the Job-Resourced-Average, represented around one-fifth of the sample. Thus, around 1 in 5 teachers in the sample experienced average job demands and self-efficacy, but simultaneously felt well-supported (above average job resources). The fourth profile, the Balanced-Average, represented 15% of the sample. These teachers appear to experience relatively similar (and average) levels of demands and resources. Finally, the Struggler represented around one-third of each sample. Thus, 1 in 3 teachers in the sample experienced high job demands, low job resources, and low self-efficacy.

Taken together, the findings are important as they provide insight into the groups of teachers that work in schools. The findings are also significant because of the close overlap in the profiles demonstrated across the two countries—which may have occurred given the historical and socio-cultural similarities of the two contexts (e.g., Bulle, 2011). Importantly, there is some commonality between our five profiles and those found in prior research among other types of employees (e.g., Van den Broeck et al., 2012). However, unlike Van den Broeck et al. (2012), we did not find evidence of a low job demands and low job resources profile, nor a high job demands and high job resources profile. At the same time, the Balanced-Average profile (with average demands and resources) seems to exist between these two extremes. Thus, it may be that matching levels of demands and resources are apparent among teachers, just not at such extreme levels as among other employees. This finding might have occurred because demands and resources are often reflective of the broader means available to a school. Schools with high demands are often under-resourced, whereas the reverse is true for schools with low demands (e.g., Muijs et al., 2004). As such, balanced profiles at extreme levels may be less likely to occur among teachers than in other professions. Additional research with different samples is needed to further test this. In future research, it will also be important to model the extent of change over time in profile membership to ascertain the stability of such groupings.

The findings provide a nuanced understanding that complements prior variable-centered results showing that job and personal resources are typically positively correlated. For instance, Collie et al. (2012) found that teacher self-efficacy and input in decision making were positively correlated (r = 0.22). Conversely, our findings showed that while personal resources and job resources appeared to be at similar levels for most profiles (i.e., the two Flourishers, the Balanced-Average, the Struggler), they were less aligned for the Job-Resourced-Average profile. Taken together, these findings highlight the merits of person-centered research given that it is able to access the experiences of different subpopulations of teachers.

In both countries, the background characteristics were associated with the profiles in similar ways, showing that male teachers were more likely to correspond to the Job-Resourced-Average, Balanced-Average, or Struggler profiles than to either of the two Flourisher profiles. It is possible this finding occurred because, relative to male teachers, female teachers have been shown to report stronger perceptions of job resources generally (Skaalvik and Skaalvik, 2018), greater teacher collaboration (Ronfeldt et al., 2015), and greater self-efficacy (Klassen and Chiu, 2010). Less experienced teachers were typically more likely to be in the Job-Resourced-Average, Balanced-Average, or Struggler profiles than in either of the two Flourisher profiles. These findings are not surprising. Teaching is a complex job and beginning teachers must navigate this complexity with less knowledge and less first-hand experience to draw upon (Mansfield et al., 2014). It is thus understandable that less experienced teachers tended to appear within profiles characterized by higher demands and lower resources. Of importance, this finding highlights the salience of providing higher levels of support for early career teachers as shown in other research (e.g., De Neve et al., 2015).

Turning to the outcomes, the Mixed-Demand-Flourisher profile displayed the highest levels of job satisfaction and occupational commitment in both countries. This was followed by the Low-Demand-Flourisher profile. These results are understandable given the nature of the two Flourisher profiles, and given that low demands and high resources have been associated with well-being in prior variable-centered research (e.g., Skaalvik and Skaalvik, 2018). The significant differences between the two Flourisher profiles in our results are also interesting and hold potential contributions for theory. Notably, the boosting process in JD-R theory stipulates that resources play an even stronger role in promoting well-being outcomes when demands are high (Bakker and Demerouti, 2017). Perhaps the Mixed-Demand-Flourisher profile displayed more positive outcomes than the Low-Demand-Flourisher profile because with relatively higher levels of disruptive student behavior, the resources available to the Mixed-Demand-Flourisher profile became more important for their well-being. Conversely, the Low-Demand-Flourisher profile experienced low job demands and thus the resources were perhaps less relevant and then less salient for well-being (Bakker et al., 2007). What is interesting about these results is they suggest that average levels of job demands are not necessarily problematic. As along as demands are outweighed by resources, teachers may still experience high levels of well-being. Future research is needed to test whether this suggestion replicates with other samples.

Moving along to the other profiles, the Job-Resourced-Average profile typically displayed the third highest levels of the well-being outcomes, followed by the Balanced-Average profile. In comparing these two profiles, the major differences occurred in job resources, which underscores prior research on the importance of contextual supports for teachers (e.g., Lee and Nie, 2014; Desrumaux et al., 2015). The Job-Resourced-Average profile had access to greater job resources, which may have meant a boost to their well-being. In contrast, the relatively equal levels of demands and resources for the Balanced-Average profile may have meant the boosting effect was not as evident (because this profile did not have particularly high resources to drawn upon). Future research is needed to disentangle these results and see if they are replicated.

Finally, the Struggler profile displayed the lowest outcomes. Alongside the mismatch between (high) demands and (low) resources that this profile experienced, low collaboration and input in decision-making may mean that teachers in this profile experience a reduced sense of autonomy (e.g., Ryan and Deci, 2017) and lower professional fit at work (e.g., Kristof, 1996). Both a sense of autonomy and professional fit have been identified as important for teachers' job satisfaction (e.g., Collie et al., 2016). Moreover, if teachers' feel their professional growth is not being fostered (e.g., via barriers to professional development), they typically have less desire to remain in the profession (e.g., Ford et al., 2019).

Taken together, the findings involving the outcomes complement knowledge gained from prior variable-centered research by providing a more nuanced understanding of the associations that demands and resources have with outcomes related to well-being. Prior research has clearly documented that job demands are typically negatively associated with well-being, whereas the reverse is true for resources (e.g., Skaalvik and Skaalvik, 2018). In a complementary manner, our results highlight how varying combinations of demands and resources are also related to differences in levels of well-being. This knowledge provides a clearer picture of the simultaneous role of multiple factors in teachers' work.

In terms of cross-country comparisons, significantly higher levels of job satisfaction were evident in most of the profiles from the Australian sample when compared with the matching profiles from the English sample. For occupational commitment, only the Struggler profile displayed higher levels in the Australian sample than in the English sample. Additional research is needed to understand precisely why these findings occurred, but it may be related to increases in compliance and reductions in professional autonomy that have been documented in England over the past decade (e.g., Adams, 2017)—such working conditions are known to be unsatisfying for teachers (e.g., Skaalvik and Skaalvik, 2014). Perhaps occupational commitment was not significantly lower among the teachers from England (except for the Struggler profile) because this construct is reflective of teachers' longer-term motives for being in the teaching profession (e.g., helping students), which are somewhat more distal from day-to-day working conditions. Given that job satisfaction is associated with lower motivation to quit the profession (Skaalvik and Skaalvik, 2017) and lower burnout (Malinen and Savolainen, 2016), the low levels of this outcome among the English sample might have longer-term ramifications. More precisely, even though the English teachers were committed to the profession, they may not be functioning as effectively as possible at work due to their lower job satisfaction and this may result in negative outcomes later. Going forward, it will be important for longitudinal research to explore this.

Findings of Note From the School-Level Results

As noted earlier, large scale datasets from surveys like TALIS enable insights into phenomena at a school-level that are often inaccessible with smaller datasets. In our study, two profiles were evident at the school-level in both countries. The first we called the Unsupportive school profile and comprised 58% of participating teachers. The second was the Supportive school profile (comprising around 42% of participating teachers), and included a substantially greater proportion of the two Flourisher profiles, and a smaller proportion of the Struggler profile than the Unsupportive school profile. The higher proportion of schools corresponding to the Unsupportive school profile is in accord with the growing attention toward the escalating demands faced by many teachers and schools in Australia and England. Indeed, there is growing attention to teachers' workload, burnout, and attrition from government working groups (UK Department for Education, 2018; Parliament of Australia, 2019) and not-for-profit organizations (e.g., Education Support Partnership, 2018) in both countries.

Notably, our results are some of the first to examine demand-resource profiles at the school-level among teachers, and the first to test profile similarity at the school-level across country. Our findings are important because limited research has examined multilevel associations using JD-R theory (Bakker and Demerouti, 2018). Moreover, whereas multilevel variable-centered research reveals knowledge about how variables are associated at the school-level, the current study identifies the types of teachers that predominate in different schools. By revealing types of schools, our findings add to knowledge about particular variables that might be important at the school level (e.g., school climate; Klassen et al., 2012). Taken together, our teacher- and school-level results provide important knowledge relevant for practice and policy on teacher well-being. More precisely, by considering all findings simultaneously, it is apparent that efforts to address not only the individual, but also the school, are warranted (further details below). Indeed, attending to one level (teacher or school), but not to both levels simultaneously, might result in efforts that are less effective in the longer term.

Turning to the outcomes, the Supportive school profile was associated with significantly higher levels of school-average job satisfaction and occupational commitment. This finding contributes to the literature by highlighting that the particular combination of teacher types within a school is associated with school-average teacher well-being. It is possible that this finding occurred because the Supportive school profile was characterized by a higher proportion of the two Flourisher profiles, which had

more positive outcome levels. When teachers' resources outweigh their demands, then teachers are better equipped to undertake their work and are likely more satisfied with their jobs and committed to their profession (e.g., Simbula et al., 2012; Collie et al., 2015; Skaalvik and Skaalvik, 2018). It is also possible that social contagion occurs in schools with more satisfied and committed teachers, helping to further promote these outcomes—though additional research is needed to examine this. Of note, these findings suggest that there may be merit in promoting school-wide approaches to addressing teacher well-being (in addition to efforts focused on the individual; discussed below).

Implications for Practice

Because the current study is one of the first to consider demand-resource profiles among teachers, we emphasize that more research is needed to ascertain the generalizability of our findings. Nonetheless, we do provide some tentative suggestions for practice. Notably, a key contribution of person-centered research is that it allows implications for practice that are targeted more closely to the needs of particular subpopulations of teachers and schools. For example, teachers in the Flourisher profiles would likely benefit from efforts to further boost (or at least maintain) their access to resources. For the Struggler profile, efforts may want to focus on reducing demands and increasing resources. Turning to the school-level, efforts targeted at the Unsupportive school profile may focus on school-wide efforts to boost resources and reduce job demands, whereas efforts focused on the Supportive school profile may focus on further boosting job and personal resources.

In terms of practices that can help boost resources and reduce demands, reducing barriers to professional development is likely important. As noted earlier, barriers include budget constraints, availability, support from leadership, and time (e.g., Kwakman, 2003). Efforts by schools to reduce each of these barriers will support teachers in accessing the professional development and learning they require. For example, in rural and remote areas, schools might focus more on online professional learning (e.g., see Broadley, 2010). By reducing barriers to professional development, teachers may gain more access to additional strategies and learning opportunities that are relevant to the other demands and resources we examined. For example, professional learning via reflection can help to improve teachers' capacity to effectively navigate disruptive behavior and improve teacher-student relationships (e.g., Spilt et al., 2012).

In terms of resources, engagement in professional learning communities and instructional rounds (observing other teachers) are two effective methods for encouraging teacher collaboration (e.g., Durksen et al., 2017) and boosting teacher self-efficacy (e.g., Chong and Kong, 2012). Finally, a growing body of research has shown that input in decision-making is important for teachers (e.g., Klassen et al., 2012). School leaders can promote teachers' input by listening to teachers' needs, attempting to understand issues from teachers' perspectives, and seeking teachers' suggestions for decisions that are made (e.g., Ware and Kitsantas, 2011).

Limitations and Future Directions

Consideration of several limitations is important in interpreting findings from the current study. First, the use of TALIS 2013 provides significant strengths (e.g., nationally representative samples, matching teacher and school-level data). Nonetheless, this type of data does come with some limitations. In particular, our study was cross-sectional in nature, which means that we are unable to ascertain causal ordering between the profiles and the outcomes, nor whether teachers' profile membership fluctuates over time. Longitudinal modeling (e.g., latent transition analysis) will be an important avenue to explore in future. Second, we focused on five demands and resources. As noted earlier, our selection of factors was firmly based in JD-R theory, conceptual reasoning, and prior empirical research. Nonetheless, in future it will be important to consider a different range of factors to see what else is salient in teachers' experiences. Third, our study employed data from teachers. Of course, teachers' perceptions are essential given that it is their interpretations that may influence their well-being outcomes. Nonetheless, in future it will be interesting to consider other markers of demands and resources to see how perceptions of demands and resources align with measures taken from other informants (e.g., school principals). Fourth, our study was conducted among teachers from Australia and England. Examining the extent to which similar profiles can identified, or not, in other countries (including non-English speaking countries) should be an important upcoming research focus. As noted earlier, teachers' experiences of demands and resources in Australia and England are mirrored in many other countries (e.g., Skaalvik and Skaalvik, 2018). Nonetheless, more evidence is needed before it is possible to argue that these results apply to broader contexts.

CONCLUSION

The aim of the current study was to establish whether different profiles of teachers could be identified based on their experiences of demands and resources, and, if so, to ascertain which profiles are more aligned with well-being. We conducted our examination at the teacher- and school-level among teachers from Australia and England. Findings revealed five teacher profiles that were similar across the two countries: the Low-Demand-Flourisher, Mixed-Demand-Flourisher, Job-Resourced-Average, Balanced-Average, and Struggler. Notably, the profiles differed in relation to two well-being outcomes, with the Mixed-Demand-Flourisher typically evincing the highest levels of job satisfaction and occupational commitment. Two school-level profiles that were similar in both countries were identified based on the prevalence of the five teacher profiles: the Unsupportive and Supportive school profiles. Of note, the Supportive school profile was associated with higher school-average teacher job satisfaction and occupational commitment. Taken together, the findings yield novel understanding about different subgroups of teachers and schools, and hold implications for practice at the teacher- and school-level.

DATA AVAILABILITY STATEMENT

The datasets used in our research are available from the OECD.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

AJSM was supported by a grant from the Social Sciences and Humanities Research Council of Canada (435-2018-0368).

ACKNOWLEDGMENTS

The authors want to thank Tihomir Asparouhov for his invaluable help in devising the procedure used to test profile similarity at the school level.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.00626/full#supplementary-material

REFERENCES

Adams, G. (2017). Using a narrative approach to illuminate teacher professional learning in an era of accountability. *Teach. Teach. Educ.* 67, 161–170. doi: 10.1016/j.tate.2017.06.007

Arens, A. K., and Morin, A. J. S. (2016). Relations between teachers' emotional exhaustion and students' educational outcomes. *J. Educ. Psychol.* 108, 800–813. doi: 10.1037/edu0000105

Bakker, A. B., and Demerouti, E. (2017). Job demands–resources theory: taking stock and looking forward. *J. Occup. Health Psychol.* 22, 273–285. doi: 10.1037/ocp0000056

Bakker, A. B., and Demerouti, E. (2018). "Multiple levels in job demands-resources theory: implications for employee well-being and performance," in *Handbook of Wellbeing*, eds E. Diener, S. Oishi, and L. Tay (Salt Lake City, UT: DEF Publishers).

Bakker, A. B., Hakanen, J. J., Demerouti, E., and Xanthopoulou, D. (2007). Job resources boost work engagement, particularly when job demands are high. *J. Educ. Psychol.* 99, 274–284. doi: 10.1037/0022-0663.99.2.274

Bliese, P., Maltarich, M., and Hendricks, J. (2018). Back to basics with mixed-effects models: nine take-away points. *J. Bus. Psychol.* 33, 1–23. doi: 10.1007/s10869-017-9491-z

Broadley, T. (2010). Digital revolution or digital divide: will rural teachers get a piece of the professional development pie? *Educ. Rural Aust.* 20, 63–76.

Bulle, N. (2011). Comparing OECD educational models through the prism of PISA. *Comp. Educ.* 47, 503–521. doi: 10.1080/03050068.2011.555117

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Struct. Equ. Modeling* 14, 464–504. doi: 10.1080/10705510701301834

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equ. Modeling* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Chong, W. H., and Kong, C. A. (2012). Teacher collaborative learning and teacher self-efficacy: the case of lesson study. *J. Exp. Educ.* 80, 263–283. doi: 10.1080/00220973.2011.596854

Collie, R. J., Granziera, H., and Martin, A. J. (2018). Teachers' perceived autonomy support and adaptability: an investigation employing the job demands-resources model as relevant to workplace exhaustion, disengagement, and commitment. *Teach. Teach. Educ.* 74, 125–136. doi: 10.1016/j.tate.2018.04.015

Collie, R. J., and Martin, A. J. (2017). Adaptive and maladaptive work-related motivation among teachers: a person-centered examination and links with well-being. *Teach. Teach. Educ.* 64, 199–210. doi: 10.1016/j.tate.2017.02.010

Collie, R. J., Shapka, J. D., and Perry, N. E. (2012). School climate and social-emotional learning: predicting teacher stress, job satisfaction, and efficacy. *J. Educ. Psychol.* 104, 1189–1204. doi: 10.1037/a0029356

Collie, R. J., Shapka, J. D., Perry, N. E., and Martin, A. J. (2015). Teachers' beliefs about social-emotional learning: identifying teacher profiles and their relations with job stress and satisfaction. *Learn. Instr.* 39, 148–157. doi: 10.1016/j.learninstruc.2015.06.002

Collie, R. J., Shapka, J. D., Perry, N. E., and Martin, A. J. (2016). Teachers' psychological functioning in the workplace: exploring the roles of contextual beliefs, need satisfaction, and personal characteristics. *J. Educ. Psychol.* 108, 788–799. doi: 10.1037/edu0000088

De Neve, D., Devos, G., and Tuytens, M. (2015). The importance of job resources and self-efficacy for beginning teachers' professional learning in differentiated instruction. *Teach. Teach. Educ.* 47, 30–41. doi: 10.1016/j.tate.2014.12.003

Desrumaux, P., Lapointe, D., Ntsame Sima, M., Boudrias, J.-S., Savoie, A., and Brunet, L. (2015). The impact of job demands, climate, and optimism on well-being and distress at work: what are the mediating effects of basic psychological need satisfaction? *Eur. Rev. Appl. Psychol.* 65, 179–188. doi: 10.1016/j.erap.2015.06.003

Dicke, T., Stebner, F., Linninger, C., Kunter, M., and Leutner, D. (2018). A longitudinal study of teachers' occupational well-being: applying the job demands-resources model. *J. Occup. Health Psychol.* 23, 262–277. doi: 10.1037/ocp0000070

Durksen, T. L., Klassen, R. M., and Daniels, L. M. (2017). Motivation and collaboration: the keys to a developmental framework for teachers' professional learning. *Teach. Teach. Educ.* 67, 53–66. doi: 10.1016/j.tate.2017.05.011

Duyar, I., Gumus, S., and Bellibas, M. S. (2013). Multilevel analysis of teacher work attitudes: the influence of principal leadership and teacher collaboration. *Int. J. Edu. Manag.* 27, 700–719. doi: 10.1108/IJEM-09-2012-0107

Education Support Partnership (2018). *Teacher Wellbeing Index 2018*. London: Education Support Partnership.

Eid, M., Langeheine, R., and Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis. *J. Cross Cult. Psychol.* 34, 195–210. doi: 10.1177/0022022102250427

Fackler, S., and Malmberg, L. (2016). Teachers' self-efficacy in 14 OECD countries: teacher, student group, school and leadership effects. *Teach. Teach. Edu.* 56, 185–195. doi: 10.1016/j.tate.2016.03.002

Ford, T. G., Olsen, J., Khojasteh, J., Ware, J., and Urick, A. (2019). The effects of leader support for teacher psychological needs on teacher burnout, commitment, and intent to leave. *J. Educ. Admin.* 57, 615–634. doi: 10.1108/JEA-09-2018-0185

Gallup (2014). *State of America's Schools: The Path to Winning Again in Education*. Washington, DC: Gallup.

Granziera, H., and Perera, H. N. (2019). Relations among teachers' self-efficacy beliefs, engagement, and work satisfaction: a social cognitive view. *Contemp. Educ. Psychol.* 58, 75–84. doi: 10.1016/j.cedpsych.2019.02.003

Hakanen, J. J., Bakker, A. B., and Schaufeli, W. B. (2006). Burnout and work engagement among teachers. *J. Sch. Psychol.* 43, 495–513. doi: 10.1016/j.jsp.2005.11.001

Klassen, R. M., and Chiu, M. M. (2010). Effects on teachers' self-efficacy and job satisfaction: teacher gender, years of experience, and job stress. *J. Educ. Psychol.* 102, 741–756. doi: 10.1037/a0019237

Klassen, R. M., Perry, N. E., and Frenzel, A. C. (2012). Teachers' relatedness with students: an underemphasized component of teachers' basic psychological needs. *J. Educ. Psychol.* 104 150–165. doi: 10.1037/a0026253

Klusmann, U., Kunter, M., Trautwein, U., Lüdtke, O., and Baumert, J. (2008). Teachers' occupational well-being and quality of instruction: the important role of self-regulatory patterns. *J. Educ. Psychol.* 100, 702–715. doi: 10.1037/0022-0663.100.3.702

Kristof, A. L. (1996). Person-organization fit: an integrative review of its conceptualizations, measurement, and implications. *Pers. Psychol.* 49, 1–49. doi: 10.1111/j.1744-6570.1996.tb01790.x

Kwakman, K. (2003). Factors affecting teachers' participation in professional learning activities. *Teach. Teach. Educ.* 19, 149–170. doi: 10.1016/S0742-051X(02)00101-4

Lee, A. N., and Nie, Y. (2014). Understanding teacher empowerment: teachers' perceptions of principal's and immediate supervisor's empowering behaviours, psychological empowerment and work-related outcomes. *Teach. Teach. Educ.* 41, 67–79. doi: 10.1016/j.tate.2014.03.006

Leithwood, K., Harris, A., and Hopkins, D. (2008). Seven strong claims about successful school leadership. *Sch. Leadersh. Manag.* 28, 27–42. doi: 10.1080/13632430701800060

Litalien, D., Gillet, N., Gagné, M., Ratelle, C. F., and Morin, A. J. S. (2019). Self-determined motivation profiles among undergraduate students: a robust test of profile similarity as a function of gender and age. *Learn. Individ. Differ.* 70, 39–52. doi: 10.1016/j.lindif.2019.01.005

Mäkikangas, A., Tolvanen, A., Aunola, K., Feldt, T., Mauno, S., and Kinnunen, U. (2018). Multilevel latent profile analysis with covariates: identifying job characteristics profiles in hierarchical data as an example. *Organ. Res. Methods* 21, 931–954. doi: 10.1177/1094428118760690

Malinen, O., and Savolainen, H. (2016). The effect of perceived school climate and teacher efficacy in behavior management on job satisfaction and burnout: a longitudinal study. *Teach. Teach. Educ.* 60, 144–152. doi: 10.1016/j.tate.2016.08.012

Mansfield, C., Beltman, S., and Price, A. (2014). 'I'm coming back again!' the resilience process of early career teachers. *Teach. Teach.* 20, 547–567. doi: 10.1080/13540602.2014.937958

McInerney, D. M., Korpershoek, H., Wang, H., and Morin, A. J. S. (2018). Teachers' occupational attributes and their psychological wellbeing, job satisfaction, occupational self-concept and quitting intentions. *Teach. Teach. Educ.* 71, 145–158. doi: 10.1016/j.tate.2017.12.020

Meyer, J. P., Allen, N. J., and Smith, C. A. (1993). Commitment to organizations and occupations: extension and test of a three-component conceptualization. *J. Appl. Psychol.* 78, 538–551. doi: 10.1037/0021-9010.78.4.538

Meyer, J. P., and Morin, A. J. S. (2016). A person-centered approach to commitment research: theory, research, and methodology. *J. Organ. Behav.* 37, 584–612. doi: 10.1002/job.2085

Meyer, J. P., Morin, A. J. S., Stanley, L. J., and Maltin, E. R. (2019). Profiles of organizational and occupational commitment: implications for well-being and career intentions. *Teach. Teach. Educ.* 79, 100–111. doi: 10.1016/j.tate.2018.09.009

Moeller, J., Ivcevic, Z., White, A. E., Menges, J. I., and Brackett, M. A. (2018). Highly engaged by burned out: intraindividual profiles in the US workforce. *Career Dev. Intl.* 23, 86–105. doi: 10.1108/CDI-12-2016-0215

Morin, A. J. S., Boudrias, J.-S., Marsh, H. W., McInerney, D. M., Dagenais-Desmarais, V., and Litalien, D. (2017). Complementary variable- and person-centered approaches to the dimensionality of psychometric constructs: approaches to psychological wellbeing at work. *J. Bus. Psychol.* 32, 395–419. doi: 10.1007/s10869-016-9448-7

Morin, A. J. S., and Litalien, D. (2017). *Longitudinal Tests of Profile Similarity and Latent Transition Analyses*. Montreal, QC: Substantive Methodological Synergy Research Laboratory.

Morin, A. J. S., Meyer, J. P., Creusier, J., and Biétry, F. (2016). Multiple-group analysis of similarity in latent profile solutions. *Organ. Res. Methods* 19, 231–254. doi: 10.1177/1094428115621148

Morin, A. J. S., Meyer, J. P., McInerney, D. M., Marsh, H. W., and Ganotice, F. (2015). Profiles of dual commitment to the occupation and organization: relations to wellbeing and turnover intentions. *Asia Pacific J. Manag.* 32, 717–744. doi: 10.1007/s10490-015-9411-6

Muijs, D., Harris, A., Chapman, C., Stoll, L., and Russ, J. (2004). Improving schools in socioeconomically disadvantaged areas – A review of research evidence. *Sch. Eff. Sch. Improv.* 15, 149–175. doi: 10.1076/sesi.15.2.149.30433

Muthén and Muthén (2017). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

OECD (2009). *Creating Effective Teaching and Learning Environments: First Results From TALIS*. Paris: OECD.

OECD (2014). *TALIS 2013 Technical Report*. Paris: OECD.

OECD (2019). *Education Policy Outlook 2019: Working Togethers to Help Students Achieve Their Potential*. Paris: OECD.

Parliament of Australia (2019). *Inquiry Into the Status of Teaching: Summary*. Canberra, ACT: Parliament of Australia.

Perera, H. N., Calkins, C., and Part, R. (2019). Teacher self-efficacy profiles: determinants, outcomes, and generalizability across teaching level. *Contemp. Educ. Psychol. 58*, 186–203. doi: 10.1016/j.cedpsych.2019.02.006

Reeves, P. M., Pun, W. H., and Chung, K. S. (2017). Influence of teacher collaboration on job satisfaction and student achievement. *Teach. Teach. Educ.* 67, 227–236. doi: 10.1016/j.tate.2017.06.016

Ronfeldt, M., Farmer, S. O., McQueen, K., and Grissom, J. A. (2015). Teacher collaboration in instructional teams and student achievement. *Am. Educ. Res. J.* 52, 475–514. doi: 10.3102/0002831215585562

Ryan, R. M., and Deci, E. L. (2017). *Self-Determination Theory: Basic Psychological Needs in Motivation, Development, and Wellness*. New York, NY: Guilford Press.

Scherer, R., Nilsen, T., and Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: an investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Front. Psychol.* 7:00110. doi: 10.3389/fpsyg.2016.00110

Schleicher, D. J., Hansen, S. D., and Fox, K. E. (2011). "Job attitudes and work values," in *APA Handbook of Industrial and Organizational Psychology: Vol. 3. Maintaining, Expanding, and Contracting the Organization*, ed. S. Zedeck (Washington, DC: American Psychological Association), 137–189.

Simbula, S., Panari, C., Guglielmi, D., and Fraccaroli, F. (2012). Teachers' well-being and effectiveness: the role of the interplay between job demands and job resources. *Procedia Soc. Behav. Sci.* 69, 729–738. doi: 10.1016/j.sbspro.2012.11.467

Skaalvik, E., and Skaalvik, S. (2018). Job demands and job resources as predictors of teacher motivation and well-being. *Soc. Psychol. Educ.* 21, 1251–1275. doi: 10.1007/s11218-018-9464-8

Skaalvik, E. M., and Skaalvik, S. (2014). Teacher self-efficacy and perceived autonomy: relations with teacher engagement, job satisfaction, and emotional exhaustion. *Psychol. Rep.* 114, 68–77. doi: 10.2466/14.02.PR0.114k14w0

Skaalvik, E. M., and Skaalvik, S. (2017). Still motivated to teach? A study of school context variables, stress and job satisfaction among teachers in senior high school. *Soc. Psychol. Educ.* 20, 15–37. doi: 10.1007/s11218-016-9363-9

Spilt, J. L., Koomen, H. M. Y., Thijs, J. T., and van der Leij, A. (2012). Supporting teachers' relationships with disruptive children: the potential of relationship-focused reflection. *Attach. Hum. Dev.* 14, 305–318. doi: 10.1080/14616734.2012.672286

Taylor, I. M., Ntoumanis, N., and Standage, M. (2008). A self-determination theory approach to understanding the antecedents of teachers' motivational strategies in physical education. *J. Sport Exerc. Psychol.* 30, 75–94. doi: 10.1123/jsep.30.1.75

Tschannen-Moran, M., and Woolfolk Hoy, A. E. (2001). Teacher efficacy: capturing an elusive construct. *Teach. Teach. Educ.* 17, 783–805. doi: 10.1016/S0742-051X(01)00036-1

UK Department for Education (2018). *Making Data Work: Report of the Teacher Workload Advisory Group*. London: UK Department for Education.

Urick, A. (2016). The influence of typologies of school leaders on teacher retention. *J. Educ. Adm.* 54, 434–468. doi: 10.1108/JEA-08-2014-0090

Van den Broeck, A., De Cuyper, N., Luyckx, K., and De Witte, H. (2012). Employees' job demands–resources profiles, burnout and work engagement: a person-centred examination. *Econ. Ind. Democracy* 33, 691–706. doi: 10.1177/0143831X11428228

Vangrieken, K., Dochy, F., Raes, E., and Kyndt, E. (2015). Teacher collaboration: a systematic review. *Educ. Res. Rev.* 15, 17–40. doi: 10.1016/j.edurev.2015.04.002

Vermunt, J. K. (2010). Latent class modeling with covariates: two improved three-step approaches. *Polit. Anal.* 18, 450–469. doi: 10.1093/pan/mpq025

von der Embse, Nathaniel P, Pendergast, L. L., Segool, N., Saeki, E., and Ryan, S. (2016). The influence of test-based accountability policies on school climate and teacher stress across four states. *Teach. Teach. Educ.* 59, 492–502. doi: 10.1016/j.tate.2016.07.013

Ware, H. W., and Kitsantas, A. (2011). Predicting teacher commitment using principal and teacher efficacy variables: an HLM approach. *J. Educ. Res.* 104, 183–193. doi: 10.1080/00220671003638543

Watt, H. M. G., and Richardson, P. W. (2008). Motivations, perceptions, and aspirations concerning teaching as a career for different types of beginning teachers. *Learn. Instr.* 18, 408–428. doi: 10.1016/j.learninstruc.2008.06.002

Watt, H. M. G., Richardson, P. W., Klusmann, U., Kunter, M., Beyer, B., Trautwein, U., et al. (2012). Motivations for choosing teaching as a career: an international comparison using the FIT-choice scale. *Teach. Teach. Educ.* 28, 791–805. doi: 10.1016/j.tate.2012.03.003

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 August 2020
doi: 10.3389/fpsyg.2020.02257

Different Patterns of Relationships Between Principal Leadership and 15-Year-Old Students' Science Learning: How School Resources, Teacher Quality, and School Socioeconomic Status Make a Difference

Cheng Yong Tan[1], Peng Liu[2] and Wai Lun Vincent Wong[1]*

[1] Faculty of Education, The University of Hong Kong, Pokfulam, Hong Kong, [2] Faculty of Education, University of Manitoba, Winnipeg, MB, Canada

OPEN ACCESS

Edited by:
Ronnel B. King,
The Education University of Hong Kong, Hong Kong

Reviewed by:
Joni Tzuchen Tang,
National Taiwan University of Science and Technology, Taiwan
Daniela Raccanello,
University of Verona, Italy

****Correspondence:***
Cheng Yong Tan
tancy@hku.hk

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *24 May 2020*
Accepted: *11 August 2020*
Published: *27 August 2020*

Citation:
Tan CY, Liu P and Wong WLV (2020) Different Patterns of Relationships Between Principal Leadership and 15-Year-Old Students' Science Learning: How School Resources, Teacher Quality, and School Socioeconomic Status Make a Difference. Front. Psychol. 11:2257. doi: 10.3389/fpsyg.2020.02257

The present study critically evaluates whether school leadership influences student learning homogenously regardless of school contexts. It examined relationships between four principal leadership variables (envisioning, instructional management, promoting professional development, empowerment) and two types of student outcomes (enjoyment in learning science, science achievement) in different school contexts [in terms of the availability of science resources, quality of science teachers, and school socioeconomic status (SES)]. The sample comprised 248,620 students and 9,370 principals in 35 developed countries who participated in the Programme for International Student Assessment (PISA) 2015. Latent profile analysis (LPA) showed that schools operated in three types of school contexts with different levels of science resources, proportion of quality science teachers, and school SES. There were also differences in the pattern of leadership practices across the three types of school contexts. Three-level hierarchical linear modeling (HLM) showed that among the four leadership variables, only instructional management was positively associated with students' enjoyment of science in schools with less science resources and quality science teachers. Therefore, instructional management had compensatory effects for students in less-endowed schools. In contrast, principal leadership related to envisioning, teacher professional development, and empowerment was not positively related to students' science learning in all three school contexts.

Keywords: academic achievement, learning enjoyment, PISA, principals, school context, school leadership

INTRODUCTION

There is a clarion call for school leadership researchers to pay more attention to understanding school contexts and contextualize research in these contexts that school leadership is enacted (Close and Raynor, 2010; Hallinger, 2018). There are many aspects that collectively characterize the complex environments that schools operate in but there are three contextual variables that

are especially important because of their proximity to teaching-and-learning. These aspects are namely, adequacy of school resources, quality of teachers, and average school SES level. Indeed, some schools may be more ready to implement leader-initiated changes to improve student learning because teachers are qualified and they embrace student-centered pedagogies (Ingvarson and Rowley, 2017), there are updated resources to support teaching-and-learning (Cohen et al., 2003), and students' parents have the resources to support school programs (Archer et al., 2015; Tan, 2018b).

The inexorable association between school leadership and contexts has two implications. First, school leaders have to arguably adapt to their contexts. For example, the contingency opportunities theory argues that the agency and effectiveness of leadership depends on environmental opportunities and challenges (Wasserman et al., 2010). Specifically, the theory asserts that leaders adapt organizational variables to their environments and that challenges can impede leadership imperatives. In the context of schools, principals have to adjust their leadership priorities to capitalize on opportunities and address challenges that impact teaching-and-learning. Second, school leadership practices may vary in their effectiveness according to contextual conditions.

Notwithstanding these implications, we do not have an informed understanding of the different types of school contexts that schools are operating in the first place. This may partly explain why there is a paucity of research that examines school leadership in context. To elaborate, a literature review indicates that there are few studies whereby leadership effects are examined for specific school contexts. For example, Jacobson's (2011) synthesis of findings from the school leadership literature and the International Successful School Principalship Project found that successful principals in challenging, high-poverty schools employed practices such as direction-setting, developing people, and redesigning the organization. They also used distributed leadership and professional self-renewal to sustain their school success over time. Stosich (2016) identified principal leadership and job-embedded support from experts as crucial for teachers undergoing attending professional development to translate their learning to instruction and collaboration in two high-poverty schools in the United States. Day et al. (2009) study of improving schools in England found that heads of schools operating in challenging contexts developed and sustained school policies for pupil behavior, motivation, and engagement; instructional standards; physical environments; teaching-and-learning improvements; and cultures of care and achievement. They also worked closely with parents and the community to improve student outcomes. Notman and Henry's (2011) qualitative study of six New Zealand primary and secondary school principals showed that principals highlighted the importance of developing learning and social assistance programs to address student and family needs for schools in lower-SES areas. The principals also demonstrated contingent leadership in response to external influences related to community support, inadequacy of financial resources from the central government, and the school's SES context, in order to sustain their leadership success over time. Walker and Ko's (2011) study reported that principals in Hong Kong schools leveraged the professional development of teachers and key staff to foster within-school alignment and the congruence between the school and government policies when faced with increasing accountability pressures. Researchers in these studies focus on specific school contexts instead of examining a range of contexts.

There are even fewer studies that compare leadership effects in different school contexts. For example, Hallinger and Murphy's (1986) study of elementary schools in California found that principals of effective, low-SES schools focused on students' mastery of basic skills in their school mission, exercised tight instructional control and task orientation, and buffered between home and school to minimize parental involvement. In contrast, principals of effective, high-SES schools focused on students' academic achievement, exercised low-to-moderate instructional control and relationship orientation, and promoted home-school cooperation. Tan (2018a) compared principal leadership effects on 15-year-old students' mathematics achievement using international data from PISA 2012. Results showed that the student sample could be divided into three sub-groups (disadvantaged, average, privileged) varying in their SES levels, teacher quality and educational resources in school, and parental academic expectations. Furthermore, principal leadership explained a greater proportion of the between-school achievement variance for disadvantaged as compared to other two sub-groups of students. Specifically, instructional management had the strongest positive association with student achievement for disadvantaged students. The other leadership variables were mostly negatively related to student achievement but the strength of association varied among the different sub-groups of students. Bottery et al. (2008) qualitative study of English headteachers and Hong Kong principals found the leadership of both groups of school leaders was influenced by contextual challenges arising from legislation, government inspection procedures, marketization, parental choice, and competition. For example, English headteachers were more adversely impacted by legislation and tough school inspections than were Hong Kong principals.

The present study addresses these knowledge gaps by identifying a typology of school contexts (as measured by the combination of variables measuring teacher quality, adequacy of school resources, and average school SES level) and examining how relationships between principal leadership and student science learning differ in these different contexts.

LITERATURE REVIEW

School Contexts

The literature review first discusses the three school context variables and their relationships with teaching and learning. Context is an elucidative element of school effectiveness and improvement at levels of both the school and entire education system (Harris et al., 2015). According to Hallinger et al. (1996), school context can be viewed as having three main aspects: school resources, teacher quality, and student SES.

The first aspect of school context is school resources. Findings from earlier studies showed that there is a consistently positive relationship between school resources and student achievement (Greenwald et al., 1996; Houtenville and Conway, 2008). According to Reynolds (2010), school effectiveness and improvement research has linked promoting school-community collaboration and developing the school culture (both contributing indirectly to school resources) to school turnaround and improvement. As for teacher expenditures (another investment in school resources), scholars including Hedges et al. (1994); Greenwald et al. (1996) believed that there is a strong and significant relationship between such expenditures and student achievement. However, there have also been differing voices in this discussion. Some researchers argued that the effect of school resources can be negligible or even negative (Hanushek, 1996; Häkkinen et al., 2003). After analyzing data from a large sample of matriculation examination scores of Finnish senior secondary school students from 1990 to1998, Häkkinen et al. (2003) concluded that changes in teaching expenditure did not make any difference to students' test scores. In fact, evidence did not support the idea that low levels of student performance in poor school districts was related to inadequate spending levels. Students in rural areas might be associated with higher per capita teaching expenditure but their performance still lagged that of students from urban areas.

Teachers' teaching experience and teaching quality, the second aspect of school contexts according to Hallinger et al. (1996), also has an influence on students' learning outcomes. A meta-analysis by Davis-Beggs (2013), synthesizing results from studies published from 1996 to 2009, reported a significant positive relationship between teachers' teaching experience and student achievement. Moreover, she suggested that the quality of professional development and the coherence of programs are the strongest predictors of high school students' academic achievement.

Student SES, Hallinger et al.' (1996) third aspect of school contexts, has also been regarded as the one of the main factors contributing to student achievement (Organization for Economic Co-operation and Development [OECD], 2004; Perry and McConney, 2010). Many reasons explain this relationship, such as demographics and other characteristics of students' parents, school community, and peers. For example, Zhang et al. (2011) showed that high-SES students had higher parental engagement than low-SES students, and that the level of parent engagement made a difference to student learning. As for school SES, Van Ewijk and Sleegers (2010) suggested that schools with high SES have higher average student achievement scores, so grouping high-SES students together could result in better student achievement than would be expected from the individual students alone.

The literature examined above has shown that school contexts, including factors such as school resources, teachers' quality and teaching experience, and students' SES, are crucial to student learning. Therefore, it is important for researchers to identify and account for school contexts in their studies.

Principal Leadership

The principal plays an important role in improving the quality of education in a school (Gurr et al., 2006; Waldron and McLeskey, 2010). Principal leadership can influence the trajectory of a school's development. The literature has emphasized the impact that school principals have on their students' learning and achievement through their leadership behaviors (Murillo and Hernández-Castilla, 2015; Day et al., 2016). Studies suggest that principal leadership is characterized by two major foci, namely instructional management and teacher capacity-building (Hoppey and McLeskey, 2013).

Leithwood et al.' (2006) conceptualization of four principal leadership functions encapsulate these two foci. First, principals galvanize schools' resources to realize the school vision, mission, and goals. The shared vision and mission enable resources to be aligned to achieve school goals (Murphy and Torre, 2015). For example, principals may focus on developing teachers and designing teachers' work to achieve school goals. Second, principals manage the instructional program by leading teachers in teaching–and-learning, promulgating effective instructional practices, and emphasizing students' holistic development (Hitt and Tucker, 2016). Envisioning and instructional management are arguably related to the notion of instructional leadership.

Third, principals facilitate teachers' professional development. This leadership practice builds on individual teacher strengths and needs, inculcates teacher responsibility for professional development, and cultivates teacher professional learning communities (Opfer and Pedder, 2011; Murphy, 2015). For example, principals can have regular discussions with teachers on instructional effectiveness and challenges faced during teaching. They can also provide professional development support to help struggling teachers (Yariv and Kass, 2019). Lastly, principals empower teachers by promoting collaborative decision-making processes. For example, they can distribute leadership roles to involve teachers in reviewing management practices and contributing to school improvement. Promoting professional development and teacher empowerment are related to the notion of enhancing teacher capacity.

Students' Enjoyment of Learning and Academic Achievement

Student learning comprises attitudinal and achievement indicators. Accordingly, the present study focuses on students' enjoyment in learning science and their science achievement. According to Fredrickson (2001), students who enjoy learning have "the urge to play, push the limits, and be creative" (p. 220). It is reasonable to expect then that students who enjoy their learning in a specific subject may be more interested to learn about different topics in that subject. Students' enjoyment in science learning is an important variable to study because it predicts their interest in science (Ainley and Hidi, 2014), engagement in science learning (Hampden-Thompson and Bennett, 2013), participation in science extracurricular (Lin et al., 2012), science achievement and career aspirations (Jeffries et al., 2020), and collaborative problem-solving (Camacho-Morles et al., 2019). Additionally, students who enjoy their learning

may internalize scientific principles (i.e., epistemological beliefs) more easily and therefore, are expected to have higher levels of science achievement (Acosta and Hsu, 2014).

According to the control-value theory, students' enjoyment of learning is dependent on their control and value appraisals (Pekrun, 2006; Pekrun et al., 2007). Control appraisals relate to students' competence perceptions of the degree to which learning outcomes are controllable (e.g., ability self-concepts). Value appraisals comprise intrinsic (e.g., perceiving learning activities as interesting and important in itself) and extrinsic (e.g., perceiving activities as important for achievement or relevant to daily life) components. More generally, students' achievement emotions can be understood in terms of object focus (activity or outcome), valence (positive or negative), activation (activating or deactivating), and prospective-retrospective dimensions. For instance, students' enjoyment of science learning comprises positive emotions of pleasure activated through participation in science learning activities. It can be associated with positive anticipation of expecting high test performance in science or retrospective joy after experiencing academic success in science.

Influence of Principal Leadership on Student Learning

The review will next discuss how principal leadership influences student learning in different contexts. When teachers participate in envisioning led by the principal, they are imbued with a shared sense of purpose, higher academic expectations, and commitment. These attributes, in turn, eventuates in a positive student learning climate in the school (Hendriks and Scheerens, 2013). In a related vein, when principals focus on instructional management, students have more learning opportunities because teachers may employ student-centered pedagogies that promote enjoyment in learning (Hendriks and Scheerens, 2013). Students can further benefit from these leadership practices (envisioning and instructional management) if the school context is favorable (e.g., well-resourced schools with qualified teachers) because they can learn from qualified, motivated teachers who employ engaging instructional practices with requisite, up-to-date educational resources. Students are more likely to receive reinforcement from their parents if the latter are more educated (e.g., higher-SES parents) because the latter are more likely to appreciate the importance of science learning and careers in science, technology, engineering, and mathematics or STEM (Archer et al., 2015).

From the perspective of control-value theory, these principal leadership functions (envisioning and instructional management) and supportive school conditions (well-resourced schools, qualified teachers, and higher-SES parents) are likely to eventuate in higher levels of student-perceived control (e.g., self-regulation and efficacy beliefs; Adams and Olsen, 2017; Zheng et al., 2017) and value (e.g., arising from autonomy support that teachers provide to support student learning; Adams and Olsen, 2019). These enhanced appraisals then enhance students' enjoyment of learning and achievement in science. Therefore, students whose principals exercise leadership in envisioning and instructional management in favorable school contexts are expected to have more positive science learning attitudes which eventuate in higher levels of science achievement.

Next, teachers in schools whereby principals invest in providing professional development may be more cohesive, professional, competent, and efficacious (Hendriks and Scheerens, 2013). A professional learning community comprising these teachers contributes to the school academic and improvement capacity, thereby benefiting student learning. However, if principals are promoting teacher professional development to address the problem of poor teacher quality (an unfavorable school contextual indicator), it is difficult to predict how this leadership practice will impact student learning for two reasons. First, multiple factors need to accompany teachers' professional development to contribute to student learning (Opfer and Pedder, 2011). These factors include simultaneous, mutually reinforcing changes in teachers' professional beliefs and practices after professional development and acquisition of subject-specific pedagogical skills. Therefore, professional development is more likely to be successful if the teacher participants are certified specialist teachers (e.g., qualified science teachers) in the first place. Second, professional development impacts teaching-and-learning positively if there is accompanying organizational support. In science education, if there is a severe deficit in science teacher quality and principals' science education knowledge, then professional development may not benefit students' science learning. The latter is evident in Lochmiller's (2016) qualitative study which demonstrated how high school principals who had limited understanding in mathematics and science education resorted to focusing on pedagogy instead of content when they gave feedback to teachers, relying on their past experience as teachers to inform the feedback provided, and using student assessment to make their feedback more meaningful.

Lastly, teacher empowerment creates positive school conditions (trust, care, risk-taking, continuous learning) that promote student learning (Hunzicker, 2012). Separately, students learn better when they are taught by qualified teachers, a school context variable examined in the present study. To illustrate, Woolnough's (1994) study of A-level students found that the quality of science teaching was one of the most important variables predicting whether students chose science as a subject. It can be argued that quality science teaching is more likely to come from well-qualified graduate science teachers who have the requisite science expertise and subject affiliation, enthusiasm in their science teaching, ability to contextualize science lessons in daily life, capacity to conduct structured yet stimulating science lessons, and willingness to spend time beyond class to have conversations with students about science. Therefore, we can expect qualified teachers, when empowered by principals, to make better decisions that cater to the learning needs of students. Empowered qualified teachers are also more adept at supporting different aspects of peers' professional growth such as sharing quality, relevant professional learning and providing support on pedagogical content knowledge issues (Wenner and Campbell, 2017), thereby benefiting student learning. In sum,

empowering qualified teachers is expected to benefit students in their learning.

The Present Study

The present study (a) identifies a typology of school contexts (as measured by the combination of variables measuring adequacy of school resources, teacher quality, and school SES); and (b) examines relationships between four core leadership practices (envisioning, instructional management, promoting professional development, empowerment; Leithwood et al., 2006) and student science learning using data from PISA 2015.

The PISA student sample comprises fifteen-year-old students in participating countries. This sample is appropriate for the present study because most students reach the end of their compulsory education by this age in many education systems (Organization for Economic Co-operation and Development [OECD], 2009), so it is important to ascertain the influence of school variables such as principal leadership and school contexts on the science achievement of this group of students. Students of this age group will also have developed the cognitive capacities to conduct appraisals of competence and value (Pekrun and Stephens, 2012; Pekrun, 2017) and therefore able to report their emotions (e.g., enjoyment of science learning) more accurately (Raccanello et al., 2018).

Principal, as opposed to teacher, leadership is examined as principal leadership constitutes the most important source of leadership in schools. For example, Leithwood and Jantzi's (2000) study found that principal, instead of teacher, leadership contributed to student engagement in Canadian schools. Day et al. (2009) reported that headteachers are regarded by teachers, governors, and parents as the key source of leadership impacting teaching processes in improving English schools. Principal leadership is measured using Leithwood et al. (2006) four core leadership functions because there is evidence that they characterize leadership practices of most school leaders, including successful principals leading schools in challenging contexts (Day et al., 2009; Jacobson, 2011).

The present study analyzed PISA 2015 which focused on students' science learning for three reasons. First, understanding how principal leadership and school contexts affect students' science learning is important given that the need for students to have the requisite literacy in STEM in modern society (Xie et al., 2015). Specifically, it is important to increase students' science participation to nurture what Irwin (2001) referred to as "science citizens" who can use scientific knowhow and technology to solve daily problems (Claussen and Osborne, 2013). Second, such students are also better placed to exploit opportunities in higher education and occupational markets in STEM disciplines characterizing KBEs in the longer term. Occupational opportunities in STEM are diverse, so students can select specific jobs that match their interests, afford them a better quality of life, and enhance their social mobility (Xie et al., 2015). Third, compared to other subjects such as reading, science learning is more susceptible to school teaching and resources (Reynolds et al., 2014) than family socialization, thereby enabling the present study to unravel contextualized principal leadership effects (if any) on student learning.

Students' enjoyment in science learning is examined in addition to their academic achievement in line with a more holistic conception of education. Students' learning attitudes are important because individuals need to be life-long learners in STEM economies and continuous learning requires students to enjoy learning. Furthermore, students' positive learning attitudes contribute to their academic achievement (Lam and Lau, 2014). Notwithstanding the salience of holistic learning outcomes, there are few studies examining how principals contribute to students' learning attitudes [e.g., self-regulated learning (Adams and Olsen, 2017), language self-efficacy (Zheng et al., 2017), engagement (Leithwood and Jantzi, 2000), and perceived autonomy support (Adams and Olsen, 2019)]. Therefore, the present study ascertains if principal leadership practices can improve students' learning attitudes and achievement in science.

Latent profile analysis (Oberski, 2016) is used to empirically derive a typology of schools operating in different contexts in the present study. It achieves this by examining the pattern of contextual variables in the sample, uncovering underlying heterogeneity, and identifying distinct sub-groups of schools varying in the three school contextual variables. This approach is useful because of its objective, data-driven approach and because it allows researchers to simultaneously incorporate multiple, correlated contextual indicators in identifying a typology of schools. The less-effective alternative approach will be to assign schools in the sample to arbitrary sub-groups based on *a priori* considerations one indicator at a time (classifying schools as low-, average-, and high-SES schools or as schools with low, average, or high proportions of qualified teachers) and drawing separate conclusions for each typology of schools.

MATERIALS AND METHODS

Participants

The sample comprised 248,620 students and 9,370 school principals in 35 OECD countries who participated in PISA 2015 (Organization for Economic Co-operation and Development [OECD], 2017). The majority of these students were in Grade 10 (55.9%) with the rest were from Grades 7−13. 69.7% of the schools were public schools while 14.8% were private schools (15.5% of the schools were unclassified). Participating students were selected to represent the complete population of 15-year-old students who were attending public or private schools in grade 7 or higher in the participating countries. PISA 2015 measured 15-year-old students' proficiency in applying their knowledge and skills learned in science (the focal domain) in addition to reading and mathematics. In addition, PISA 2015 collected background data from students, parents, principals, and teachers, on student/home/family, classroom, and school variables.

Measures

The present study analyzed data from principals' and students' responses to the School and Student Questionnaires,

respectively[1], and students' science performance[2]. Data on the following PISA 2015 variables were used in the analysis.

Science Resource Availability

The availability of science resources in schools (SciRes) was measured by summing up principals' responses (*Yes*, *No*) to eight items on the availability of resources for the science department (e.g., "Compared to similar schools, we have a well-equipped laboratory.") The items pertained to equipment in the science department, allocation of extra funding to science teaching, educational levels of science teachers, materials for laboratory and hands-on learning, laboratory support staff, and up-to-date science equipment.

Science Teacher Quality

The quality of science teachers (TrQua) was measured by the proportion of science teachers with a bachelor/master and science major qualifications in schools (principal-reported).

Principal Leadership

Principal leadership was measured with four scales using principals' responses to 13 items asking about the frequency of specific leadership behaviors with a six-point scale (1 = *Did not occur*, 2 = *1–2 times during the year*, 3 = *3–4 times during the year*, 4 = *Monthly*, 5 = *Weekly*, 6 = *More than once a week*). The scales corresponded to the four core principal leadership functions identified by Leithwood et al. (2006). The first scale (Envisioning) measured envisioning ($\alpha = 0.99$) — principals' framing and communication of school goals and curricular development — with four items pertaining to principals using student results to develop school academic goals, aligning teachers' professional development and work to school goals, and discussing school goals with teachers (e.g., "I use student performance results to develop the school's educational goals.") The second scale (Instructional-Mgmt; $\alpha = 0.98$) measured instructional management using three items related to principals promoting research-based teaching practices, praising teachers whose students were learning actively, and emphasizing to teachers the development of critical and social capacities in students (e.g., "I promote teaching practices based on recent educational research.").The third scale (Professional-Devt; $\alpha = 0.98$) measured principals' promotion of teachers' professional development using three items pertaining to principals taking the initiative to discuss problems teachers encountered in classrooms, paying attention to students' disruptive behavior, and solving classroom problems with teachers collaboratively (e.g., "When a teacher has problems in his/her classroom, I take the initiative to discuss matters."). The fourth scale (Empowerment; $\alpha = 0.98$) measured empowerment — principals' facilitation of teachers' participation in leadership — using three items related to principals engaging staff to participate in school decision-making, building a school culture of continuous improvement, and reviewing management practices (e.g., "I provide staff with opportunities to participate in school decision-making."). These four leadership scales corresponded to the four core principal leadership functions identified by Leithwood et al. (2006). Confirmatory factor analysis (CFA) showed that the four scales explained the variation in the 13 items satisfactorily [$\chi^2(59) = 3{,}627.96$, $p < 0.01$; CFI = 0.93; TLI = 0.91; RMSEA = 0.085; SRMR = 0.05].

Student and School SES

Student SES (StuSES) was measured by the index of economic, social, and cultural status computed by PISA 2015 (Organization for Economic Co-operation and Development [OECD], 2017). The index represented the first principal component derived from student data on parents' highest education level, parents' highest occupational status, and students' home possessions. Data on parents' highest education level were derived from student responses on their parents' highest levels of schooling completed (two questions for each parent). The response categories corresponded to "*no education*," "*primary education*," "*lower secondary*," "*vocational/pre-vocational upper secondary*," "*general upper secondary and/or non-tertiary post-secondary*," "*vocational tertiary*," and "*theoretically oriented tertiary and postgraduate*." Data on parents' highest occupational status were derived from students' responses on the nature of their parents' main jobs (two questions for each parent). PISA 2015 coded these data and mapped the codes onto the international socioeconomic index of occupational status (Ganzeboom and Treiman, 2003). Data on student home possessions were derived from student responses to three questions asking about the availability of different home resources such as study desk, own room, quiet place to study, computer for study, educational software, Internet connectivity, classic literature, poetry books, art works, books to support study, reference books, dictionary, books on art/music/design, televisions, cars, rooms with bath/shower, cell phones with Internet access, tablet computers, e-book readers, and musical instruments. The present study averaged students' SES levels within a school to obtain a measure of school SES (SchSES).

Student Gender

A variable identifying student gender (Male) was coded 1 and 0 for boys and girls, respectively.

Confucian Heritage Culture (CHC)

Countries were coded 1 if they were CHCs (Japan and Korea) and 0 otherwise.

Students' Enjoyment in Learning Science

Dependent variables comprised students' enjoyment in learning science and their science achievement. Students' enjoyment in learning science (Enjoy; $\alpha = 0.99$) was computed from student responses to five items measuring the extent to which they enjoyed learning science using a four-point scale (*Strongly disagree*, *Disagree*, *Agree*, *Strongly agree*). These items pertained to students learning, reading on, and working on science topics (e.g., "I generally have fun when I am learning < broad science > topics."). CFA showed satisfactory model fit with the five items specified to load on a single latent construct

[1]https://read.oecd-ilibrary.org/education/pisa-2015-assessment-and-analytical-framework/pisa-2015-background-questionnaires_9789264281820-9-en#page37

[2]https://read.oecd-ilibrary.org/education/pisa-2015-assessment-and-analytical-framework/pisa-2015-science-framework_9789264281820-3-en#page1

[$\chi^2(5)$ = 6,679.53, $p < 0.01$; CFI = 0.99; TLI = 0.99; RMSEA = 0.076; SRMR = 0.01].

Students' Science Achievement

Students' science achievement was the focal dependent variable measured in PISA 2015. Students were not administered the complete set of test items by design, and therefore each item had missing responses. This made it impossible to estimate achievement scores for each student. To overcome this limitation, PISA 2015 aggregated the results of individual students to produce scores for groups of students. It also used a set of ten "plausible values" (PV1-PV10) for each student to represent the estimated distribution of science scores of students similar to him or her in terms of responses to the assessment and background items.

Procedure

PISA 2015 used a two-stage stratified sampling design, with schools first selected from a national sampling frame of schools with probabilities proportional to size and students next selected from within each of the schools (Organization for Economic Co-operation and Development [OECD], 2017). PISA 2015 was sponsored internationally by the OECD. All participating countries followed standardized procedures outlined in the technical standards and manuals provided.

According to Organization for Economic Co-operation and Development [OECD] (2017), PISA was managed by a large international team comprising the PISA Governing Board (PGB), experts in working groups, National project Managers (NPMs). OECD Secretariat, Educational Testing Service (ETS) in the United States, and other external contractors. Specifically, the implementation of PISA was informed by a framework established by the PGB that comprised senior policymakers from all participating countries. The PGS oversaw the establishment of policy priorities and standards for developing indicators, for establishing assessment instruments, and for reporting results. Experts from participating countries formed working groups to relate PISA policy objectives to the best international available technical expertise to ensure that the instruments were internationally valid, culturally sensitive, sound in measurement potential, authentic, and educationally valid. An NPM was appointed in each of the participating countries to ensure that internationally established technical and administrative procedures were adopted. The NPMs developed and validated the assessment instruments and evaluated the survey results, analyses, and reports. The OECD Secretariat was overall responsible for the management of PISA. It provided day-to-day monitoring, provided secretariat services for the PGB, facilitated consensus-building among participating countries, and mediated between the PGB and international contractors. The ETS was responsible for the overall management of external contractors. These contractors were responsible for designing and implementing the surveys.

Missing Value Imputation

Missing values may compromise estimation efficiency and produce biased results (Cheema, 2014). Therefore, imputation by fully conditional specification was used to address the methodological challenge arising from missing values in the variables in SPSS 25. Imputation of variables with missing values using all other variables as predictors continued until the maximum number of iterations was reached. The imputed dataset then comprised imputed values at the maximum iteration. A set of missing values was imputed separately for school-level variables (SciRes 19.31% missing; TrQua 18.20% missing; Envisioning 12.28% missing; Instructional-Mgmt 12.27% missing; Professional-Devt 12.41% missing; and Empowerment 12.48% missing) and student-level variables (StuSES 2.31% missing; and Enjoy 9.42% missing). There were no missing values for other variables (Male, SchSES, CHC, and SciPV1-SciPV10).

LPA

Latent profile analysis (Oberski, 2016), using MPlus8, was employed to uncover underlying heterogeneity in schools within the sample and identify distinct groups of schools varying in the three school contextual variables (availability of science resources, quality of science teachers, school SES). There is no clear-cut way of determining the "correct" number of latent classes underlying the population but most analysts make their decision by examining different indicators such as information criteria indicators (Akaike Information Criteria or AIC, Baysesian Information Criteria or BIC, sample-adjusted BIC), entropy, and model parsimony (Nylund et al., 2007). A single-level LPA was performed for the school-level contextual variables because simulation studies have indicated that it yields results similar to those from multilevel LPA as long as the sample size is reasonably large and latent classes are distinct from each other (Park and Yu, 2018). In the context of the LPA for the present study (results to be reported in greater detail later), the total number of schools (9,370) was deemed to be sufficiently large with each latent class having between 1,782 (19.02% of total number of schools) and 5,713 schools (60.97% of schools) and mean levels of the three school context variables varying substantially among the latent classes identified.

Three-Level HLM

For each latent class of schools, three-level fixed effect HLM with full maximum likelihood estimation and robust standard errors was performed using HLM7.03 to examine relationships between school leadership and student outcomes (Raudenbush and Bryk, 2002). The use of robust standard errors enables unbiased standard errors to be computed even when model assumptions are violated and therefore, mitigates the problem of model misspecification. The four school leadership variables were entered into the models separately as the CFA results indicated high intercorrelations among them (0.58 to 0.90). All independent variables were rescaled by subtracting the grand mean of the entire sample from the respective raw scores for ease of interpretation. After the rescaling, each HLM parameter represents the "effect" of the respective variable for a student with values equal to the grand mean for the other variables. Senate weights for student- and school-level variables were included in the HLM.

Students and schools in each of latent classes were analyzed separately. The following set of nested HLM models was fitted for each student outcome (PV1-PV10, Enjoy). Three sets of nested models were estimated:

- Model 1 — baseline with no predictors;
- Model 2 — random intercepts model with student-level (Male, StuSES) and country-level (CHC) control variables; and
- Model 3 – random intercepts model with Model 2 variables and one of the four school leadership variables (Envisioning, Instructional-Mgmt, Professional-Devt, Empowerment).
- The mathematical formulation for Model 3 predicting students' science achievement is:

$$PV_{ijk} = \gamma_{000} + \gamma_{001}{}^{*}CHC_{k} + \gamma_{010}{}^{*}LeadershipVariable_{jk} + \gamma_{100}{}^{*}MALE_{ijk} + \gamma_{200}{}^{*}StuSES_{ijk} + r_{0jk} + u_{00k} + e_{ijk}$$

The mathematical formulation for Model 3 predicting students' enjoyment of science is:

$$Enjoy_{ijk} = \gamma_{000} + \gamma_{001}{}^{*}CHC_{k} + \gamma_{010}{}^{*}LeadershipVariable_{jk} + \gamma_{100}{}^{*}MALE_{ijk} + \gamma_{200}{}^{*}StuSES_{ijk} + r_{0jk} + u_{00k} + e_{ijk}$$

for student i from school j in country k. e_{ijk}, r_{0jk}, and u_{00k} = level-1,level-2, and level-3 residuals respectively. LeadershipVariable = Envisioning, Instructional-Mgmt, Professional-Devt, or Empowerment.

For the HLM involving students' science achievement, HLM7.03 estimates parameters for each plausible value separately and averages the ten estimates. It also combines the average of the sampling error from the ten estimates with the variance between the ten estimates multiplied by a factor related to the number of plausible values to yield the measurement error.

RESULTS

Typology of Schools

Various information criteria (AIC, BIC, and sample size-adjusted BIC) showed decreasing values when the number of latent classes was increased from 1 to 5 (**Table 1**). However, the percentage decrease in the information criteria was marginal when the number of latent classes was increased from 3 to 4 (1.04%, 1.00%, 1.02% for AIC, BIC, sample size-adjusted BIC, respectively) as compared to the case when the number of latent classes was increased from 1 to 2 (6.83%, 6.79%, 6.81%, respectively) or from 2 to 3 (5.01%, 4.97%, 4.99%, respectively). Entropy was the highest at.95 for 3 as compared to 1, 2, 4, or 5 latent classes. Therefore, a 3-class solution best characterized the typology of school contexts based on the three contextual indicators. Results of the final class counts and proportions for the latent classes based on their most likely latent class membership showed that 5,713 schools belonged to Class 1 (60.97%), 1,782 schools belonged to Class 2 (19.02%), and 1,875 schools belonged to Class 3 (20.01%). The classification quality was satisfactory as evident by the high mean "dominant" probability (i.e., highest probability of belonging to a class) of 0.98 and 97.6% of the sample having a high dominant probability of at least 0.70.

Mean levels for the three school context indicators (**Table 2**) were all significantly different from zero except for school SES in Class 3 ($p = 0.69$). Schools in Class 1 (named EquippedSch-AveSES) had the highest mean levels of science resources and quality of science teachers and average level of mean school SES. Schools in Class 2 (named NeedySch-LowSES) had the lowest mean levels of science resources, quality of science teachers, and school SES. Schools in Class 3 (named AveSch-HighSES) had average mean levels of science resources and quality of science teachers but the highest mean level of school SES.

Comparison of Principal Leadership Among Latent Classes

ANOVA showed that there were overall differences in mean levels of the four school leadership variables among the three latent classes (F for Envisioning = 40.09, $p < 0.01$; F for Instructional-Mgmt = 31.05, $p < 0.01$; Professional-Devt = 3.07, $p < 0.05$; Empowerment = 20.28, $p < 0.01$; **Table 3**). Tamhane *post hoc* tests indicated that compared to schools in the other two latent classes, schools with the highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES) had the highest levels in all four school leadership variables except for Professional-Devt where there were no differences between EquippedSch-AveSES and schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES) (Mean difference or MD = 0.01, $p = 0.99$). In addition to similar mean levels of Professional-Devt between NeedySch-LowSES and EquippedSch-AveSES, there were no significant differences in mean levels of the four leadership variables between NeedySch-LowSES and schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES) (MD for Envisioning = −0.01, $p = 0.96$; MD for Instructional-Mgmt = 0.01, $p = 1.00$; MD for Profdev = 0.06, $p = 0.21$; MD for Empowerment = 0.02, $p = 0.96$).

Comparison of Student Learning Among Latent Classes

Schools with highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES), schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES), and schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-highSES) had the highest, lowest, and average mean levels of students' science achievement and enjoyment in learning science respectively (**Table 4**). ANOVA and Tamhane *post hoc* tests showed that differences in mean levels of these student outcomes among the latent classes and all pairwise comparisons were significant at the 0.01 level.

Relationships Between School Leadership and Students' Science Learning

Table 5 summarizes HLM results for relationships between school leadership and the science achievement of students from

TABLE 1 | LPA model fit indicators.

No. of latent classes (n)	Information criteria			Entropy	Percentage decrease in information criteria for n classes as compared to (n-1) classes		
	AIC	BIC	Sample size- adjusted BIC		AIC	BIC	Sample size- adjusted BIC
(1)	82,645.75	82,688.63	82,669.56	-			
(2)	76,999.03	77,070.48	77,038.70	0.92	6.83	6.79	6.81
(3)	73,137.87	73,237.91	73,193.42	0.95	5.01	4.97	4.99
(4)	72,374.58	72,503.20	72,446.00	0.92	1.04	1.00	1.02
(5)	71,953.09	72,110.28	72,040.37	0.89	0.58	0.54	0.56

TABLE 2 | Descriptives for latent classes.

	M (*SE*) for latent classes		
	Schools with highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES)	Schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES)	Schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES)
Availability of science resources	0.62**(0.00)	0.47**(0.01)	0.56**(0.01)
Quality of science teachers	4.83**(0.01)	0.39**(0.02)	2.65**(0.03)
School SES	−0.06* (0.03)	−0.49**(0.05)	0.02 (0.04)

*$p < 0.05$, **$p < 0.01$.

TABLE 3 | Principal leadership for different latent classes.

	Envisioning	Instructional-Mgmt	Professional-Devt	Empowerment
	M (*SE*)			
Schools with highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES)	3.47 (0.01)	3.85 (0.01)	4.47 (0.01)	3.95 (0.01)
Schools with lowest levels of resources, least qualifed teachers, and lowest SES levels (NeedySch-LowSES)	3.29 (0.02)	3.67 (0.03)	4.46 (0.02)	3.82 (0.02)
Schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES)	3.30 (0.02)	3.66 (0.03)	4.40 (0.02)	3.81 (0.02)
	Comparison of mean levels			
ANOVA (F statistics)	40.09**	31.05**	3.07*	20.28**
	Tamhane *post hoc* tests (Mean differences)			
EquippedSch-AveSES vs. NeedySch-LowSES	0.18**(0.02)	0.18**(0.03)	0.01 (0.03)	0.13**(0.03)
EquippedSch-AveSES vs. AveSch-HighSES	0.16**(0.02)	0.19**(0.03)	0.07*(0.03)	0.14**(0.03)
NeedySch-LowSES vs. AveSch-HighSES	−0.01 (0.03)	0.01 (0.04)	0.06 (0.03)	0.02 (0.03)

*$p < 0.05$ and **$p < 0.01$.

schools with the highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES). Results showed that after controlling for students' gender (Male) and SES (StuSES) and CHC, Professional-Devt ($\beta = -7.70$, $p < 0.01$) and Empowerment ($\beta = -3.63$, $p < 0.05$) were negatively associated with students' science achievement. In contrast, Envisioning ($\beta = -3.67$, $p = 0.07$) and Instructional-Mgmt ($\beta = -2.39$, $p = 0.10$) were not related to students' science achievement.

Table 6 summarizes HLM results for relationships between school leadership and students' enjoyment of science in schools with the highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES). After controlling

TABLE 4 | Students' science achievement and enjoyment in learning science for different latent classes.

	PV1	PV2	PV3	PV4	PV5	PV6	PV7	PV8	PV9	PV10	Enjoy
						***M* (*SE*)**					
Schools with the highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES)	498.96 (0.25)	499.07 (0.25)	498.95 (0.25)	498.81 (0.25)	499.00 (0.25)	499.04 (0.25)	498.92 (0.25)	498.80 (0.25)	498.69 (0.25)	498.98 (0.25)	2.65 (0.00)
Schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES)	479.96 (0.46)	479.62 (0.46)	479.62 (0.46)	479.62 (0.46)	479.72 (0.46)	479.91 (0.46)	479.89 (0.46)	470.60 (0.46)	479.39 (0.46)	479.87 (0.46)	2.57 (0.00)
Schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES)	495.66 (0.44)	495.64 (0.44)	495.69 (0.44)	495.66 (0.44)	495.49 (0.44)	495.65 (0.44)	495.68 (0.44)	495.45 (0.44)	495.64 (0.44)	495.59 (0.44)	2.60 (0.00)
						Comparison of mean levels					
ANOVA (F statistics)	658.27**	687.51**	679.31**	670.85**	675.37**	664.88**	660.12**	670.33**	678.45**	664.51**	179.83**
						Tamhane *post hoc* tests (Mean differences)					
EquippedSch-AveSES vs. NeedySch-LowSES	19.00** (0.52)	19.45** (0.52)	19.32** (0.52)	19.19** (0.52)	19.28** (0.52)	19.13** (0.52)	19.03** (0.52)	19.20** (0.52)	19.29** (0.52)	19.11** (0.52)	0.08** (0.00)
EquippedSch-AveSES vs. AveSch-HighSES	3.30** (0.50)	3.43** (0.50)	3.26** (0.50)	3.15** (0.50)	3.51** (0.50)	3.39** (0.50)	3.24** (0.50)	3.35** (0.50)	3.04** (0.50)	3.39** (0.50)	0.05** (0.00)
NeedySch-LowSES vs. AveSch-HighSES	−15.71** (0.63)	−16.02** (0.63)	−16.07** (0.63)	−16.04** (0.63)	−15.77** (0.63)	−15.74** (0.63)	−15.79** (0.63)	−15.85** (0.63)	−16.25** (0.63)	−15.73** (0.63)	−0.03** (0.01)

**$p < 0.01$.

TABLE 5 | School leadership and students' science achievement for schools with highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES).

Parameter	Model 1	Model 2	Model 3A	Model 3B	Model 3C	Model 3D
Fixed effects						
Intercept	482.75**(6.25)	484.48**(5.16)	483.85**(5.14)	483.97**(5.08)	483.88**(4.96)	483.96**(5.13)
Student variables						
Male		6.00**(1.60)	5.99**(1.60)	6.00**(1.60)	6.00**(1.61)	6.00**(1.60)
StuSES		20.79**(1.31)	20.79**(1.31)	20.80**(1.30)	20.78**(1.30)	20.78**(1.31)
School variables						
Envisioning			−3.67 (2.05)			
Instructional-Mgmt				−2.39 (1.45)		
Professional-Devt					−7.70**(1.50)	
Empowerment						−3.63*(1.68)
Country variable						
CHC		39.76**(12.62)	38.33**(11.58)	38.24**(11.66)	37.99**(11.71)	39.65**(11.42)
Intercepts						
Level 1	6,396.98	6,151.70	6,151.66	6,151.61	6,151.43	6,151.58
Level 2	3,087.84**	2,364.00**	2,356.11**	2,359.34**	2,306.45**	2,353.45**
Level 3	1,285.47**	814.31**	789.20**	799.49**	750.87**	803.72**
% variance						
Level 1	59.39	65.93	66.17	66.07	66.80	66.08
Level 2	28.67	25.34	25.34	25.34	25.05	25.28
Level 3	11.94	8.73	8.49	8.59	8.15	8.63

*Standard errors in parentheses. *$p < 0.05$; **$p < 0.01$.*

TABLE 6 | School leadership and students' enjoyment of science for schools with highest levels of resources and most qualified teachers but average SES levels (EquippedSch-AveSES).

Parameter	Model 4	Model 5	Model 6A	Model 6B	Model 6C	Model 6D
Fixed effects						
Intercept	2.61**(0.02)	2.62**(0.02)	2.62**(0.02)	2.62**(0.02)	2.62**(0.02)	2.62**(0.02)
Student variables						
Male		0.08**(0.01)	0.08**(0.01)	0.08**(0.01)	0.08**(0.01)	0.08**(0.01)
StuSES		0.10**(0.01)	0.10**(0.01)	0.10**(0.01)	0.10**(0.01)	0.10**(0.01)
School variables						
Envisioning			−0.00 (0.01)			
Instructional-Mgmt				−0.00 (0.00)		
Professional-Devt					−0.01 (0.00)	
Empowerment						−0.00 (0.01)
Country variable						
CHC		−0.19**(0.05)	−0.19**(0.05)	−0.19**(0.05)	−0.19**(0.05)	−0.19**(0.05)
Intercepts						
Level 1	0.60	0.59	0.59	0.59	0.59	0.59
Level 2	0.04**	0.03**	0.03**	0.03**	0.03**	0.03**
Level 3	0.02**	0.02**	0.02**	0.02**	0.02**	0.02**
% variance						
Level 1	90.91	92.19	92.19	92.19	92.19	92.19
Level 2	6.06	4.69	4.69	4.69	4.69	4.69
Level 3	3.03	3.13	3.13	3.13	3.13	3.13

*Standard errors in parentheses. **$p < 0.01$.*

for students' gender (Male) and SES (StuSES) and CHC, none of the school leadership variables was significantly related to enjoyment at the 0.05 level (Envisioning, $\beta = -0.00$, $p = 0.91$; Instructional-Mgmt, $\beta = -0.00$, $p = 0.58$; Professional-Devt, $\beta = -0.01$, $p = 0.13$; Empowerment, $\beta = -0.00$, $p = 0.76$).

Next, **Table 7** summarizes HLM results for relationships between school leadership and the science achievement of students from schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES). Results showed that after controlling for students'

TABLE 7 | School leadership and students' science achievement for schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES).

Parameter	Model 1	Model 2	Model 3A	Model 3B	Model 3C	Model 3D
Fixed effects						
Intercept	464.49**(6.58)	465.68**(5.88)	465.49**(5.81)	465.61**(5.93)	465.82**(5.66)	465.44**(5.89)
Student variables						
Male		7.08**(1.94)	7.08**(1.93)	7.08**(1.93)	7.09**(1.94)	7.10**(1.93)
StuSES		15.29**(1.54)	15.29**(1.54)	15.29**(1.54)	15.27**(1.54)	15.28**(1.54)
School variables						
Envisioning			−3.27 (2.97)			
Instructional-Mgmt				−0.46 (1.91)		
Professional-Devt					−6.75**(2.01)	
Empowerment						−3.35 (2.22)
Country variable						
CHC		33.40*(12.15)	31.98**(11.56)	33.13*(12.04)	31.12**(10.78)	33.11**(11.29)
Intercepts						
Level 1	5,721.90	5,581.76	5,581.87	5,581.77	5,581.61	5,581.73
Level 2	3,167.71**	2,698.37**	2,692.61**	2,697.97**	2,659.10**	2,689.06**
Level 3	1,253.74**	915.22**	877.44**	912.23**	848.20**	894.34**
% variance						
Level 1	56.41	60.70	60.99	60.72	61.41	60.90
Level 2	31.23	29.34	29.42	29.35	29.26	29.34
Level 3	12.36	9.95	9.59	9.92	9.33	9.76

*Standard errors in parentheses. *p < 0.05; **p < 0.01.*

TABLE 8 | School leadership and students' enjoyment of science for schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES).

Parameter	Model 4	Model 5	Model 6A	Model 6B	Model 6C	Model 6D
Fixed effects						
Intercept	2.54**(0.02)	2.55**(0.02)	2.55**(0.02)	2.56**(0.02)	2.55**(0.02)	2.55**(0.03)
Student variables						
Male		0.09**(0.02)	0.09**(0.02)	0.09**(0.02)	0.09**(0.02)	0.09**(0.02)
StuSES		0.08**(0.01)	0.08**(0.01)	0.09**(0.01)	0.08**(0.01)	0.08**(0.01)
School variables						
Envisioning			0.02 (0.02)			
Instructional-Mgmt				0.03*(0.01)		
Professional-Devt					0.00 (0.01)	
Empowerment						0.00(0.01)
Country variable						
CHC		−0.19**(0.05)	−0.18**(0.05)	−0.18**(0.05)	−0.19**(0.05)	−0.19**(0.05)
Intercepts						
Level 1	0.59	0.59	0.59	0.59	0.59	0.59
Level 2	0.04**	0.04**	0.04**	0.04**	0.04**	0.04**
Level 3	0.02**	0.02**	0.02**	0.02**	0.02**	0.02**
% variance						
Level 1	90.77	90.77	90.77	90.77	90.77	90.77
Level 2	6.15	6.15	6.15	6.15	6.15	6.15
Level 3	3.08	3.08	3.08	3.08	3.08	3.08

*Standard errors in parentheses. *p < 0.05; **p < 0.01.*

gender (Male) and SES (StuSES) and CHC, Professional-Devt ($\beta = -6.75$, $p < 0.01$) was negatively associated with students' science achievement. The other three school leadership variables were not related to students' science achievement (Envisioning, $\beta = -3.27$, $p = 0.27$; Instructional-Mgmt, $\beta = -0.46$, $p = 0.81$; Empowerment, $\beta = -3.52$, $p = 0.13$).

Table 8 summarizes HLM results for relationship between school leadership and students' enjoyment of science in schools with lowest levels of resources, least qualified teachers, and lowest

SES levels (NeedySch-LowSES). After controlling for students' gender (Male) and SES (StuSES) and CHC, only Instructional-Mgmt was significantly related to enjoyment at the 0.05 level ($\beta = 0.03$, $p < 0.05$) whereas the other three school leadership variables were not (Envisioning, $\beta = 0.02$, $p = 0.16$; Professional-Devt, $\beta = 0.00$, $p = 0.73$; Empowerment, $\beta = 0.00$, $p = 0.87$).

Moving on, **Table 9** summarizes HLM results for relationships between school leadership and the science achievement of students from schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES). Results showed that after controlling for students' gender (Male) and SES (StuSES) and CHC, only Professional-Devt ($\beta = -6.58$, $p < 0.01$) was negatively associated with students' science achievement whereas Envisioning ($\beta = -4.17$, $p = 0.07$), Instructional-Mgmt ($\beta = -2.27$, $p = 0.16$), and Empowerment ($\beta = -2.62$, $p = 0.20$) were not.

Lastly, **Table 10** summarizes HLM results for relationships between school leadership and students' enjoyment of science in schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES). After controlling for students' gender (Male) and SES (StuSES) and CHC, only Instructional-Mgmt ($\beta = 0.02$, $p < 0.05$) was significantly related to students' enjoyment. In contrast, the other three school leadership variables (Envisioning, $\beta = 0.01$, $p = 0.19$; Professional-Devt, $\beta = -0.00$, $p = 0.60$; Empowerment, $\beta = 0.01$, $p = 0.40$) were not associated with students' enjoyment.

DISCUSSION

School SES and Different Types of School Resources

Results from the present study showed that there were three types of schools differing in their contexts as measured by the adequacy of science resources, proportion of qualified science teachers, and school SES. Interestingly, schools with the highest SES were not those with the highest level of qualified science teachers and science resources. This finding may arise because schools vary in the specific types of resources that they have, and high-SES parents may send their children to attend schools with the resources that they value. For example, in Japan, some royal families and very renowned political leaders may aspire their children to study in Gakushûin in order to be socially connected to the most powerful elite in the Japanese society. These schools are endowed in its socio-political capital and general resources but not necessarily science resources. In the same vein, parents in the UK who are alumni of Eton College may enroll their children in their alma mater in view of the prospect for economic connections instead of other considerations such as the adequacy of science resources. Therefore, the present study does not categorically "refute" the association between school SES and resources reported in some studies (Willms, 2010; Liu et al., 2015) but instead point to possible nuances in the specific types of resources that schools have. Future studies can ascertain the different types of resources that characterize high-SES schools.

Inextricable Relationships Between School Contexts and Leadership

Results from the present study also showed that levels of the four principal leadership variables varied with the three types of schools. These results are consistent with the refrain in the school leadership scholarship regarding the need to examine leadership effects in the school context that the leadership is enacted (Hallinger, 2018). However, we do not have a clear understanding of how school contexts and leadership are related to each other. Contexts may shape leadership (as is assumed in this study), so school leaders have to adapt their practice to the school environment that they are operating in Wasserman et al. (2010). Alternatively, leadership may shape contexts, so school leaders have the agency to develop their "ideal" school environment to support their school improvement plans (Hendriks and Scheerens, 2013). Lastly, it can be the case that contexts and leadership may exert a mutual influence on each other. Obviously, the three scenarios carry different implications for school leadership, so future research can clarify the causal relationship between school contexts and leadership.

Envisioning, Instructional Management, and Empowerment in More-Endowed School Contexts

In terms of specificity, levels of principals' envisioning, instructional management, and empowerment were the highest in well-endowed schools (i.e., highest proportion of qualified science teachers and adequate science resources) as compared to the other two types of schools. Principals of schools staffed by qualified teachers may be more likely to focus on setting shared goals (i.e., envisioning) because these teachers have greater capacity to achieve these goals (Notman and Henry, 2011). Principals of well-resourced schools may also be more involved in envisioning because there are adequate resources for realizing school academic goals. The importance of school resources is highlighted by Murphy and Torre (2015) who argued for the alignment between school visions, improvement, and organization including budgets (for resource allocation), operating procedures, structures, and policies.

Next, principals leading schools with more qualified teachers are more likely to focus on instructional management since these teachers are equipped to implement instructional initiatives that promote student-centered pedagogies. Principals leading well-resourced schools may also be more likely to focus on instructional management given the availability of resources to support the implementation of innovative pedagogies (Cohen et al., 2003). The importance of school resources can be inferred from Chang et al. (2008) study of Taiwanese elementary schools which reported that the successful implementation of school plans for technology-enabled instruction required adequate budgets, technological, and other resources. In the case of science education, updated science teaching resources are especially crucial for teachers to deliver effective student-centered lessons.

Principals leading schools with qualified teachers may be more likely to empower teachers to leverage the professional knowledge

TABLE 9 | School leadership and students' science achievement for schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES).

Parameter	Model 1	Model 2	Model 3A	Model 3B	Model 3C	Model 3D
Fixed effects						
Intercept	476.10**(6.52)	478.55**(5.52)	478.63**(5.38)	478.56**(5.45)	478.90**(5.56)	478.66**(5.43)
Student variables						
Male		6.83**(2.19)	6.83**(2.19)	6.83**(2.20)	6.85**(2.20)	6.84**(2.20)
StuSES		17.02**(2.07)	17.02**(2.07)	17.02**(2.07)	16.99**(2.07)	17.02**(2.07)
School variables						
Envisioning			−4.17 (2.30)			
Instructional-Mgmt				−2.27 (1.61)		
Professional-Devt					−6.58**(1.63)	
Empowerment						−2.62 (2.03)
Country variable						
CHC		40.02**(8.47)	37.11**(7.31)	37.61**(7.42)	36.12**(8.87)	39.29**(7.86)
Intercepts						
Level 1	6,138.74	5,982.62	5,982.49	5,982.61	5,982.53	5,982.47
Level 2	3,118.62**	2,601.10**	2,594.13**	2,596.92**	2,559.92**	2,596.78**
Level 3	1,151.92**	781.38**	750.59**	770.09**	768.80**	775.30**
% variance						
Level 1	58.97	63.88	64.14	63.99	64.25	63.95
Level 2	29.96	27.77	27.81	27.78	27.49	27.76
Level 3	11.07	8.34	8.05	8.24	8.26	8.29

*Standard errors in parentheses. **$p < 0.01$.*

TABLE 10 | School leadership and students' enjoyment of science for schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES).

Parameter	Model 4	Model 5	Model 6A	Model 6B	Model 6C	Model 6D
Fixed effects						
Intercept	2.60**(0.03)	2.61**(0.02)	2.61**(0.02)	2.61**(0.02)	2.61**(0.02)	2.61**(0.02)
Student variables						
Male		0.10**(0.03)	0.10**(0.03)	0.10**(0.03)	0.10**(0.03)	0.10**(0.03)
StuSES		0.09**(0.01)	0.09**(0.01)	0.09**(0.01)	0.09**(0.01)	0.09**(0.01)
School variables						
Envisioning			0.01 (0.01)			
Instructional-Mgmt				0.02*(0.01)		
Professional-Devt					−0.00 (0.01)	
Empowerment						0.01 (0.01)
Country variable						
CHC		−0.15**(0.05)	−0.14**(0.04)	−0.13**(0.04)	−0.15**(0.05)	−0.15**(0.05)
Intercepts						
Level 1	0.60	0.59	0.59	0.59	0.59	0.59
Level 2	0.04**	0.04**	0.04**	0.04**	0.04**	0.04**
Level 3	0.02**	0.02**	0.02**	0.02**	0.02**	0.02**
% variance						
Level 1	90.91	90.77	90.77	90.77	90.77	90.77
Level 2	6.06	6.15	6.15	6.15	6.15	6.15
Level 3	3.03	3.08	3.08	3.08	3.08	3.08

*Standard errors in parentheses. *$p < 0.05$; **$p < 0.01$.*

of these teachers to improve the school. This relationship is evident in Lochmiller and Acker-Hocevar's (2016) qualitative study of US public high school administrators which found that principals when confronted with a lack of content knowledge in specialized subject areas such as mathematics and science resorted to hiring teachers who could teach effectively and work collaboratively, allocating resources to support teacher collaboration, and providing professional development.

Importance of Promoting Teachers' Professional Development Across Different School Contexts

The finding for the leadership practice of principals promoting teachers' professional development was more nuanced — principals were more involved in promoting teachers' professional development in well-endowed schools (i.e., schools with qualified teachers and adequate resources) than in schools with average levels of these resources but there was no difference between the most and least endowed schools. Compared to the pattern of results for the other leadership practices (envisioning, instructional management, empowerment) in different types of school contexts, these results suggest that school leaders are more likely to focus on teachers' professional development regardless of their levels of school resources and teacher quality. This leadership imperative reflects the difficulty of hiring new teachers as compared to training existing ones (Hitt and Tucker, 2016) and hence the need for principals to make the best use of available teacher capacity in the school. For example, Lai's (2014) qualitative research in Hong Kong found that principals promoted teachers' professional development by sending teachers to attend external courses when there were teacher resource and institutional constraints. Professional development is particularly helpful if teachers have specific developmental needs to be addressed. For example, principals can also help struggling teachers by providing professional development programs, guidance in classroom management, and organizational, financial, and human support (Yariv and Kass, 2019). Indeed, teachers who undergo professional development get to become more cohesive, professional, competent, and efficacious (Hendriks and Scheerens, 2013). These teachers can contribute to the school academic and improvement capacity.

Principal Leadership and Students' Holistic Learning Outcomes

The present study examined academic and non-academic student learning outcomes in science, namely science achievement and enjoyment in learning science. This more comprehensive conception of student learning, beyond academic achievement alone, is in line with the aims of high-performing education systems worldwide to equip students with knowledge, competencies, and skills that are fit for purpose in the 21st century knowledge-based economies. Notwithstanding the salience of holistic learning outcomes, there are few studies examining the contribution of principal leadership to different students' learning attitudes. Some leadership researchers only focused on specific students' variables [e.g., self-efficacy in Zheng et al. (2017); student engagement in Leithwood and Jantzi (2000); self-concepts, participation, and engagement in Silins and Mulford (2002)]. For example, Zheng et al. (2017) reported that, compared to other leadership factors pertaining to visibility and direct participation, organization of school environment, planning and personnel, and external relations, principals' role in developing teaching-learning most highly predicted grade 8 students' reading achievement and self-efficacy in China. However, Kruger et al. (2007) failed to find a relationship between principal leadership and student commitment as measured by students' perceptions of their relationships with teachers, of the school organization, and of the school culture. The present study therefore addresses the lacuna in our knowledge base on whether principal leadership practices can improve students' learning attitudes and achievement in the area of science. Additionally, results from the present study showing that only instructional management exercised in schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES) and schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES) provide nuanced insights on the types of specific distal antecedents (principal leadership and school contexts) that may influence students' control and value appraisals and consequently their enjoyment of science learning. These contextual insights complement the set of psychological variables in the control-value theory that researchers are increasingly using to explain student experiences of emotions in their learning (Pekrun et al., 2007; Mercan, 2020).

Instructional Management for Promoting Educational Equity

The present study showed that among the four leadership practices, only instructional management was positively related to students' enjoyment of science learning in schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES) and schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES) but not in schools with highest levels of resources and most qualified teachers but average levels of SES (EquippedSch-AveSES); none of the leadership practices was significantly related to students' science achievement. Instructional management can contribute to students' enjoyment of science learning when teachers' instructional practices are informed by the latest research focusing on student-centered learning, when principals emphasize the importance of teachers developing students' critical and creative thinking capacities (beyond textbook knowledge), and when principals recognize teachers' efforts to provide effective student-centered pedagogies in their teaching. These aspects of instructional management are encapsulated in the leadership items used to measure principals' instructional management practices in the present study.

The finding that principals' instructional management was only positively related to students' science enjoyment in less-endowed (schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES) and schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES) is consistent with that reported in an evolving body of literature. For example, Tan's (2018a) analysis of PISA 2012 data found that principal instructional leadership was most strongly associated with the mathematics achievement of students who attended the least-resourced schools in OECD countries; these students were also from the lowest SES families, had the lowest prior achievement level, had parents with the lowest academic expectations of schools.

There are many reasons why principals' instructional management may contribute more to student learning in less-endowed schools. First, principals in these schools may exercise greater instructional control and are more focused on teaching-and-learning than building relationships (Hallinger and Murphy, 1986). Some studies indicate that principals also leveraged on collaborative instructional leadership focusing on teaching-and-learning (Hallinger and Heck, 2011). Second, principals may facilitate more instructional discussions among teachers, protect teachers from classroom disruptions, leverage test results more frequently to improve instructional programs, establish more systematic monitoring of student progress, and communicate instructional goals to teachers more effectively (Heck, 1992). Lastly, from a school improvement perspective, Day et al.' (2016) mixed-method, longitudinal study of effective and improving English schools underscored the need for principals to have high levels of expectations in classroom teaching, emphasize student behavior and achievement, conduct more classroom observations, and coach less effective teachers.

The finding that principals' instructional management was positively related to student learning in less-endowed schools contributes to the policy discourse on school equity. There is an expectation that effective school systems achieve high levels of student performance (educational excellence) for different groups of students (educational equity) (Schleicher, 2009). In the context of the present study, greater equity means closing the learning gap between students in advantaged and disadvantaged school contexts. Therefore, principals in less-endowed schools can focus more on instructional management (vis-à-vis other leadership practices) to improve student learning even in less-endowed school contexts (Dimmock and Tan, 2016).

Negative Relationships Between Enhancing Teacher Capacity and Student Learning

The present study examines two principal leadership practices that are aimed at enhancing teacher capacity, namely promoting professional development and teacher empowerment. Results showed that the two leadership practices were not positively related to student learning. Instead, results from the present study showed that in all three school contexts, students whose principals promoted professional development for teachers had lower levels of science achievement.

These results may arise because principals have limited time and energies to manage myriad school needs (Goldring et al., 2008; May et al., 2012), so if they focus on teachers' professional development they will have less capacity to spearhead instructional initiatives which may have a more direct impact on student learning. More fundamentally, it is important to ascertain what drives higher levels of teachers' professional development in the first place. If schools suffer from a deficit in teacher quality and principals attempt to address this capacity issue through professional development, then it takes time for effects of enhanced teacher capacity via professional development to manifest in student learning. Indeed, teachers need to change their pedagogical beliefs and practices simultaneously to effect changes in students' learning outcomes (Clarke and Hollingsworth, 2002). Therefore, if teacher capacity constraints are severe and professional development fails to change complex systems of influences simultaneously, then student learning may not improve (Opfer and Pedder, 2011). Another possible reason to explain the negative relationships between principals' promotion of teachers' professional development and student learning is that some teachers may perceive greater professional development as undue influences to shape their professional practice. This argument reinforces Opfer and Pedder's (2011) thesis to appreciate the complex interplay among school factors, teacher factors, and the learning activity that collectively impact the effectiveness of teachers' professional development. If this is so, teachers who are asked to attend professional development may perceive an erosion of professional autonomy and be less motivated (Hallinger and Lu, 2014). The decreased motivation may impact the quality of teaching adversely.

As for teacher empowerment, results showed that this leadership practice, just as in the case for teachers' professional development, was also not positively related to student achievement. Specifically, teacher empowerment was negatively related to student achievement in EquippedSch-AveSES schools and not related to student achievement in the other two types of schools. These results may happen because teachers who are expected to contribute to organizational improvement may not be able to commit their energies and resources to improving teaching-and-learning. The tension between leadership and teaching responsibilities is evident in Brooks et al. (2004) study where teacher leaders perceived their leadership responsibilities as "a source of frustration that pried them from the essential, instructional tasks of their profession" (p. 253). As a result, students may not benefit directly from increased teacher empowerment.

Contributions, Limitations, and Future Research

The present study elucidates the different types of contexts that schools operate in and clarifies how some leadership practices differentially impact students' science learning depending on these school contexts. Data from 248,620 students and 9,370 school principals in 35 OECD countries who participated in PISA 2015 were analyzed using LPA, ANOVA and Tamhane post-hoc comparisons, and three-level HLM. The study contributes to theory and practice in three ways.

First, it is one of the few studies to provide empirical evidence that schools do not operate in homogeneous contexts by clarifying how these different school contexts look like in terms of the availability of science resources, quality of science teachers, and school SES. Among the three types of school contexts identified in the LPA, schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES) represents the most challenging contexts that 19.02% of the schools in the sample operate in. These schools are confronted with having less science resources, less qualified science teachers, and students from lower-SES families who are likely to receive less parental support for their

learning. The study makes a second contribution by identifying leadership practices associated with specific types of school contexts. For example, results showed principals were more likely to have envisioning, instructional management, and teacher empowerment in schools that had the most science resources and best science teacher quality (EquippedSch-AveSES). Future studies can unravel what principals focus on in their leadership in less-endowed schools and whether these leadership priorities contribute to student learning. The study makes a third contribution by identifying principals' instructional management as being more effective than promoting professional development or empowering teachers for students' science learning in schools with less science resources and lower science teacher quality [schools with lowest levels of resources, least qualified teachers, and lowest SES levels (NeedySch-LowSES), schools with average levels of resources and moderately qualified teachers but highest SES levels (AveSch-HighSES)]. Instructional management thus seems to have a compensatory effect on students' learning in less-endowed schools. However, professional development and empowerment are means to addressing teachers' competence and autonomy needs (Eyal and Roth, 2011; Shepherd-Jones and Salisbury-Glennon, 2018) and thereby, building teacher capacity which will in the long term also benefit student learning. How then should principals strike a balance between focusing on instructional management and building teacher capacity? Future research can examine how principals negotiate these different leadership priorities.

As with all empirical studies, results from the present study should be read with some limitations in mind. First, the PISA sample comprised only 15-year-old students who were mostly in Grade 10 (55.9%) with the rest were from Grades 7−13, so results reported are applicable only to this student population. Second, it examined only four core leadership practices in Leithwood et al. (2006) conceptualization, so future studies can examine other leadership practices. Third, the study relied on principals' self-reported data for their leadership practices, so future studies can complement these with teacher-reported data to reduce bias (Urick, 2016). Fourth, the focus on students' learning in science instead of other subject areas assumes that schools generally value science learning but there are schools which may value learning in other domains (e.g., aesthetics in Waldorf School) as much as, if not more than, in science. Therefore, results from the present study have to be interpreted with this caveat in mind. Lastly, the data analyzed were correlational in nature, so causal inferences should be made with caution. Causal, or at least longitudinal, research designs in future research can be used to ascertain the relationships reported.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.oecd.org/pisa/data/2015database/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

CT conceptualized the study and wrote 60% of the article (including Sections "Introduction, Materials and Methods, Results, and Conclusion"). PL contributed to the literature review. WW contributed to the discussion of the article. All authors contributed to the article and approved the submitted version.

FUNDING

The study was supported by a grant from The University of Hong Kong Faculty Research Fund. Funds for the open access publication fees are received from The University of Hong Kong.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.02257/full#supplementary-material

REFERENCES

Acosta, S., and Hsu, H. Y. (2014). Shared academic values: testing a model of the association between Hong Kong parents' and adolescents' perception of the general value of science and scientific literacy. *Educ. Stud.* 40, 174–195. doi: 10.1080/03055698.2013.866889

Adams, C., and Olsen, J. (2017). Principal support for student psychological needs: a social- psychological pathway to a healthy learning environment. *J. Educ. Admin.* 55, 510–525. doi: 10.1108/JEA-05-2016-0045

Adams, C. M., and Olsen, J. J. (2019). Principal support of student psychological needs and a functional instructional core. *J. Educ. Admin.* 57, 243–260. doi: 10.1108/JEA-04-2018-0076

Ainley, M., and Hidi, S. (2014). "Interest and enjoyment," in *International Handbook of Emotions in Education* eds R. Pekrun and L. Linnebrink-Garcia (New York, NY: Routledge) 205–227.

Archer, L., Dawson, E., DeWitt, J., Seakins, A., and Wong, B. (2015). "Science capital": a conceptual, methodological, and empirical argument for extending Bourdieusian notions of capital beyond the arts. *J. Res. Sci. Teach.* 52, 922–948. doi: 10.1002/tea.21227

Bottery, M., Ngai, G., Wong, P. M., and Wong, P. H. (2008). Leaders and contexts: comparing English and Hong Kong perceptions of educational challenges. *ISEA* 36, 56–71.

Brooks, J. S., Scribner, J. P., and Eferakorho, J. (2004). Teacher leadership in the context of whole school reform. *J. Sch. Leaders.* 14, 242–265. doi: 10.1177/105268460401400301

Camacho-Morles, J., Slemp, G. R., Oades, L. G., Morrish, L., and Scoular, C. (2019). The role of achievement emotions in the collaborative problem-solving performance of adolescents. *Learn. Indiv. Differ.* 70, 169–181. doi: 10.1016/j.lindif.2019.02.005

Chang, I.-H., Chin, J. M., and Hsu, C.-M. (2008). Teachers' perceptions of the dimensions and implementation of technology leadership of principals in Taiwanese elementary schools. *Educ. Technol. Soc.* 11, 229–245.

Cheema, J. (2014). A review of missing data handling methods in education research. *Rev. Educ. Res.* 84, 487–508. doi: 10.3102/0034654314532697

Clarke, D., and Hollingsworth, H. (2002). Elaborating a model of teacher professional growth. *Teach. Teach. Educ.* 18, 947–967. doi: 10.1016/S0742-051X(02)00053-7

Claussen, S., and Osborne, J. (2013). Bourdieu's notion of cultural capital and its implications for the science curriculum. *Sci. Educ.* 97, 58–79. doi: 10.1002/sce.21040

Close, P., and Raynor, A. (2010). Five literatures of organisation: putting the context back into educational leadership. *Sch. Leaders. Manag.* 30, 209–224. doi: 10.1080/13632434.2010.485182

Cohen, D. K., Raudenbush, S. W., and Ball, D. L. (2003). Resources, instruction, and research. *Educ. Eval. Policy Anal.* 25, 119–142. doi: 10.3102/01623737025002119

Davis-Beggs, K. D. (2013). *The Effects of School Resources on Student achievement (Publication No. 3561096)*. Doctoral dissertation. Harrogate, TN: Lincoln Memorial University.

Day, C., Gu, Q., and Sammons, P. (2016). The impact of leadership on student outcomes: how successful school leaders use transformational and instructional strategies to make a difference. *Educ. Admin. Q.* 52, 221–258. doi: 10.1177/0013161X15616863

Day, C., Sammons, P., Hopkins, D., Harris, A., Leithwood, K., Gu, Q., et al. (2009). *The Impact of School Leadership on Pupil Outcomes: Final Report (Research Report DCSF-RR108)*. Nottingham: Department for Children, Schools and Families.

Dimmock, C., and Tan, C. Y. (2016). Re-conceptualizing learning-centred (instructional) leadership: an obsolete concept in need of renovation. *Lead. Manag.* 22, 1–17.

Eyal, O., and Roth, G. (2011). Principals' leadership and teachers' motivation: self-determination theory analysis. *J. Educ. Admin.* 49, 256–275. doi: 10.1108/09578231111129055

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: the broaden- and-build theory of positive emotions. *Am. Psychol.* 56, 218–226. doi: 10.1037/0003-066X.56.3.218

Ganzeboom, H. B. G., and Treiman, D. J. (2003). "Three internationally standardised measures for comparative research on occupational status," in *Advances in Cross-National Comparison: A European Working Book for Demographic and Socio-Economic Variables*, eds J. H. P. Hoffmeyer-Zlotnik and C. Wolf (New York, NY: Kluwer Academic Press), 159–193. doi: 10.1007/978-1-4419-9186-7_9

Goldring, E., Huff, J., May, H., and Camburn, E. (2008). School context and individual characteristics: what influences principal practice? *J. Educ. Admin.* 46, 332–352. doi: 10.1108/09578230810869275

Greenwald, R., Hedges, L. V., and Laine, R. D. (1996). The effect of school resources on student achievement. *Rev. Educ. Res.* 66, 361–396. doi: 10.3102/00346543066003361

Gurr, D., Drysdale, L., and Mulford, B. (2006). Models of successful principal leadership. *Sch. Leaders. Manag.* 26, 371–395. doi: 10.1080/13632430600886921

Häkkinen, I., Kirjavainen, T., and Uusitalo, R. (2003). School resources and student achievement revisited: new evidence from panel data. *Econ. Educ. Rev.* 22, 329–335. doi: 10.1016/S0272-7757(02)00060-2

Hallinger, P. (2018). Bringing context out of the shadows of leadership. *Educ. Manag. Admin. Leaders.* 46, 5–24. doi: 10.1177/1741143216670652

Hallinger, P., Bickman, L., and Davis, K. (1996). School context, principal leadership, and student reading achievement. *Element. Sch. J.* 96, 527–549. doi: 10.1086/461843

Hallinger, P., and Heck, R. H. (2011). Exploring the journey of school improvement: classifying and analyzing patterns of change in school improvement processes and learning outcomes. *Sch. Effect. Sch. Improve.* 22, 1–27. doi: 10.1080/09243453.2010.536322

Hallinger, P., and Lu, J. (2014). Modelling the effects of principal leadership and school capacity on teacher professional learning in Hong Kong primary schools. *Sch. Leaders. Manag.* 34, 481–501. doi: 10.1080/13632434.2014.938039

Hallinger, P., and Murphy, J. F. (1986). The social context of effective schools. *Am. J. Educ.* 94, 328–355. doi: 10.1086/443853

Hampden-Thompson, G., and Bennett, J. (2013). Science teaching and learning activities and students' engagement in science. *Int. J. Sci. Educ.* 35, 1325–1343. doi: 10.1080/09500693.2011.608093

Hanushek, E. A. (1996). "School resources and student performance," in *Does Money Matter? The Effect of School Resources on Student Achievement and Adult Success* ed G. Burtless (Washington, DC: Brookings Institution Press), 43–73.

Harris, A., Adams, D., Jones, M. S., and Muniandy, V. (2015). System effectiveness and improvement: the importance of theory and context. *Schl. Effect. Schl. Improv.* 26, 1–3. doi: 10.1080/09243453.2014.987980

Heck, R. H. (1992). Principals' instructional leadership and school performance: implications for policy development. *Educ. Eval. Policy Anal.* 14, 21–34. doi: 10.3102/01623737014001021

Hedges, L. V., Laine, R. D., and Greenwald, R. (1994). An exchange: Part I: does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educ. Res.* 23, 5–14. doi: 10.2307/1177220

Hendriks, M. A., and Scheerens, J. (2013). School leadership effects revisited: a review of empirical studies guided by indirect-effect models. *Schl. Leaders. Manag.* 33, 373–394. doi: 10.1080/13632434.2013.813458

Hitt, D. H., and Tucker, P. D. (2016). Systematic review of key leader practices found to influence student achievement: a unified framework. *Rev. Educ. Res.* 86, 531–569. doi: 10.3102/0034654315614911

Hoppey, D., and McLeskey, J. (2013). A case study of principal leadership in an effective inclusive school. *J. Spec. Educ.* 46, 245–256. doi: 10.1177/0022466910390507

Houtenville, A. J., and Conway, K. S. (2008). Parental effort, school resources, and student achievement. *J. Hum. Resour.* 43, 437–453. doi: 10.2307/40057353

Hunzicker, J. (2012). Professional development and job-embedded collaboration: how teachers learn to exercise leadership. *Prof. Dev. Educ.* 38, 267–289. doi: 10.1080/19415257.2012.657870

Ingvarson, L., and Rowley, G. (2017). Quality assurance in teacher education and outcomes: a study of 17 countries. *Educ. Res.* 46, 177–193. doi: 10.3102/0013189X17711900

Irwin, A. (2001). Constructing the scientific citizen: science and democracy in the biosciences. *Publ. Understand. Sci.* 10, 1–18. doi: 10.3109/a036852

Jacobson, S. (2011). Leadership effects on student achievement and sustained school success. *Int. J. Educ. Manag.* 25, 33–44. doi: 10.1108/09513541111100107

Jeffries, D., Curtis, D. D., and Conner, L. N. (2020). Student factors influencing STEM subject choice in year 12: a structural equation model using PISA/LSAY data. *Int. J. Sci. Math. Educ.* 18, 441–461. doi: 10.1007/s10763-019-09972-5

Kruger, M. L., Witziers, B., and Sleegers, P. (2007). The impact of school leadership on school level factors: validation of a causal model. *Schl. Effect. Schl. Improv.* 18, 1–20. doi: 10.1080/09243450600797638

Lai, E. (2014). Principal leadership practices in exploiting situated possibilities to build teacher capacity for change. *Asia Pac. Educ. Rev.* 15, 165–175. doi: 10.1007/s12564-014-9314-0

Lam, T. Y. P., and Lau, K. C. (2014). Examining factors affecting science achievement of Hong Kong in PISA 2006 using hierarchical linear modelling. *Int. J. Sci. Educ.* 36, 2463–2480. doi: 10.1080/09500693.2013.879223

Leithwood, K., and Jantzi, D. (2000). Principal and teacher leadership effects: a replication. *Schl. Leaders. Manag.* 20, 415–434. doi: 10.1080/713696963

Leithwood, K., Jantzi, D., and McElheron-Hopkins, C. (2006). The development and testing of a school improvement model. *Schl. Effect. Schl. Improve.* 17, 441–464. doi: 10.1080/09243450600743533

Lin, H., Lawrenz, F., Lin, S.-F., and Hong, Z.-R. (2012). Relationships among affective factors and preferred engagement in science-related activities. *Publ. Understand. Sci.* 22, 941–954. doi: 10.1177/0963662511429412

Liu, H., van Damme, J., Gielen, S., and van den Noortgate, W. (2015). School processes mediate school compositional effects: model specification and estimation. *Br. Educ. Res. J.* 41, 423–447. doi: 10.1002/berj.3147

Lochmiller, C. R. (2016). Examining administrators' instructional feedback to high school math and science teachers. *Educ. Admin. Q.* 52, 75–109. doi: 10.1177/0013161X15616660

Lochmiller, C. R., and Acker-Hocevar, M. (2016). Making sense of principal leadership in content areas: the case of secondary math and science instruction. *Leaders. Policy Schl.* 15, 273–296. doi: 10.1080/15700763.2015.1073329

May, H., Huff, J., and Goldring, E. (2012). A longitudinal study of principals' activities and student performance. *Schl. Effective. Schl. Improve.* 23, 417–443. doi: 10.1080/09243453.2012.678866

Mercan, F. C. (2020). Control-value theory and enjoyment of science: a cross-national investigation with 15-year-olds using PISA 2006 data. *Learn. Indiv. Differ.* 80:101889. doi: 10.1016/j.lindif.2020.101889

Murillo, F. J., and Hernández-Castilla, R. (2015). Liderazgo para el aprendizaje:? Qué tareas de los directores y directoras escolares son las que más inciden en el aprendizaje de los estudiantes? *Relieve Rev. Electrón. Invest. Eval. Educ.* 21, doi: 10.7203/relieve.21.1.5015

Murphy, J. (2015). Creating communities of professionalism: addressing cultural and structural barriers. *J. Educ. Admin.* 53, 154–176. doi: 10.1108/JEA-10-2013-0119

Murphy, J., and Torre, D. (2015). Vision: essential scaffolding. *Educ. Manag. Admin. Leaders.* 43, 177–197. doi: 10.1177/1741143214523017

Notman, R., and Henry, D. (2011). Building and sustaining successful school leadership in New Zealand. *Leaders. Policy Schl.* 10, 373–394. doi: 10.1080/15700763.2011.610555

Nylund, K. L., Asparouhov, T., and Muthen, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: a Monte Carlo simulation study. *Struct. Equ. Model.* 14, 535–569. doi: 10.1080/10705510701575396

Oberski, D. (2016). "Mixture models: latent profile and latent class analysis," in *Modern Statistical Methods for HCI*, eds J. Robertson and M. Kaptein (Switzerland: Springer), 275–287. doi: 10.1007/978-3-319-26633-6_12

Opfer, V. D., and Pedder, D. (2011). Conceptualizing teacher professional learning. *Rev. Educ. Res.* 81, 376–407. doi: 10.3102/0034654311413609

Organization for Economic Co-operation and Development [OECD] (2004). *Learning for Tomorrow's World: First Results From PISA 2003*. Paris: OECD.

Organization for Economic Co-operation and Development [OECD] (2009). *PISA 2006 Technical Report*. Paris: OECD.

Organization for Economic Co-operation and Development [OECD] (2017). *PISA 2015 Technical Report*. Paris: OECD.

Park, J., and Yu, H.-T. (2018). A comparison of approaches for estimating covariate effects in nonparametric multilevel latent class models. *Struct. Equ. Model. Multidiscipl. J.* 25, 778–790. doi: 10.1080/10705511.2018.1448711

Pekrun, R. (2006). The control-value theory of achievement emotions: assumptions, corollaries, and implications for educational research and practice. *Educ. Psychol. Rev.* 18, 315–341. doi: 10.1007/s10648-006-9029-9

Pekrun, R. (2017). Emotion and achievement during adolescence. *Child Dev. Perspect.* 11, 215–221. doi: 10.1111/cdep.12237

Pekrun, R., Frenzel, A. C., Goetz, T., and Perry, R. P. (2007). "The control-value theory of achievement emotions: an integrative approach to emotions in education," in *Emotion in Education*, eds P. A. Schutz and R. Pekrun (Amsterdam: Academic Press), 13–36. doi: 10.1016/b978-012372545-5/50003-4

Pekrun, R., and Stephens, E. J. (2012). "Academic emotions," in *APA Educational Psychology Handbook, Vol 2: Individual Differences and Cultural and Contextual Factors*, eds K. R. Harris, S. Graham, T. Urdan, S. Graham, J. M. Royer, and M. Zeidner (Washington, DC: American Psychological Association), 3–31.

Perry, L. B., and McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teach. Coll. Rec.* 112, 1137–1162.

Raccanello, D., Hall, R., and Burro, R. (2018). Salience of primary and secondary school students' achievement emotions and perceived antecedents: interviews on literacy and mathematics domains. *Learn. Indiv. Differ.* 65, 65–79. doi: 10.1016/j.lindif.2018.05.015

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edn. Thousand Oaks, CA: Sage.

Reynolds, D. (2010). *Failure-Free Education? The Past, Present and Future of School Effectiveness and School Improvement*. London: Routledge.

Reynolds, D., Sammons, P., de Fraine, B., van Damme, J., Townsend, T., Teddlie, C., et al. (2014). Educational effectiveness research (EER): a state-of-the-art review. *Schl. Effective. Schl. Improve.* 25, 197–230. doi: 10.1080/09243453.2014.885450

Schleicher, A. (2009). Securing quality and equity in education: lessons from PISA. *Prospects* 39, 251–263. doi: 10.1007/s11125-009-9126-x

Shepherd-Jones, A. R., and Salisbury-Glennon, J. D. (2018). Perceptions matter: the correlation between teacher motivation and principal leadership styles. *J. Res. Educ.* 28, 93–131.

Silins, H., and Mulford, B. (2002). Schools as learning organisations: the case for system, teacher and student learning. *J. Educ. Admin.* 40, 425–446. doi: 10.1108/09578230210440285

Stosich, E. L. (2016). Building teacher and school capacity to teach to ambitious standards in high-poverty schools. *Teach. Teach. Educ.* 58, 43–53. doi: 10.1016/j.tate.2016.04.010

Tan, C. Y. (2018a). Examining school leadership effects on student achievement: the role of contextual challenges and constraints. *Cambridge J. Educ.* 48, 21–45. doi: 10.1080/0305764X.2016.1221885

Tan, C. Y. (2018b). Involvement practices, socioeconomic status, and student science achievement: insights from a typology of home and school involvement patterns. *Am. Educ. Res. J.* 56, 899–924. doi: 10.3102/0002831218807146

Urick, A. (2016). The influence of typologies of school leaders on teacher retention: a multilevel latent class analysis. *J. Educ. Admin.* 54, 434–468. doi: 10.1108/JEA-08-2014-0090

Van Ewijk, R., and Sleegers, P. (2010). The effect of peer socioeconomic status on student achievement: a meta-analysis. *Educ. Res. Rev.* 5, 134–150. doi: 10.1016/j.edurev.2010.02.001

Waldron, N., and McLeskey, J. (2010). Establishing a collaborative culture through comprehensive school reform. *J. Educ. Psychol. Consult.* 20, 58–74. doi: 10.1080/10474410903535364

Walker, A., and Ko, J. (2011). Principal leadership in an era of accountability: a perspective from the Hong Kong context. *Schl. Leaders. Manag.* 31, 369–392. doi: 10.1080/13632434.2011.606269

Wasserman, N., Nohria, N., and Anand, B. (2010). "When does leadership matter? A contingent opportunities view of CEO leadership," in *Handbook of Leadership Theory and Practice: An HBS Centennial Colloquium on Advancing Leadership*, eds N. Nohria and R. Khurana (Boston, MA: Harvard Business Press), 27–63.

Wenner, J. A., and Campbell, T. (2017). The theoretical and empirical basis of teacher leadership: a review of the literature. *Rev. Educ. Res.* 87, 134–171. doi: 10.3102/0034654316653478

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teach. Coll. Rec.* 112, 1008–1037.

Woolnough, B. (1994). *Effective Science Teaching*. Buckingham: Open University Press.

Xie, Y., Fang, M., and Shauman, K. (2015). STEM education. *Annu. Rev. Sociol.* 41, 331–357. doi: 10.1146/annurev-soc-071312-145659

Yariv, E., and Kass, E. (2019). Assisting struggling teachers effectively. *Educ. Manag. Admin. Leaders.* 47, 310–325. doi: 10.1177/1741143217725323

Zhang, D., Hsu, H. Y., Kwok, O. M., Benz, M., and Bowman-Perrott, L. (2011). The impact of basic-level parent engagements on student achievement: patterns associated with race/ethnicity and socioeconomic status (SES). *J. Disabil. Policy Stud.* 22, 28–39. doi: 10.1177/1044207310394447

Zheng, Q., Li, L., Chen, H., and Loeb, S. (2017). What aspects of principal leadership are most highly correlated with school outcomes in China? *Educ. Admin. Q.* 53, 409–447. doi: 10.1177/0013161X17706152

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 30 September 2020
doi: 10.3389/fpsyg.2020.579545

When Large-Scale Assessments Meet Data Science: The Big-Fish-Little-Pond Effect in Fourth- and Eighth-Grade Mathematics Across Nations

*Ze Wang**

Department of Educational, School & Counseling Psychology, University of Missouri, Columbia, MO, United States

The programming language of R has useful data science tools that can automate analysis of large-scale educational assessment data such as those available from the United States Department of Education's National Center for Education Statistics (NCES). This study used three R packages: EdSurvey, MplusAutomation, and tidyverse to examine the big-fish-little-pond effect (BFLPE) in 56 countries in fourth grade and 46 countries in eighth grade for the subject of mathematics with data from the Trends in International Mathematics and Science Study (TIMSS) 2015. The BFLPE refers to the phenomenon that students in higher-achieving contexts tend to have lower self-concept than similarly able students in lower-achieving contexts due to social comparison. In this study, it is used as a substantive theory to illustrate the implementation of data science tools to carry out large-scale cross-national analysis. For each country and grade, two statistical models were applied for cross-level measurement invariance testing, and for testing the BFLPE, respectively. The first model was a multilevel confirmatory factor analysis for the measurement of mathematics self-concept using three items. The second model was multilevel latent variable modeling that decomposed the effect of achievement on self-concept into between and within components; the difference between them was the contextual effect of the BFLPE. The BFLPE was found in 51 of the 56 countries in fourth grade and 44 of the 46 countries in eighth grade. The study provides syntax and discusses problems encountered while using the tools for modeling and processing of modeling results.

Keywords: big-fish-little-pond effect, data science, latent variable modeling, large-scale assessment, R, TIMSS

OPEN ACCESS

Edited by:
Ronnel B. King,
University of Macau, China

Reviewed by:
Lu Wang,
Ball State University, United States
Jesús-Nicasio García-Sánchez,
Universidad de León, Spain

***Correspondence:**
Ze Wang
wangze@missouri.edu

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *02 July 2020*
Accepted: *25 August 2020*
Published: *30 September 2020*

Citation:
Wang Z (2020) When Large-Scale Assessments Meet Data Science: The Big-Fish-Little-Pond Effect in Fourth- and Eighth-Grade Mathematics Across Nations. Front. Psychol. 11:579545. doi: 10.3389/fpsyg.2020.579545

INTRODUCTION

Data science tools, particularly those developed with the statistical language of R (R Core Team, 2020), have been increasingly used in educational and social sciences. For scholarly articles, R is the second most frequently used data science software following SPSS (Muenchen, n.d.). Given its integrated system of data wrangling, statistical modeling, visualization, and communication (Grolemund and Wickham, 2018), R is appealing to those conducting empirical analysis (i.e., using

real data) as well as those interested in simulation studies. Currently, there are over 16,000 R packages available on the Comprehensive R Archive Network (CRAN) – R's main repository of packages – and more packages in other repositories (such as GitHub). Packages are developed for various topics (for example, see "Task Views" at the CRAN). They, together with R's core packages, provide tools for researchers to work with different aspects of using data. There are also search engines (e.g., RSeek, Nabble), online communities (e.g., Stack Overflow, Cross Validated, RStudio Community), and mailing lists (e.g., R-help, R-devel) that are available for additional help for using R. At the same time, the sheer amount of R resources seems daunting to beginner users, let alone its sometimes unfamiliar or non-user-friendly ways of "doing" things.

Large-scale assessments (LSAs) are great data sources (Rutkowski et al., 2014). An LSA typically involves complex design frameworks for the development of items, sampling participants, data collection, and variable creation. The United States Department of Education's National Center for Education Statistics (NCES) houses multiple international LSA studies across the lifespan from early childhood to adults (National Center for Education Statistics, n.d.). These studies are sponsored by two organizations: The International Association for the Evaluation of Educational Achievement (IEA) and the Organization for Economic Cooperation and Development (OECD), although the work is typically directed by testing firms and research institutes in cooperation with national research institutions and governmental agencies. In the United States, the National Assessment of Educational Progress (NAEP) is an LSA that was first conducted in 1969 (National Center for Education Statistics, 2020). LSAs allow researchers to use nationally and internationally representative data to answer research questions and even for policymaking (Wagemaker, 2014).

Despite its rich data, LSAs have been used only to the extent that is far from its potential (Wang, 2017). It takes quite some time for one to get familiar with an LSA. Substantive researchers may be unaware of the relevant content in LSAs that can be used for their research. Or, they may lack the expertise to go through the database which may contain hundreds of datasets, or to run large-scale analysis. At the same time, when a researcher does use an LSA, many times only data from a single or a few countries/regions are used for analysis (e.g., Wang et al., 2012; Smith et al., 2020).

In this article, I illustrate how to use a few R packages that I have found particularly useful for conducting large-scale cross-national analysis using NCES data. Those packages are EdSurvey (Bailey et al., 2020), MplusAutomation (Hallquist and Wiley, 2018), and tidyverse (Wickham et al., 2019). Several other packages were used for this study but the main functions are from these three packages.

The goal of this study is twofold. First, it examines and continues to document the big-fish-little-pond effect (BFLPE) using the Trends in International Mathematics and Science Study (TIMSS), an international LSA by IEA. Second, it demonstrates the implementation of data science tools to carry out large-scale cross-national analysis. I provide syntax so that interested readers can replicate the analysis. The syntax can also be modified for similar analyses.

A Few Data Science Tools

Traditional tools tend to treat different aspects of the whole data manipulation and statistical analysis in compartments. Each tool is for a special purpose and the user has to piece all the different elements together to use multiple tools for a more complex problem. To illustrate my point, think about the different statistics courses a doctoral student in educational psychology typically takes. The student may take courses that cover regression, analysis of variance, factor analysis, structural equation modeling, etc. For those courses, the professor may provide data for homework problems and/or projects, or the student may be encouraged to work on projects with their "own" data. In the latter case, the data may be from the student's advisor or another fellow graduate student. Most likely, the data are already cleaned/managed in the sense that the variables are ready to be used to apply the learned statistical techniques. The student may be unappreciative of the data management steps that lead to the cleaned data until they are involved in a bigger grant project or at the dissertation/thesis stage. However, data wrangling is time-consuming. Data scientists spend from 50 to 80% of their time collecting and preparing unruly digital data (Lohr, 2014). With the amount of data available in every field, the toolbox of quantitative researchers, especially those working with empirical data, needs to include tools that allow them to handle various types and quantities of data.

R and RStudio

The programming language R has increasingly become a popular statistical analysis software. It is open source meaning that everyone can access and contribute to its development. Despite its relatively long history (The first publicly available version of R was released in 2000), R has only gained more acceptance among social science researchers in the past decade or so. R was born out of S, which was intended to be a programming language focused on data analysis, and has evolved into a system used not only by computer programmers and data analysts but also by physical scientists, psychologists, journalists, etc. Researchers use R because (a) it is free and open-source; (b) it has many packages built to meet various needs of statistical analysis; (c) there are freely provided useful resources among the R community; (d) collaboration using R is easy; (e) analysis with R can be highly reproducible; and (f) data wrangling using R can be fast, dependable, and highly replicable (Barrett, 2019).

RStudio is an integrated development environment (IDE) for R. It uses R to develop codes and analysis that can be executed and has greater usability than R. Essentially RStudio can be thought of as the interface between the user and R. It depends on and adds onto R, which means that the R program has to be installed before RStudio for RStudio to implement R. Any R package or function can be used in RStudio. RStudio has many features for good usability. One basic feature I particularly like is auto-completion. When the user types the first few characters of an R command, function, or the name of a data object that has been created, RStudio will show a list of complete names

from which the user can choose to insert. This saves a lot of time typing and finding typos. For more experienced users who would like to develop their own packages, RStudio provides tools that automatically organize the structure necessary for package development. Interested readers can check out the "Advanced R" book (Wickham, 2019).

EdSurvey

One particular challenge of using LSAs is to access and browse the data. A researcher may have some idea about the LSA after reading its description online or the user's guide, but getting hands-on with the data usually means downloading big zipped files, unzipping them, and making them viewable using statistical software such as SPSS or SAS. Sometimes, there are hundreds of datasets that can be explored. The R EdSurvey package, recently developed by Bailey et al. (2020), makes accessing and transforming LSAs data to be R-ready a breeze.

EdSurvey was developed for data downloading, processing, manipulation, and analysis of LSAs by NCES and incorporates special survey methodologies (complex sampling, sampling weights, replicate weights, etc.) in a single package. In addition to data procuration, EdSurvey has methods developed for statistical analysis. However, these methods are for analysis of observed variables. Researchers interested in using latent variable modeling techniques such as factor analysis has to rely on other packages. The R package lavaan (Rosseel, 2012) is widely used for latent variable modeling but its capabilities are still limited for analysis using LSAs. For example, sampling weights in lavaan can only be used for non-clustered data. Although it is possible to use lavaan for multilevel structural equation modeling, only listwise deletion can be used for handling missing data.

MplusAutomation

Mplus (Muthén and Muthén, 1998-2017) is a comprehensive program for structural equation modeling (SEM) including latent variable modeling. Mplus is especially popular among applied researchers. It is syntax-based but relatively easy to use. It has many capabilities for advanced analysis (e.g., multilevel latent variable modeling, intensive longitudinal data analysis, Bayesian analysis) and can handle many data issues (e.g., missing data, non-normality, clustered data, complex survey designs); new methodologies are routinely added for its development. The most recent version is Version 8. Unlike R, it is not open source, and the user purchases licenses for the software and technical support services.

One drawback of Mplus is that the input or output of every model is stored as a separate file (.inp for input files and .out for output files). If one is to run many models, extracting information from the many output files can be a problem. The process can be tedious and error-prone. In addition, while Mplus is great for modeling, it has very limited capability for data management either to prepare data for the model or to further process data contained in the output files. To address these problems, Hallquist developed the R MplusAutomation package that can create, batch run, and extract results from many models (Hallquist and Wiley, 2018). Data to be analyzed can be managed in R like other R objects; sections of the Mplus input syntax are embedded in the object created by calling the mplusObject function; the mplusModeler function creates Mplus input files as well as dataset files if requested; the runModels function runs a group of Mplus models; and the Mplus output files (i.e., those with.out extension) can be extracted using the readModels. In addition, the MplusAutomation package provides functions to tabulate summary statistics, compare models, and extract parameters.

Tidyverse

Another useful R package for programming large-scale analysis with LSAs is tidyverse (Wickham et al., 2019). Technically, tidyverse is not a single R package; rather, it is a collection of R packages that share an underlying design philosophy, grammar, and data structures, which makes data wrangling, analysis, and visualization relatively easy.

For cross-national analysis using data from LSAs, it is necessary to process data before and after modeling them. While it is possible to use other packages (e.g., the "data.table" package; Dowle and Srinivasan, 2019) or the R base package to get the same results, I chose tidyverse because of its comprehensiveness and because it is relatively easy to use. The functions used in the present study are a tiny part of all the capacity of tidyverse. Here I would like to particularly point out the pipe operator (%>%) and the map function. The pipe operator comes from the magrittr (Bache and Wickham, 2014) package but is loaded automatically with tidyverse. It chains sequential operations to avoid creating intermediate objects and nested function calls and to make the syntax more readable. The map function is from the purr package (Henry and Wickham, 2020a) which is also loaded automatically with tidyverse. It takes a vector and a function as function inputs (i.e., arguments), applies the function to each element of the vector, and returns the results in a list of the same length. It is an efficient way of eliminating "for" loops so that the code is easier to write and read. If the output is more desired in a vector format, there are four variants which return a specific type of results: map_lgl (for a logical vector), map_int (for an integer vector), map_dbl (for a double vector), and map_chr (for a character vector).

The Big-Fish-Little-Pond Effect

When students compare their ability in an academic subject, they tend to compare themselves in their immediate context. As a result, students in higher-achieving contexts have lower self-concept than similarly able students in lower-achieving contexts. This phenomenon is called the big-fish-little-pond effect (BFLPE; Marsh, 1990). The BFLPE can be explained by the social comparison theory. According to this theory, individuals evaluate themselves by comparing themselves to others (Festinger, 1954; Suls et al., 2002). For such comparisons, those in an individual's immediate social group often serve as the comparison target (Rogers et al., 1978). To evaluate one's academic ability, the student may compare his/her academic position with their classmates when they form their academic self-concept. As a result, students from different classes may have different self-evaluations even when their academic abilities are the same.

Due to its social comparison nature, the BFLPE is a contextual effect, which occurs when the aggregate of a person-level characteristic (e.g., mathematics ability) is related to the outcome (e.g., mathematics self-concept) even after controlling for the effect of the individual characteristic; in other words, the "context" has an additional effect on the individual. Contextual effects can be examined using multilevel modeling statistical techniques (Raudenbush and Bryk, 2002). In a two-level modeling framework (e.g., students nested within classes), if the predictor variable is grand-mean centered, the between-level effect is the contextual effect; if the predictor is group-mean centered, the difference between the between-level effect and the within-level effect is the contextual effect.

TIMSS 2015

TIMSS is an international assessment of student achievement in mathematics and science in fourth and eighth grades. It is sponsored by IEA and directed by the TIMSS & PIRLS International Study Center at Boston College. The first TIMSS was administered in 1995 and has been administered every 4 years since then. TIMSS 2015 was the sixth cycle and is the most recent administration with data released to the public (TIMSS 2019 results are expected to be released in December 2020). In addition to tests measuring achievement, background and non-cognitive information is collected from students, teachers, and school principals, allowing researchers to examine relationships between achievement and personal and contextual factors across countries/regions.

Large-scale assessments have been used to study the BFLPE across countries. Marsh and Hau (2003) used the Program of Student Assessment (PISA) 2000 data collected in 26 countries; Seaton et al. (2009) used PISA 2003 data collected in 41 countries; Nagengast and Marsh (2011) used PISA 2006 data and examined the BFLPE with a total international sample from 57 countries, a total United Kingdom sample, and four samples from United Kingdom counties. Using TIMSS 2007, Wang (2015) examined the BFLPE in 49 countries in eighth-grade mathematics. Wang and Bergin (2017) further examined the BFLPE in 59 countries using TIMSS 2011 in eighth-grade mathematics. However, no study has investigated the BFLPE across many countries using TIMSS 2015.

MATERIALS AND METHODS

Samples, Variables, and Data

The present study used TIMSS 2015 data from 56 countries at the fourth-grade level and 46 countries at the eighth-grade level (Foy, 2017). The total sample consisted of 330,204 students from 15,740 classes in 10,964 schools in fourth grade and 285,190 students from 11,856 classes in 8,500 schools in eighth grade (see **Tables 1**, **2**).

Mathematics self-concept was measured by three items in each grade: (a) I usually do well in mathematics; (b) I am just not good at mathematics (for fourth-graders) / Mathematics is not one of my strengths (for eighth-graders); and (c) I learn things quickly in mathematics. This conceptualization of mathematics self-concept is consistent with Wang (2015); Wang and Bergin (2017) but differs from other articles using TIMSS data such as Marsh et al. (2014, 2015), which included a perceived relative standing item, *Mathematics is more difficult for me than for many of my classmates* for eighth-grade. A similar item in fourth grade is *Mathematics is harder for me than for many of my classmates.* Wang and Bergin (2017) argued that the perceived relative standing item should be separated from the self-concept items.

The three mathematics self-concept items were rated on a 1 to 4 Likert-scale (1 = Agree a lot, 2 = Agree a little, 3 = Disagree a little, 4 = Disagree a lot) and positively worded items were reverse coded so that a higher value corresponded to a higher level of self-concept. Mathematics self-concept was modeled as a latent variable with the three items as indicators and decomposed as having a within and between components during statistical modeling.

TIMSS databases use matrix sampling for the design of test administration where each student answered some but not all items on the test. Student achievement was estimated using item response theory together with a multiple imputation technique. Each student's mathematics achievement was measured by five plausible values. Those plausible values are not appropriate for reporting individual achievement and are suitable for estimating group characteristics (Wu, 2005). When used for statistical analysis, the five plausible values are treated as multiply imputed values: the analysis is run five times, each time using a single plausible value, and the five sets of results are then combined for point estimates and statistical inference (Enders, 2010).

Data collected in each country are hierarchical because schools were selected first and then classes were selected within schools and either all or sampled students responded to the student survey and the achievement test. For the three-sampling-stage process, TIMSS used probability proportional to size (PPS) sampling to select schools, classes, and students so that schools with more students had a higher probability of being selected and each individual student in the population had roughly the same probability of being selected. Probability weights and adjustment variables for non-responses were calculated for each sampling stage. For analysis using data from each country, a two-level modeling technique was adopted: the within-level was the student level and the between-level was the class level, further clustering at the school-level was accommodated at the between-level by incorporating the probability weights and adjustment factors of selecting schools.

Statistical Modeling

Two statistical models are used corresponding to the first two models in Wang and Bergin (2017). The first statistical model is the multilevel confirmatory factor analysis (CFA) model, which was applied for cross-level measurement invariance testing and separately for each country and grade (see **Figure 1**). The second statistical model is the multilevel SEM model where there are within and between effects of mathematics achievement on mathematics self-concept (see **Figure 2**). The rescaled difference between the between and within effects is the BFLPE.

TABLE 1 | Results of Model 1 in Fourth Grade.

				χ^2					Within variance		Between variance		ICC	
Country	**#Schools**	**#Classes**	**#Students**	**Est.**	**p**	**CFI**	**TLI**	**RMSEA**	**Est.**	***p***	**Est.**	***p***	**Est.**	***p***
Abu Dhabi, United Arab Emirates	163	219	5001	6.07	0.05	0.99	0.97	0.021	0.332	<0.001	0.023	<0.001	0.065	<0.001
Buenos Aires, Argentina	136	292	6435	2.21	0.33	1.00	1.00	0.004	0.638	<0.001	0.043	<0.001	0.062	<0.001
Dubai, United Arab Emirates	168	316	7453	1.30	0.52	1.00	1.00	0.000	0.454	<0.001	0.048	<0.001	0.095	<0.001
United Arab Emirates	558	891	21177	5.70	0.06	1.00	0.99	0.009	0.348	<0.001	0.034	<0.001	0.090	<0.001
Armenia	148	234	5384	6.81	0.03	1.00	0.99	0.022	0.724	<0.001	0.047	0.004	0.061	0.003
Australia	287	498	6057	2.63	0.27	1.00	1.00	0.007	0.628	<0.001	0.009	0.071	0.015	0.070
Belgium (Flemish)	153	295	5404	36.99	0.00	0.99	0.97	0.057	0.758	<0.001	0.023	0.001	0.030	<0.001
Bulgaria	149	233	4228	0.19	0.91	1.00	1.00	0.000	0.629	<0.001	0.097	<0.001	0.134	<0.001
Bahrain	182	345	8575	0.97	0.62	1.00	1.00	0.000	0.084	0.010	0.005	0.040	0.057	0.003
Canada	441	696	12283	2.34	0.31	1.00	1.00	0.004	0.666	<0.001	0.030	0.001	0.042	<0.001
Chile	179	179	4756	0.10	0.95	1.00	1.00	0.000	0.612	<0.001	0.043	<0.001	0.066	<0.001
ON, Canada	151	271	4574	1.58	0.45	1.00	1.00	0.000	0.669	<0.001	0.034	0.007	0.049	0.002
QC, Canada	121	152	2798	7.71	0.02	0.99	0.98	0.032	0.638	<0.001	0.024	0.034	0.036	0.032
Cyprus	148	243	4125	7.01	0.03	1.00	0.99	0.025	0.660	<0.001	0.036	<0.001	0.052	<0.001
Czechia	159	265	5202	50.40	0.00	0.98	0.93	0.068	0.626	<0.001	0.009	0.090	0.014	0.081
Germany	204	213	3948	1.91	0.38	1.00	1.00	0.000	0.676	<0.001	0.017	0.012	0.025	0.012
Denmark	193	194	3710	0.60	0.74	1.00	1.00	0.000	0.697	<0.001	0.025	0.001	0.034	0.001
England	147	176	4006	2.69	0.26	1.00	1.00	0.009	0.615	<0.001	0.040	<0.001	0.061	<0.001
Spain	358	379	7764	6.38	.04	1.00	0.99	0.017	0.632	<0.001	0.039	<0.001	0.058	<0.001
Finland	158	300	5015	2.41	0.30	1.00	1.00	0.006	0.670	<0.001	0.015	0.025	0.022	0.025
France	164	273	4873	53.12	0.00	0.97	0.91	0.073	0.615	<0.001	0.022	0.004	0.034	0.002
Georgia	153	188	3919	8.56	0.01	0.99	0.96	0.029	0.342	<0.001	0.029	0.013	0.078	0.003
Hong Kong SAR	132	145	3600	0.55	0.76	1.00	1.00	0.000	0.666	<0.001	0.021	0.005	0.030	0.005
Croatia	163	223	3985	2.60	0.27	1.00	1.00	0.009	0.597	<0.001	0.017	0.008	0.028	0.007
Hungary	144	241	5036	4.09	0.13	1.00	1.00	0.014	0.704	<0.001	0.021	0.012	0.030	0.012
Indonesia	230	312	8319	2.60	0.27	1.00	1.00	0.006	0.255	<0.001	0.097	0.001	0.277	<0.001
Ireland	149	214	4344	1.81	0.40	1.00	1.00	0.000	0.678	<0.001	0.011	0.088	0.017	0.082
Iran, Islamic Rep. of	248	291	7928	0.46	0.80	1.00	1.00	0.000	0.381	<0.001	0.038	0.005	0.090	<0.001
Italy	164	257	4373	102.62	0.00	0.93	0.79	0.108	0.626	<0.001	0.019	0.009	0.030	0.010
Jordan	254	272	7861	3.05	0.22	0.99	0.98	0.008	0.026	0.326	0.004	0.379	0.138	0.003
Japan	148	148	4383	0.62	0.73	1.00	1.00	0.000	0.587	<0.001	0.020	0.003	0.033	0.003
Kazakhstan	171	239	4702	0.44	0.80	1.00	1.00	0.000	0.489	<0.001	0.078	<0.001	0.138	<0.001
Korea, Rep. of	149	188	4669	1.42	0.49	1.00	1.00	0.000	0.733	<0.001	0.042	<0.001	0.054	<0.001
Kuwait	166	294	7296	0.08	0.96	1.00	1.00	0.000	0.076	0.003	0.015	0.016	0.161	<0.001
Lithuania	225	290	4529	4.86	0.09	1.00	0.99	0.018	0.692	<0.001	0.013	0.040	0.018	0.044
Morocco	358	374	10428	2.05	0.36	1.00	1.00	0.002	0.325	<0.001	0.118	<0.001	0.267	<0.001
Northern Ireland	118	153	3116	0.35	0.84	1.00	1.00	0.000	0.655	<0.001	0.019	0.050	0.028	0.044
Netherlands	129	223	4515	8.91	0.01	1.00	0.99	0.028	0.765	<0.001	0.003	0.430	0.004	0.430
Norway (4th grade)	139	219	4164	1.44	0.49	1.00	1.00	0.000	0.550	<0.001	0.032	0.002	0.055	0.001
Norway	140	222	4329	1.64	0.44	1.00	1.00	0.000	0.656	<0.001	0.017	0.030	0.025	0.028
New Zealand	174	459	6322	8.67	0.01	1.00	0.99	0.023	0.645	<0.001	0.028	<0.001	0.041	<0.001
Oman	300	353	9105	1.58	0.45	1.00	1.00	0.000	0.025	0.777	0.002	0.781	0.060	0.145
Poland	150	254	4747	29.55	0.00	0.99	0.97	0.054	0.643	<0.001	0.030	0.013	0.045	0.008
Portugal	217	321	4693	0.19	0.91	1.00	1.00	0.000	0.599	<0.001	0.042	<0.001	0.066	<0.001
Qatar	211	224	5194	8.23	0.02	0.99	0.97	0.025	0.235	<0.001	0.016	<0.001	0.065	<0.001
Russian Federation	208	217	4921	5.33	0.07	1.00	1.00	0.018	0.539	<0.001	0.022	0.001	0.039	<0.001
Saudi Arabia*	189	189	4337	6.77	0.03	0.97	0.90	0.024	−0.135	0.256	−0.018	0.275	0.120	0.003
Singapore	179	358	6517	39.19	0.00	0.98	0.95	0.053	0.602	<0.001	0.137	<0.001	0.185	<0.001
Serbia	160	192	4036	0.85	0.65	1.00	1.00	0.000	0.595	<0.001	0.036	0.001	0.056	<0.001
Slovak Republic	198	327	5773	18.39	0.00	0.99	0.96	0.038	0.620	<0.001	0.033	0.001	0.051	<0.001
Slovenia	148	255	4445	1.61	0.45	1.00	1.00	0.000	0.621	<0.001	0.016	0.014	0.025	0.013
Sweden	144	211	4142	2.73	0.25	1.00	1.00	0.009	0.633	<0.001	0.019	0.008	0.029	0.008
Turkey	242	251	6456	2.56	0.28	1.00	1.00	0.007	0.497	<0.001	0.040	<0.001	0.074	<0.001
Chinese Taipei	150	177	4291	2.45	0.29	1.00	1.00	0.007	0.658	<0.001	0.019	<0.001	0.028	<0.001
United States	250	497	10029	21.54	.00	0.99	0.98	0.032	0.630	<0.001	0.032	<0.001	0.049	<0.001
South Africa	297	298	10932	3.54	0.17	1.00	0.99	0.009	0.285	<0.001	0.027	0.004	0.087	<0.001

*CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; ICC, intraclass correlation coefficient. *model run with 10 random sets of starting values.*

TABLE 2 | Results of Model 1 in Eighth Grade.

				χ^2					Within variance		Between variance		ICC	
Country	**#Schools**	**#Classes**	**#Students**	**Est.**	**p**	**CFI**	**TLI**	**RMSEA**	**Est.**	***p***	**Est.**	***p***	**Est.**	**p**
Abu Dhabi, United Arab Emirates	156	208	4838	3.02	0.22	1.00	1.00	0.010	0.503	<0.001	0.049	<0.001	0.089	<0.001
Buenos Aires, Argentina	128	138	3253	1.60	0.45	1.00	1.00	0.000	0.748	<0.001	0.054	<0.001	0.067	<0.001
Dubai, United Arab Emirates	135	264	6149	1.62	0.44	1.00	1.00	0.000	0.573	<0.001	0.060	<0.001	0.095	<0.001
United Arab Emirates	477	763	18012	1.84	0.40	1.00	1.00	0.000	0.529	<0.001	0.055	<0.001	0.094	<0.001
Armenia	145	228	5060	0.86	0.65	1.00	1.00	0.000	0.906	<0.001	0.058	<0.001	0.060	<0.001
Australia	285	645	10338	31.77	0.00	0.99	0.97	0.039	0.600	<0.001	0.099	<0.001	0.142	<0.001
Bahrain	105	197	4918	3.23	.20	1.00	0.99	0.011	0.112	0.166	0.009	0.194	0.072	<0.001
Botswana	159	169	5964	2.47	.29	1.00	1.00	0.006	0.453	<0.001	0.046	<0.001	0.092	<0.001
Canada	276	409	8757	13.80	0.00	1.00	0.99	0.026	0.648	<0.001	0.077	<0.001	0.107	<0.001
Chile	171	173	4849	5.30	0.07	1.00	1.00	0.019	0.706	<0.001	0.033	<0.001	0.044	<0.001
ON, Canada	138	217	4520	3.41	0.18	1.00	1.00	0.013	0.666	<0.001	0.063	<0.001	0.086	<0.001
QC, Canada	122	175	3950	8.92	.01	1.00	0.99	0.030	0.649	<0.001	0.094	<0.001	0.126	<0.001
Egypt	211	215	7822	2.23	0.33	1.00	1.00	0.004	1.586	0.279	0.152	0.330	0.087	<0.001
England	143	213	4814	55.00	0.00	0.98	0.93	0.075	0.508	<0.001	0.160	<0.001	0.239	<0.001
Georgia	153	187	4035	4.82	0.09	1.00	0.99	0.019	0.544	<0.001	0.041	<0.001	0.070	<0.001
Hong Kong SAR	133	145	4155	2.19	0.33	1.00	1.00	0.005	0.684	<0.001	0.038	<0.001	0.053	<0.001
Hungary	144	241	4893	4.42	0.11	1.00	1.00	0.016	0.711	<0.001	0.054	<0.001	0.070	<0.001
Ireland	149	204	4704	6.53	0.04	1.00	1.00	0.022	0.654	<0.001	0.031	<0.001	0.045	<0.001
Iran, Islamic Rep. of	250	251	6130	2.71	0.26	1.00	1.00	0.008	0.669	<0.001	0.066	<0.001	0.089	<0.001
Israel	200	200	5512	0.25	0.88	1.00	1.00	0.000	0.632	<0.001	0.061	<0.001	0.088	<0.001
Italy	161	230	4481	0.53	0.77	1.00	1.00	0.000	0.692	<0.001	0.038	<0.001	0.052	<0.001
Jordan*	252	260	7865	5.31	0.07	0.99	0.98	0.015	−0.067	0.470	−0.005	0.470	0.071	<0.001
Japan	147	147	4745	1.43	0.49	1.00	1.00	0.000	0.641	<0.001	0.012	0.006	0.018	0.006
Kazakhstan	172	239	4887	1.17	.56	1.00	1.00	0.000	0.540	<0.001	0.071	<0.001	0.116	<0.001
Korea, Rep. of	150	170	5309	34.13	0.00	0.99	0.97	0.055	0.850	<0.001	0.017	0.011	0.020	0.011
Kuwait	168	191	4503	1.48	0.48	1.00	1.00	0.000	0.227	<0.001	0.027	<0.001	0.105	<0.001
Lebanon	138	185	3873	9.06	0.01	0.99	0.96	0.032	0.608	<0.001	0.031	0.025	0.048	0.019
Lithuania	208	252	4347	3.78	0.15	1.00	1.00	0.014	0.719	<0.001	0.044	0.001	0.058	<0.001
Morocco	345	375	13035	2.10	0.35	1.00	1.00	0.002	0.151	0.472	0.010	0.479	0.061	<0.001
Malta	48	223	3817	3.43	0.18	1.00	1.00	0.014	0.632	<0.001	0.108	<0.001	0.146	<0.001
Malaysia	207	326	9726	8.87	0.01	0.99	0.98	0.019	0.514	<0.001	0.060	<0.001	0.104	<0.001
Norway (8th grade)	142	216	4795	0.45	0.80	1.00	1.00	0.000	0.686	<0.001	0.023	0.002	0.033	0.002
Norway	143	216	4697	9.45	0.01	1.00	0.99	0.028	0.731	<0.001	0.021	0.003	0.028	0.002
New Zealand	145	377	8142	6.23	0.04	1.00	0.99	0.016	0.613	<0.001	0.048	<0.001	0.072	<0.001
Oman	301	356	8883	1.42	0.49	1.00	1.00	0.000	0.167	<0.001	0.016	0.001	0.087	<0.001
Qatar	131	238	5403	1.58	0.45	1.00	1.00	0.000	0.418	<0.001	0.037	<0.001	0.081	<0.001
Russian Federation	204	221	4780	7.35	0.03	1.00	0.99	0.024	0.546	<0.001	0.032	<0.001	0.055	<0.001
Saudi Arabia	143	149	3759	43.17	0.00	0.92	0.77	0.075	0.000	1.000	0.000	1.000	0.093	1.000
Saudi Arabia*				12.18	0.00	0.98	0.94	0.037	−0.746	<0.001	−0.062	<0.001	0.077	0.006
Singapore	167	334	6116	0.51	0.77	1.00	1.00	0.000	0.705	<0.001	0.072	<0.001	0.093	<0.001
Slovenia	148	217	4257	4.78	0.09	1.00	1.00	0.018	0.665	<0.001	0.016	0.006	0.024	0.005
Sweden	150	206	4090	15.73	0.00	1.00	0.99	0.041	0.666	<0.001	0.039	<0.001	0.056	<0.001
Thailand	204	213	6482	1.05	0.59	1.00	1.00	0.000	0.544	<0.001	0.077	<0.001	0.124	<0.001
Turkey	218	220	6079	16.90	0.00	1.00	0.99	0.035	0.692	<0.001	0.056	<0.001	0.075	<0.001
Chinese Taipei	190	191	5711	4.93	0.09	1.00	1.00	0.016	0.744	<0.001	0.025	<0.001	0.032	<0.001
United States	246	534	10221	30.98	0.00	1.00	0.99	0.038	0.558	<0.001	0.093	<0.001	0.143	<0.001
South Africa	292	328	12514	18.26	0.00	0.99	0.97	0.026	0.475	<0.001	0.080	<0.001	0.143	<0.001

*CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; ICC, intraclass correlation coefficient. *model run with 10 random sets of starting values.*

To illustrate the models, let $\mathbf{Y_{ij}}$ be a vector with three elements, representing values of the three mathematics self-concept items and let sc_{ij} be the latent mathematics self-concept for student *i* in class *j*. In a single-level CFA model, $\mathbf{Y_{ij}}$ is the vector of indicators of sc_{ij}. In the two-level model, sc_{ij} is decomposed into a within and a between component.

$$sc_{ij} = sc_{wij} + sc_{bj} \quad (1)$$

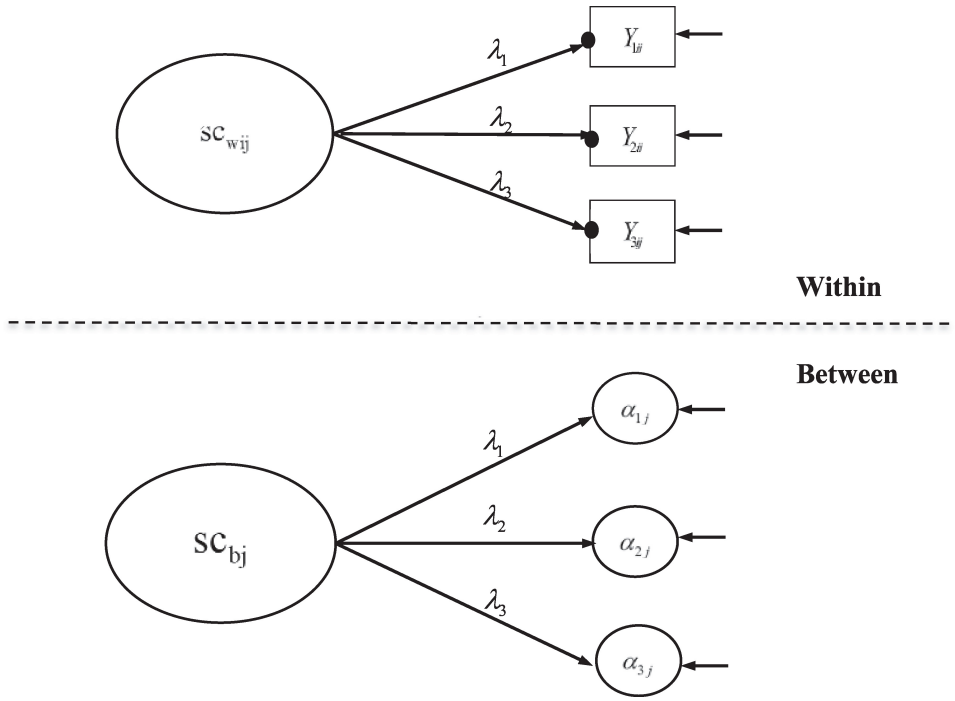

FIGURE 1 | Statistical Model 1 – two-level confirmatory factor analysis model with multilevel measurement invariance of mathematics self-concept. The solid dots indicate random intercepts for different classes. Reprinted with permission from Wang and Bergin (2017); Copyright 2017 by Elsevier.

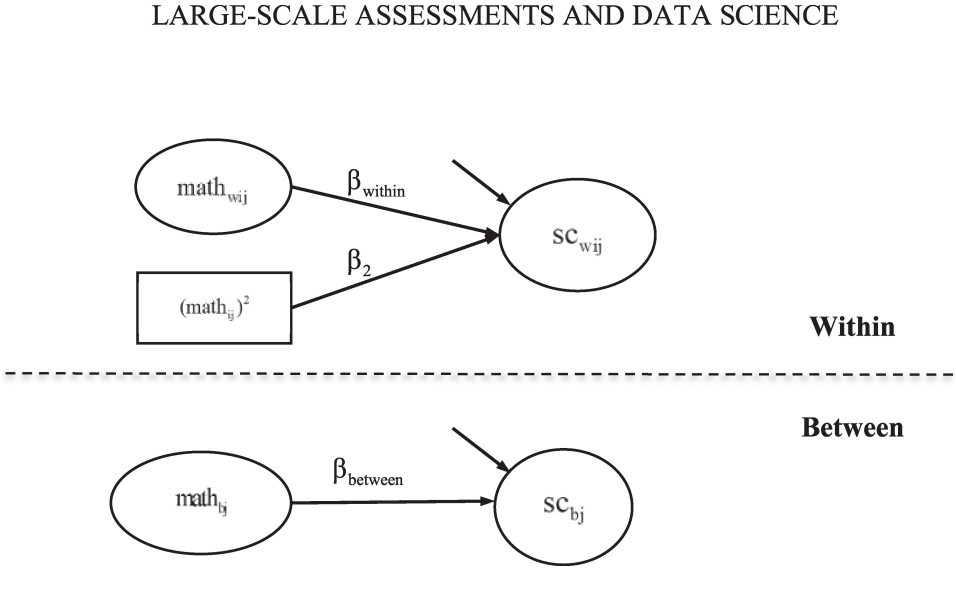

FIGURE 2 | Statistical Model 2 – Model to test the big-fish-little-pond effect. Indicators of the within- and between-level mathematics self-concept are not shown in figure. Reprinted with permission from Wang and Bergin (2017); Copyright 2017 by Elsevier.

where sc_{wij} is the within component and $sc_{bj}$is the between component. sc_{wij} and sc_{bj} each is measured by three indicators as shown in equations (2) and (3), respectively.

$$\mathbf{Y_{ij}} = \boldsymbol{\alpha_j} + \boldsymbol{\lambda} sc_{wij} + \mathbf{e_{ij}} \quad (2)$$

$$\boldsymbol{\alpha_j} = \boldsymbol{\gamma} + \boldsymbol{\lambda} sc_{bj} + \mathbf{r_j} \quad (3)$$

$\boldsymbol{\alpha_j}$ represents class-specific indicator intercepts at the within level that function as indicators of the latent factor sc_{bj} at the between level; $\boldsymbol{\gamma}$ is a vector of constants representing the grand mean indicator intercepts at the between level. $\boldsymbol{\lambda}$ is a vector of factor loadings that are invariant across levels. The invariance of cross-level factor loadings ensures that the

interpretation of mathematics self-concept at the within and between levels is the same.

The predictor, mathematics achievement, is also decomposed into a within and a between components.

$$\text{math}_{ij} = \mu_{math} + \text{math}_{wij} + \text{math}_{bj} \quad (4)$$

math_{ij} is the mathematics achievement of student i in class j; μ_{math} is a constant representing the grand mean of mathematics achievement for all students in all classes; math_{wij} is student i's mathematics achievement around the class-average mathematics achievement; and math_{bj} is the average mathematics achievement for class j, around the grand mean.

Further, the relationship between mathematics self-concept and mathematics achievement is modeled at the two levels in equations (5) and (6), respectively.

$$\text{sc}_{wij} = \beta_{within}\text{math}_{wij} + \beta_2(\text{math}_{ij})^2 + \varepsilon_{ij} \quad (5)$$

$$\text{sc}_{bj} = \beta_{between}\text{math}_{bj} + \delta_j \quad (6)$$

In equation (5), the quadratic component of student mathematics achievement is included following previous BFLPE research (e.g., Marsh and Hau, 2003). The standardized effect size of the BFLPE can be calculated as:

$$\text{ES}_{\text{BFLPE}} = 2 \times (\beta_{between} - \beta_{within}) \times \sqrt{\text{Var}(\text{math}_{bj})}/\sqrt{\text{Var}(\text{sc}_{wij}) + \text{Var}(\text{sc}_{bj})} \quad (7)$$

$\text{Var}(\text{math}_{bj})$, $\text{Var}(\text{sc}_{wij})$, and $\text{Var}(\text{sc}_{bj})$ are the variances of the between-level mathematics achievement, the within-level self-concept, and the between-level self-concept, respectively. The detailed calculations of those variances are illustrated in Wang and Bergin (2017).

The Syntax

Here I explain the R syntax to examine the BFLPE using TIMSS 2015 data. First, Mplus and R are installed. I also recommend RStudio be installed. All syntax is written using RStudio as an R script (R scripts are like text files). Next, start R or RStudio and install the packages EdSurvey, MplusAutomation, and tidyverse. I also install the rlang package (Henry and Wickham, 2020b) for two functions related to expressions that are used in extracting model fit indices. After the packages are installed, load them using the library function. Packages only need to be installed once on the computer. However, they have to be loaded every time R or RStudio is started.

```
install.packages("EdSurvey")
install.packages ("MplusAutomation")
install.packages ("tidyverse")
install.packages ("rlang")

library(EdSurvey)
library(MplusAutomation)
library(tidyverse)
library(rlang)
```

For large-scale analysis with many files, it is important to have a good file system. All related files for the present study are stored in the folder called "BFLPE study." This folder is created manually in the C: drive and set as the working directory. Alternatively, one can create an R project using RStudio and associate the R project with this working directory. Except for this main folder, all other folders and their contents are created by running syntax in RStudio.

Under this "BFLPE study" folder, there are three subfolders called, "TIMSS," "Mplus_g4," and "Mplus_g8," respectively. The "TIMSS" folder has a subfolder named "2015" inside which are TIMSS 2015 datasets downloaded via an internet connection, as well as META files and text files to be created to facilitate fast reading of data using the EdSurvey package. The "Mplus_g4" folder has two subfolders: "Model1" and "Model2," corresponding to the two statistical models. "Model 1" includes all Mplus input, output, and data files used for the first statistical model (i.e., multilevel CFA) for all countries at the fourth-grade level and "Model 2" has all Mplus input, output, and data files for the second statistical model (i.e., the BFLPE model) for all countries at the fourth-grade level. The "Mplus_g8" folder also has two subfolders "Model1" and "Model2" with similar information but for eighth grade.

```
setwd("C:/BFLPE study")
dir.create("Mplus_g8")
dir.create("Mplus_g4")
dir.create("Mplus_g8/Model1")
dir.create("Mplus_g8/Model2")
dir.create("Mplus_g4/Model1")
dir.create("Mplus_g4/Model2")

downloadTIMSS(years = 2015,
              root = "C:/BFLPE study")

TIMSS_15_g8 <-readTIMSS(path = "./TIMSS/2015",
              countries = "*", gradeLvl = 8)
```

The object TIMSS_15_g8 is a survey data frame (SDF) which stores all TIMSS 2015 information from the student survey, teacher survey, school survey, as well as achievement information in eighth grade for all participating countries. For the remainder of this section, only syntax relevant to eighth-grade analysis is presented. Interested readers can easily modify the syntax to suit for fourth grade.

For this study, students' mathematics achievement, the three items measuring their mathematics self-concept, and weight variables and adjustment factors accounting for the PPS sampling are used for analysis. Clustering within classes and schools are considered; country id and student id variables are specified as auxiliary variables just for quality control purposes (i.e., to make sure the data are created and used correctly).

The mathematics self-concept items in the SDF are stored as factors and need to be converted to numeric variables. Missing values are specified. All observed variables are standardized within each country. Weight variables and the square term of mathematics achievement used in the Mplus input files are created. For statistical model 1 (i.e., the multilevel CFA model), mathematics achievement data are not used so there is a single dataset for each country. For statistical model 2 (the BFLPE model), each plausible value of mathematics achievement is stored in a different dataset for a total of five datasets for each

country. These five datasets are used as imputed datasets in the Mplus syntax.

For analysis, two mathematics self-concept items have to be reverse coded.

```
data_g8 <- getData(data = TIMSS_15_g8,
             varnames = c("mmat", "bsbm19a",
             "bsbm19c", "bsbm19d", "idcntry",
             "idschool", "idclass", "idstud",
             "wgtadj1", "wgtadj2", "wgtadj3",
             "wgtfac1", "wgtfac2", "wgtfac3"),
addAttributes = TRUE, omittedLevels = FALSE)

length(data_g8) #2
length(data_g8[[1]]) #46
length(data_g8[[2]]) #1
data_g8[[2]] # names of 46 countries

reverse_cols_g8 <- c("bsbm19a", "bsbm19d")
                   #variables to reverse code
use_cols_g8_model1 <- c("bsbm19a", "bsbm19c",
                   "bsbm19d", "idcntry", "idschool",
                   "idclass", "idstud", "wt1", "wt2")
                   #variables to be used in model 1
use_cols_g8_model2 <- c("math", "mathsq",
                   "bsbm19a", "bsbm19c", "bsbm19d",
                   "idcntry", "idschool", "idclass",
                   "idstud", "wt1", "wt2")
                   #variables to be used in model 2
scale2 <- function(x, na.rm = FALSE)
      (x - mean(x, na.rm = na.rm)) /sd(x, na.rm)
            # a function to standardize variable
reverse <- function(x) 5-x

dd_g8 <- list()
idat_g8 <- list()
for (i in seq_along(data_g8[[1]])) {
  dd_g8[[i]] <- data_g8[[1]][[i]] %>%
as_tibble %>%
mutate_at(c("bsbm19a", "bsbm19c", "bsbm19d"),
             as.numeric) %>%
mutate_at(c("bsbm19a", "bsbm19c", "bsbm19d"),
             ~na_if(.,9)) %>%
mutate_at(c("wgtadj1", "wgtadj2", "wgtadj3",
             "wgtfac1", "wgtfac2", "wgtfac3"),
             ~na_if(.,999999.000000)) %>%
mutate_at(reverse_cols_g8, reverse) %>% #reverse code
mutate_at(c("bsbm19a", "bsbm19c", "bsbm19d",
             "bsmmat01", "bsmmat02", "bsmmat03",
             "bsmmat04", "bsmmat05"),scale2,
             na.rm = TRUE) %>%
mutate(wt1 = wgtadj3*wgtfac3) %>% #within-level weight
mutate(wt2 = wgtadj1*wgtfac1*wgtadj2*wgtfac2)
                                #between-level weight
  idata1 <- dd_g8[[i]] %>% select(-c(bsmmat02,
                   bsmmat03,bsmmat04,bsmmat05)) %>%
rename(math = bsmmat01) %>% mutate(mathsq = math**2)
  idata2 <- dd_g8[[i]] %>% select(-c(bsmmat01,
                          bsmmat03,bsmmat04,bsmmat05)) %>%
rename(math = bsmmat02) %>% mutate(mathsq = math**2)
  idata3 <- dd_g8[[i]] %>% select(-c(bsmmat01,
                          bsmmat02,bsmmat04,bsmmat05)) %>%
rename(math = bsmmat03) %>% mutate(mathsq = math**2)
  idata4 <- dd_g8[[i]] %>% select(-c(bsmmat01,
                          bsmmat02,bsmmat03,bsmmat05)) %>%
rename(math = bsmmat04) %>% mutate(mathsq = math**2)
  idata5 <- dd_g8[[i]] %>% select(-c(bsmmat01,
                          bsmmat02,bsmmat03,bsmmat04)) %>%
rename(math = bsmmat05) %>% mutate(mathsq = math**2)
  idat_g8[[i]] <- list(idata1[,use_cols_g8_model2],
                   idata2[,use_cols_g8_model2],
                   idata3[,use_cols_g8_model2],
                   idata4[,use_cols_g8_model2],
                   idata5[,use_cols_g8_model2])
  dd_g8[[i]] <- dd_g8[[i]]
      %>% select(all_of(use_cols_g8_model1))
}
```

For each of the 46 countries, an object is created – using the mplusObject function – that contains all syntax sections needed to create a Mplus input file for Model 1 for that country. Next, the Mplus input file is created and run using the mplusModeler function. Iterations on countries are done using the map function. Datasets are exported and Mplus output files are created when the model is run. The readModels function extracts information in all Mplus output files in the folder.

The next step after reading Mplus output files is usually to get some type of summary tables. However, for large-scale analysis using LSAs, oftentimes, the model for a few countries may not run properly. In that case, functions such as Summary Table and paramExtract of the MplusAutomation package will not work well and will give errors. Country 22 was the problematic one. Here I simply skip the Mplus output for this country and will manually revise the Mplus input file for this country later. To skip country 22's Mplus output, I change its Mplus output file extension to.didnotrun so that this file would not be read using the readModels function. I also calculate the number of schools, classes, and students in each country.

The readModels function imports results into R as mplus.model objects with a predictable structure. This structure, shown in Table 3 of Hallquist and Wiley (2018), serves as a guide to what can be extracted from Mplus outputs.

```
model1_g8 <- function(i) {
bflpe <- mplusObject(
TITLE = sprintf("Multilevel CFA model country%d", i),
VARIABLE = "auxiliary = idcntryidstud;
                  cluster = idschoolidclass;
                  weight is wt1;
wtscale is cluster;
bweight is wt2;
bwtscale is sample;",
ANALYSIS = "type = twolevel complex;",
MODEL = "%within%
scw by bsbm19a (1)

                    bsbm19c (3)
                    bsbm19d (4);
scw* (var1);

%between%
scb by bsbm19a (1)
       bsbm19c (3)
       bsbm19d (4);
scb* (var2);",
MODELCONSTRAINT = "NEW(ICC);
          ICC=var2/(var1+var2);",
MODELTEST = "var1=0; var2=0; ICC=0;",
OUTPUT = "",
usevariables = use_cols_g8_model1,
rdata = dd_g8[[i]])
mplusModeler(bflpe, dataout = paste0(getwd(),
             "/Mplus_g8/Model1/",sprintf("data%d", i)),
             modelout = paste0(getwd(),"/Mplus_g8/Model1/",
             sprintf("country%d.inp", i)), run = TRUE,
             writeData ='always', hashfilename = FALSE)
}
map(1:46,model1_g8)

model1_results_g8 <- readModels("./Mplus_g8/Model1")
## It is a good idea to check the model summaries
   one by one to detect problems.
# for(i in 1:46) {print(model1_results_g8[[i]]$summaries)}

## model 1 of country22 did not run successfully.
   Remove file "country22.out" from folder,
   or change the file extension to something else.
   Re-read the remaining Mplus output files.
```

```
oldname <- paste(getwd(), "Mplus_g8/Model1",
           "country22.out", sep = "/")
newname <- paste(getwd(), "Mplus_g8/Model1",
           "country22.didnotrun", sep = "/")
file.rename(oldname, newname)

# calculate # of schools, classes, and
  students in each country
n_school_g8 <- dd_g8 %>%
map("idschool") %>%
map(unique) %>%
map_int(length)

n_class_g8 <- dd_g8 %>%
map("idclass") %>%
map(unique) %>%
map_int(length)

n_student_g8 <- dd_g8 %>%
map("idstud") %>%

map(unique) %>%
map_int(length)

n_school_g8 <- n_school_g8[-22]
n_class_g8 <- n_class_g8[-22]
n_student_g8 <- n_student_g8[-22]

model1_results_g8 <- readModels("./Mplus_g8/Model1")
```

There are multiple summary and fit indices for the modeling results in each country. Extract such information can be easily done by applying the map function and its variants. To extract parameters, we need to know the position of the parameters in the results. For example, after viewing the model1_results_g8[[1]]$parameters$unstandardized object, the within-level variance of mathematics self-concept is in the fourth row. Its estimate and the p value of the estimate are extracted. All results for Model 1 are in the model1_table_g8 object.

The order of elements in R objects is important to match results for countries. The elements in edsurvey.data.frame.list objects in this study (e.g., TIMSS_15_g8 and data_g8) are in ascending order using three-letter country codes. Therefore, the first element is for Abu Dhabi, United Arab Emirates with country code "aad" and the second element is for Buenos Aires, Argentina using country code "aba." When the Mplus input and output files are created, I simply name them by their country number; therefore the Mplus input and output for Abu Dhabi, United Arab Emirates are country1.inp and country1.out, respectively; and the Mplus input and output for Buenos Aires, Argentina are country2.inp and country2.out, respectively. When reading Mplus outputs using the readModels function, the order in the resulted mplus.model object (the "model1_results_g8") is ascending alphabetically. Therefore, the first element in "model1_results_g8" has information for country1 and the second element has information for country10 (Chile) instead of country2. We reordered the elements in the mplus.model objects to be ascending according to country numbers (1, 2, 3, etc.).

```
model1_fit_indices <- c("Filename", "Observations",
                  "ChiSqM_Value", "ChiSqM_PValue",
                  "CFI", "TLI", "RMSEA_Estimate",
                  "SRMR.Within", "SRMR.Between")

model1_fit_g8 <- as_tibble(matrix(ncol
            = length(model1_fit_indices),nrow =45))
colnames(model1_fit_g8) <- model1_fit_indices

model1_fit_g8[,1] <- model1_results_g8 %>%
map("summaries") %>%
map_chr("Filename")

for (i in 2:length(model1_fit_indices)) {
  x <- model1_fit_indices[i]
  x <- sym(x) # a symbol is an expression.
            It is printed without quotes.
  model1_fit_g8[,i] <- model1_results_g8 %>%
map("summaries") %>%
map_dbl(as_string(x))
}

# order the Fit statistics tibble
according to the country index
countryi <- as.numeric(gsub("country([0-9]+).*$","\\1",
           model1_fit_g8$Filename))
model1_fit_g8 <- model1_fit_g8[order(countryi),]
country <- data_g8[[2]]$country[-22]
model1_fit_g8 <- as_tibble(cbind("country" =country,
                  "# schools" = n_school_g8,
                  "# classes" = n_class_g8,
                  "# students" = n_student_g8, model1_fit_g8))

#model1_results_g8[[1]]$parameters$unstandardized

para_unstandardized <- model1_results_g8 %>%
map("parameters") %>%
map("unstandardized")

within_variance <- para_unstandardized %>%
map_dbl(~.[4,]$est)
within_variance_pval <- para_unstandardized %>%
map_dbl(~.[4,]$pval)

between_variance <- para_unstandardized %>%
map_dbl(~.[14,]$est)
between_variance_pval <- para_unstandardized %>%
map_dbl(~.[14,]$pval)

ICC <- para_unstandardized %>%
map_dbl(~.[18,]$est)
ICC_pval <- para_unstandardized %>%
map_dbl(~.[18,]$pval)

within_variance <- within_variance[order(countryi)]
within_variance_pval <- within_variance_pval[order(countryi)]
between_variance <- between_variance[order(countryi)]
between_variance_pval <- between_variance_pval[order(countryi)]
ICC <- ICC[order(countryi)]
ICC_pval <- ICC_pval[order(countryi)]

(model1_table_g8 <- as_tibble(cbind(model1_fit_g8,
                  "within_variance" = within_variance,
                  "within_variance_pval" = within_variance_pval,
                   "between_variance" = between_variance,
                  "between_variance_pval" = between_variance_pval,
                  "ICC" =ICC, "ICC_pval" = ICC_pval)))
```

For eighth-grade Model 2, the flow is similar: create an object that contains Mplus input syntax sections, create Mplus input files, run the model in Mplus, output the data and Mplus output files, read the Mplus output files, and extract summary and parameter information from the Mplus output files.

```
model2_g8 <- function(i) {
bflpe <- mplusObject(
TITLE = sprintf("BFLPE model country%d", i),
VARIABLE = "auxiliary = idcntryidstud;
                  within = mathsq;
                  cluster = idschoolidclass;
                  weight is wt1;
wtscale is cluster;
bweight is wt2;
bwtscale is sample;",
ANALYSIS = "type = twolevel complex;",
MODEL = "%within%
scw by bsbm19a (1)
       bsbm19c (3)
       bsbm19d (4);
```

```
scw on mathsq (bb)
       math (b1);
scw (var1);
math (var2);
mathsq(var5);

%between%
scb by bsbm19a (1)
       bsbm19c (3)
       bsbm19d (4);
scb on
                 math (b2);
scb (var3);
math (var4);",
MODELCONSTRAINT = "new(esw); new(esb); new(esbflpe);
             esw=2*b1*sqrt(var2)/sqrt(b1**2*var2
                  +bb**2*var5+var1+b2**2*var4+var3);
             esb=2*b2*sqrt(var4)/sqrt(b1**2*var2
                  +bb**2*var5+var1+b2**2*var4+var3);
             esbflpe=2*(b2-b1)*sqrt(var4)/sqrt(b1**2*var2
                  +bb**2*var5+var1+b2**2*var4+var3);",
MODELTEST = "esw=0; esb=0; esbflpe=0;",
OUTPUT = "",
usevariables = use_cols_g8_model2,
rdata = idat_g8[[i]],
imputed = TRUE)
mplusModeler(bflpe, dataout =paste0(getwd(),
            "/Mplus_g8/Model2/",sprintf("data%d", i)),
            modelout =paste0(getwd(),"/Mplus_g8/Model2/",
            sprintf("country%d.inp", i)), run = TRUE,
            writeData ='always', hashfilename = FALSE)
}

map(1:46,model2_g8)


model2_results_g8 <- readModels("./Mplus_g8/Model2")

model2_fit_indices <- c("Filename", "ChiSqM_DF",
            "ChiSqM_Mean", "ChiSqM_SD",
            "ChiSqM_NumComputations", "LL_Mean",
            "LL_SD", "LL_NumComputations",
            "UnrestrictedLL_Mean", "UnrestrictedLL_SD",
            "UnrestrictedLL_NumComputations",
            "CFI_Mean", "CFI_SD", "CFI_NumComputations",
            "TLI_Mean", "TLI_SD", "TLI_NumComputations",
            "AIC_Mean", "AIC_SD", "AIC_NumComputations",
            "BIC_Mean", "BIC_SD", "BIC_NumComputations",
            "aBIC_Mean", "aBIC_SD", "aBIC_NumComputations",
            "RMSEA_Mean", "RMSEA_SD", "RMSEA_NumComputations",
            "SRMR.Within_Mean", "SRMR.Within_SD",
            "SRMR.Within_NumComputations", "SRMR.Between_Mean",
            "SRMR.Between_SD", "SRMR.Between_NumComputations")


model2_fit_g8 <- as_tibble(matrix(ncol
                  = length(model2_fit_indices),nrow =46))
colnames(model2_fit_g8) <- model2_fit_indices

model2_fit_g8[,1] <- model2_results_g8 %>%
map("summaries") %>%
map_chr("Filename")

for (i in 2:length(model2_fit_indices)) {
  x <- model2_fit_indices[i]
  x <- sym(x) #a symbol is an expression.
             It is printed without quotes.
  model2_fit_g8[,i] <- model2_results_g8 %>%
map("summaries") %>%
map_dbl(as_string(x))
}

# order the Fit statistics tibble according
  to the country index
countryi <- as.numeric(gsub("country([0-9]+).*$","\\1",
                        model2_fit_g8$Filename))
model2_fit_g8 <- model2_fit_g8[order(countryi),]
model2_fit_g8 <- as_tibble(cbind(data_g8[[2]],model2_fit_g8))

#model2_results_g8[[1]]$parameters$unstandardized

para_unstandardized <- model2_results_g8 %>%
map("parameters") %>%
map("unstandardized")

bflpe_g8 <- para_unstandardized%>%
map_dbl(~.[28,]$est)

bflpe_g8_pval <- para_unstandardized %>%
map_dbl(~.[28,]$pval)

bflpe_g8 <- bflpe_g8[order(countryi)]
bflpe_g8_pval <- bflpe_g8_pval[order(countryi)]

(model2_table_g8 <- as_tibble(cbind(model2_fit_g8,
                    "BFLPE" = bflpe_g8, "p" = bflpe_g8_pval)))
```

RESULTS

Multilevel CFA Model (Model 1)

As explained in the Syntax section, Model 1 did not run successfully for Saudi Arabia (country47) in fourth grade and Jordan (country22) in eighth grade. In both cases, Mplus output messages that the Fisher information matrix is non-positive definite and this could be due to issues with starting values for the model parameters. Non-positive definite matrices cause problems in parameter estimation of latent variable modeling (i.e., Heywood cases; see Kolenikov and Bollen, 2012), indicate lack of model fit, and could be the result of model misspecification, empirical under-identification, sampling fluctuations, or even outliers (Bollen, 1987). For parameter estimation of multilevel CFA modeling with the maximum likelihood estimation with robust standard errors (MLR), by default, Mplus uses fixed starting values. These fixed starting values could lead to non-convergence of parameter estimation. For the two problematic cases (Saudi Arabia in fourth grade and Jordan in eighth grade) of Model 1, I manually modified the Mplus input files to use 10 random sets of starting values to address the issue of non-convergence of the fixed starting value run. After the modifications, the model estimation terminated normally, although there was a warning messaging of a non-positive definite covariance matrix for the latent variables.

In addition, after examining the summary results, I decided that the model for Saudi Arabia in eighth grade needed further attention because the estimates for the within-level variance and the between-level variance were both zero. This could suggest a problem with parameter estimation and changing starting values *may* solve the problem. I manually modified the Mplus input file to use 10 random sets of starting values. After the modification, the results were more trustworthy (see **Table 1**, row "Saudi Arabia*").

Model 1 has two degrees of freedom. The model fit indices are in **Table 1** for fourth grade and in **Table 2** for eighth grade. Based on the regular model fit cutoffs (root mean square error of approximation, or RMSEA < 0.08; comparative fit index, or CFI > 0.95, Tucker–Lewis index, or TLI > 0.95) (Browne and Cudeck, 1993; Hu and Bentler, 1999; Kline, 2016), the model did not fit data from four (Czechia, France, Italy, and Saudi Arabia) of the 56 countries in fourth grade. However, only Italy had relatively poor model fit (CFI = 0.93, TLI = 0.79, RMSEA = 0.108). Czechia, France, and Saudi Arabia had relatively low TLI (0.93, 0.91, and 0.90, respectively) but their CFI values are greater than 0.95 and RMSEA values less than 0.08. In eighth grade, two (England and Saudi Arabia) out of the 46 countries did not have good model fit. Nevertheless, although both countries had relatively low TLI values (0.93 and 0.94, respectively), their CFI (0.98 for both countries) and RMSEA values (0.075 and 0.037, respectively) indicated adequate model fit.

Tables 1, **2** also include the within-level variance, the between-level variance, and the intraclass correlation coefficient (ICC) of mathematics self-concept for each country. For the three datasets (fourth grade Saudi Arabia, eighth grade Saudi Arabia, and eighth grade Jordan) that needed random sets of starting values, the estimates of the within-level variance and the between-level variance were negative. For the other datasets, in fourth grade, the within-level variance was statistically significant at the 0.05 level for all countries except Jordan (p = 0.326) and Oman (p = 0.777); the between-level variance was statistically significant at the 0.05 level for all countries except Australia (p = 0.071), Czechia (p = 0.090), Ireland (p = 0.088), and Netherlands (p = 0.430), as well as for Jordan (p = 0.379), and Oman (p = 0.781); the ICC ranged from 0.4% (Netherlands) to 27.7% (Indonesia). In eighth grade, the within-level variance was statistically significant at the 0.05 level for all countries except Bahrain (p = 0.166), Egypt (p = 0.279), and Morocco (p = 0.472); the same three countries had statistically non-significant between-level variance (p-Values were 0.194, 0.330, and 0.479 for Bahrain, Egypt, and Morocco, respectively); the ICC ranged from 1.8% (Japan) to 23.9% (England). A small ICC means that there is little between class variation compared to within class. However, a small ICC of mathematics self-concept can also be viewed as resulting from social comparison largely within the class.

Model 2 (The BFLPE Model)

Model 2 has nine degrees of freedom. **Tables 3**, **4** show modeling results for fourth grade and eighth grade, respectively. For each model fit index, there is a mean and a standard deviation. This is because the model for each country in each grade was actually run five times using the five plausible values of mathematics achievement in the TIMSS 2015 database and therefore there were five values for each model fit index. The mean and standard deviation of the five values were reported in the Mplus output. For example, the mean of CFI values for Abu Dhabi, United Arab Emirates in eighth grade was 0.93 with a standard deviation of 0.005. Using regular model fit cutoffs of CFI > 0.95 and TLI > 0.95 for the mean, most countries did not have adequate model fit. Using RMSEA < 0.08 for the mean, 52 out of 56 countries had good model fit in fourth grade and 43 out of 46 countries had good model fit in eighth grade.

In fourth grade, the BFLPE was negative and statistically significant at the 0.05 level in all but five countries (Indonesia, Jordan, Kuwait, Oman, and Saudi Arabia), ranging from −0.124 (Norway) to −1.167 (South Africa) with a mean of −0.461 and a median of −0.447 measured as the Cohen's d. In eighth grade, the BFLPE was negative and statistically significant at the 0.05 level in all but two countries (Oman and Saudi Arabia), ranging from −0.161 (Egypt) to −1.317 (Singapore) with a mean of −0.576 and a median of −0.553. The model fit indices in general were not as good as those for Model 1. It is interesting to see that the countries where the BFLPE did not manifest tended to have some of the worst model fit.

DISCUSSION

Large-scale assessments such as those available from NCES are rich data sources for researchers to study substantive research questions. One particular challenge for using such data is due to their sizes. The researcher needs to navigate various documents and datasets to identify variables and information that are useful and has to be good at data wrangling. When the analysis has to be scaled up for many groups (e.g., states, countries, regions), manually running analysis for individual groups is tedious and should be avoided. Data science tools can be particularly useful because they can automate repeated actions.

In this study, I showed how to use three R packages, EdSurvey, MplusAutomation, and tidyverse to conduct a large-scale analysis of the BFLPE across countries. Mainly, the EdSurvey was used to obtain data, MplusAutomation was used to run complex multilevel latent variable models and to extract results from Mplus outputs, and tidyverse was used for data management. Although each of the three packages is quite useful in its own way, the combination of them is a powerful toolkit for applied quantitative researchers interested in using NCES data. With these few packages learned, a researcher can do most data wrangling and analysis of LSAs.

Other R packages have been developed that may also be useful for researchers interested in analysis of LSAs. The lavaan.survey package (Oberski, 2014) combines special features of the lavaan and survey packages to allow for SEM analysis of complex survey data. However, it also has some of the same limitations as lavaan and survey. For example, missing data cannot be handled with full information maximum likelihood together with survey weights. The MplusAutomation package, because it calls and therefore has the same capacity of modeling as Mplus, can apply more advanced methods to deal with missing data, complex survey designs, and other analysis issues. It is possible to only use existing R packages without having to rely on the external Mplus software to address the missing data and other issues. For example, the semTools package (Jorgensen et al., 2020) has the runMI function that can fit a lavaan model to multiply imputed datasets or fit the lavaan model while imputing the missing values using the Amelia (Honaker et al., 2011) or the

TABLE 3 | Results of Model 2 in Fourth Grade.

	χ^2		CFI		TLI		RMSEA		BFLPE	
Country	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Est.**	***p***
Abu Dhabi, United Arab Emirates	121.49	12.74	0.77	0.012	0.59	0.022	0.050	0.003	−0.570	<0.001
Buenos Aires, Argentina	115.45	6.88	0.90	0.004	0.81	0.008	0.043	0.001	−0.439	<0.001
Dubai, United Arab Emirates	52.26	2.79	0.94	0.003	0.90	0.005	0.025	0.001	−0.462	<0.001
United Arab Emirates	218.97	6.37	0.86	0.003	0.75	0.005	0.033	0.000	−0.620	<0.001
Armenia	100.91	9.00	0.91	0.007	0.85	0.013	0.043	0.002	−0.558	<0.001
Australia	33.75	6.02	0.98	0.003	0.97	0.006	0.021	0.002	−0.527	<0.001
Belgium (Flemish)	37.60	8.84	0.99	0.003	0.98	0.005	0.024	0.004	−0.618	<0.001
Bulgaria	104.37	5.98	0.91	0.004	0.85	0.007	0.050	0.002	−0.474	<0.001
Bahrain	157.59	10.69	0.74	0.010	0.53	0.019	0.044	0.002	−0.157	0.037
Canada	76.26	6.60	0.98	0.002	0.96	0.003	0.025	0.001	−0.398	<0.001
Chile	187.95	28.67	0.88	0.013	0.79	0.024	0.064	0.005	−0.523	<0.001
ON, Canada	71.35	6.99	0.97	0.003	0.94	0.004	0.039	0.002	−0.343	<0.001
QC, Canada	9.67	2.97	1.00	0.002	1.00	0.006	0.006	0.006	−0.324	<0.001
Cyprus	174.09	15.59	0.90	0.007	0.82	0.012	0.067	0.003	−0.264	<0.001
Czechia	140.08	15.09	0.95	0.005	0.90	0.010	0.053	0.003	−0.492	<0.001
Germany	61.01	5.34	0.98	0.002	0.96	0.004	0.038	0.002	−0.447	<0.001
Denmark	62.62	9.34	0.98	0.004	0.96	0.008	0.040	0.004	−0.291	<0.001
England	49.02	4.60	0.97	0.003	0.95	0.006	0.033	0.002	−0.423	<0.001
Spain	288.50	35.02	0.84	0.014	0.72	0.025	0.063	0.004	−0.482	<0.001
Finland	116.52	11.10	0.96	0.003	0.93	0.006	0.049	0.002	−0.315	<0.001
France	39.01	1.89	0.98	0.001	0.97	0.002	0.026	0.001	−0.505	<0.001
Georgia	209.74	22.49	0.70	0.019	0.46	0.034	0.075	0.004	−0.253	0.022
Hong Kong SAR	39.42	4.93	0.97	0.004	0.95	0.007	0.031	0.003	−0.602	<0.001
Croatia	118.68	5.83	0.94	0.003	0.90	0.005	0.055	0.001	−0.402	<0.001
Hungary	349.04	23.86	0.88	0.006	0.79	0.010	0.087	0.003	−0.717	<0.001
Indonesia	250.71	23.81	0.71	0.016	0.48	0.028	0.057	0.003	−0.136	0.193
Ireland	113.65	18.64	0.94	0.009	0.90	0.016	0.052	0.005	−0.340	<0.001
Iran, Islamic Rep. of	551.97	45.96	0.54	0.021	0.18	0.037	0.087	0.004	−0.635	<0.001
Italy	66.41	10.25	0.96	0.007	0.92	0.012	0.038	0.004	−0.508	<0.001
Jordan	98.30	10.00	0.76	0.021	0.57	0.037	0.035	0.002	0.064	0.631
Japan	47.61	5.33	0.99	0.002	0.97	0.004	0.031	0.002	−0.266	<0.001
Kazakhstan	45.19	5.16	0.95	0.007	0.91	0.013	0.029	0.002	−0.615	<0.001
Korea, Rep. of	149.84	15.52	0.95	0.004	0.92	0.008	0.058	0.003	−0.150	0.014
Kuwait	55.70	1.73	0.81	0.008	0.67	0.015	0.027	0.000	−0.130	0.213
Lithuania	79.85	13.86	0.95	0.009	0.91	0.017	0.042	0.004	−0.645	<0.001
Morocco	90.19	7.67	0.82	0.012	0.67	0.021	0.029	0.001	−0.483	<0.001
Northern Ireland	119.45	24.41	0.93	0.011	0.88	0.020	0.062	0.007	−0.334	<0.001
Netherlands	150.36	10.73	0.94	0.003	0.89	0.006	0.059	0.002	−0.444	<0.001
Norway (4th grade)	85.32	15.96	0.94	0.011	0.90	0.019	0.045	0.005	−0.124	0.049
Norway	72.50	12.62	0.96	0.006	0.93	0.011	0.040	0.004	−0.314	<0.001
New Zealand	224.77	15.70	0.89	0.005	0.81	0.008	0.062	0.002	−0.619	<0.001
Oman	109.48	7.07	0.68	0.019	0.43	0.034	0.035	0.001	−0.068	0.388
Poland	347.90	18.55	0.88	0.007	0.79	0.012	0.089	0.002	−0.340	<0.001
Portugal	108.94	13.21	0.95	0.005	0.91	0.009	0.049	0.003	−0.390	<0.001
Qatar	166.81	9.98	0.75	0.011	0.55	0.019	0.058	0.002	−0.596	<0.001
Russian Federation	69.99	9.92	0.97	0.005	0.94	0.008	0.037	0.003	−0.699	<0.001
Saudi Arabia	81.79	1.99	0.70	0.012	0.47	0.021	0.043	0.001	−0.213	0.176
Singapore	201.17	21.47	0.92	0.007	0.86	0.013	0.057	0.003	−0.288	<0.001
Serbia	135.18	19.78	0.88	0.014	0.78	0.025	0.059	0.004	−0.491	<0.001
Slovak Republic	156.41	14.81	0.91	0.004	0.84	0.008	0.053	0.003	−0.760	<0.001
Slovenia	138.55	13.50	0.93	0.006	0.88	0.011	0.057	0.003	−0.372	<0.001
Sweden	118.68	20.53	0.93	0.010	0.87	0.018	0.054	0.005	−0.333	<0.001
Turkey	162.33	16.64	0.88	0.007	0.78	0.013	0.051	0.003	−0.667	<0.001
Chinese Taipei	261.43	35.20	0.89	0.007	0.81	0.012	0.081	0.006	−0.275	<0.001
United States	151.87	15.46	0.95	0.004	0.92	0.007	0.040	0.002	−0.436	<0.001
South Africa	347.18	145.98	0.49	0.219	0.09	0.390	0.058	0.011	−1.167	<0.001

CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; BFLPE, big-fish-little-pond effect.

TABLE 4 | Results of Model 2 in Eighth Grade.

	χ^2		CFI		TLI		RMSEA		BFLPE	
Country	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***	**Est.**	***p***
Abu Dhabi, United Arab Emirates	67.40	4.80	0.93	0.005	0.88	0.008	0.037	0.002	−0.852	<0.001
Buenos Aires, Argentina	89.03	11.90	0.93	0.008	0.87	0.015	0.052	0.004	−0.586	<0.001
Dubai, United Arab Emirates	35.98	1.69	0.98	0.001	0.96	0.002	0.022	0.001	−0.696	<0.001
United Arab Emirates	107.50	2.64	0.96	0.001	0.93	0.002	0.025	0.000	−0.871	<0.001
Armenia	96.43	6.22	0.95	0.004	0.90	0.007	0.044	0.002	−0.332	<0.001
Australia	117.31	9.11	0.97	0.002	0.95	0.003	0.034	0.001	−0.569	<0.001
Bahrain	97.39	7.86	0.88	0.006	0.79	0.011	0.045	0.002	−0.410	<0.001
Botswana	460.69	39.79	0.69	0.018	0.45	0.033	0.092	0.004	−0.400	<0.001
Canada	371.04	40.56	0.94	0.005	0.90	0.008	0.068	0.004	−0.511	<0.001
Chile	199.85	18.59	0.92	0.005	0.86	0.009	0.066	0.003	−0.752	<0.001
ON, Canada	117.48	22.45	0.97	0.006	0.95	0.010	0.051	0.005	−0.376	<0.001
QC, Canada	50.93	6.64	0.98	0.002	0.97	0.003	0.034	0.003	−0.420	<0.001
Egypt	165.21	11.65	0.73	0.007	0.52	0.013	0.047	0.002	−0.161	0.044
England	45.45	3.95	0.97	0.002	0.95	0.004	0.029	0.002	−0.907	<0.001
Georgia	165.56	24.32	0.87	0.015	0.77	0.026	0.065	0.005	−0.427	<0.001
Hong Kong SAR	86.79	13.45	0.96	0.006	0.93	0.011	0.045	0.004	−0.876	<0.001
Hungary	159.88	11.51	0.96	0.002	0.92	0.004	0.058	0.002	−0.790	<0.001
Ireland	129.76	14.83	0.96	0.003	0.93	0.006	0.053	0.003	−0.491	<0.001
Iran, Islamic Rep. of	186.54	14.18	0.91	0.006	0.84	0.011	0.057	0.002	−0.582	<0.001
Israel	128.14	11.52	0.94	0.005	0.89	0.010	0.049	0.002	−0.602	<0.001
Italy	72.52	10.77	0.98	0.002	0.97	0.004	0.040	0.003	−0.552	<0.001
Jordan	201.46	8.40	0.76	0.006	0.57	0.010	0.052	0.001	−0.346	<0.001
Japan	104.93	6.71	0.97	0.001	0.95	0.003	0.047	0.002	−0.554	<0.001
Kazakhstan	83.12	8.78	0.95	0.005	0.91	0.009	0.041	0.003	−0.491	<0.001
Korea, Rep. of	159.41	22.59	0.97	0.003	0.94	0.005	0.056	0.004	−0.197	<0.001
Kuwait	77.87	8.01	0.88	0.009	0.79	0.017	0.041	0.002	−0.540	<0.001
Lebanon	66.35	7.66	0.90	0.012	0.82	0.021	0.040	0.003	−0.381	<0.001
Lithuania	50.63	3.43	0.98	0.001	0.97	0.002	0.033	0.001	−0.477	<0.001
Morocco	297.04	14.03	0.80	0.006	0.65	0.010	0.050	0.001	−0.288	<0.001
Malta	21.56	1.25	0.99	0.001	0.98	0.002	0.019	0.001	−0.756	<0.001
Malaysia	584.05	58.27	0.68	0.018	0.43	0.032	0.081	0.004	−0.756	<0.001
Norway (8th grade)	134.86	20.14	0.97	0.005	0.94	0.009	0.054	0.004	−0.319	<0.001
Norway	46.93	7.15	0.99	0.002	0.98	0.003	0.030	0.003	−0.265	<0.001
New Zealand	111.00	7.18	0.96	0.002	0.93	0.004	0.037	0.001	−0.871	<0.001
Oman	142.96	5.95	0.86	0.004	0.75	0.008	0.041	0.001	−0.122	0.056
Qatar	87.93	1.93	0.90	0.004	0.83	0.006	0.040	0.000	−0.663	<0.001
Russian Federation	53.84	3.52	0.98	0.001	0.97	0.002	0.032	0.001	−0.682	<0.001
Saudi Arabia	114.39	7.42	0.83	0.010	0.69	0.017	0.056	0.002	−0.166	0.064
Singapore	34.34	4.32	0.99	0.001	0.99	0.002	0.021	0.002	−1.317	<0.001
Slovenia	30.38	5.69	0.99	0.002	0.99	0.003	0.023	0.003	−0.395	<0.001
Sweden	112.72	14.63	0.97	0.004	0.94	0.007	0.053	0.004	−0.482	<0.001
Thailand	158.42	7.17	0.87	0.005	0.77	0.008	0.051	0.001	−0.717	<0.001
Turkey	122.98	2.32	0.96	0.001	0.93	0.002	0.046	0.000	−0.670	<0.001
Chinese Taipei	585.72	21.32	0.92	0.002	0.86	0.003	0.106	0.002	−0.404	<0.001
United States	98.12	7.25	0.98	0.001	0.97	0.002	0.031	0.001	−0.568	<0.001
South Africa	70.93	3.56	0.95	0.002	0.91	0.003	0.023	0.001	−1.046	<0.001

CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error of approximation; BFLPE, big-fish-little-pond effect.

mice (van Buuren and Groothuis-Oudshoorn, 2011) package. For more experienced R users, exploring various packages for specific data and analysis issues may be a joyful learning journey. However, for less experienced users who are interested in applying latent variable modeling to large-scale educational assessment data, I recommend spending time to get familiar with the three packages discussed in this study: EdSurvey, MplusAutomation, and tidyverse.

The BFLPE was found in 51 of the 56 countries in fourth grade and 44 of the 46 countries in eighth grade for the subject of mathematics, suggesting generalizability of the effect. Earlier work using TIMSS (Wang, 2015; Wang and Bergin, 2017) and PISA (Marsh and Hau, 2003; Seaton et al., 2009) also showed the existence of the BFLPE in many countries. While the theoretical explanation of the BFLPE is social comparison, it is not clear how students compare. Huguet et al. (2009) argued that forced upward social comparison with the entire class underlies the BFLPE and found that controlling for perceived relative standing would eliminate the BFLPE; however, Wang (2015); Wang and Bergin (2017) found that students' perceived relative standing in the class did not eliminate but instead may decrease the BFLPE. All of these studies used cross-sectional data. In fact, the majority of existing BFLPE research has used cross-sectional, self-reported data. There is a need for future research based on alternative data types and formats of data collection.

Interestingly, Oman and Saudi Arabia did not have statistically significant BFLPE in both grades. Four (Cyprus, Algeria, Morocco, and Slovenia) of the 49 countries in Wang (2015) and one (Syrian Arab Republic) of the 59 countries in Wang and Bergin (2017) did not show statistically significant BFLPE for eighth-grade mathematics. It is not clear why these countries differed from other countries. It could be due to their education or social systems but a closer look at these countries may shed light on BFLPE research.

While the research on model fit of SEM is still quite active, there is little research on how these fit indices behave in large-scale analysis of complex survey data. In this project, I used traditional cutoffs for model fit indices that were developed based on single-level analysis and with the maximum likelihood estimator. The multilevel CFA model seemed to fit the data well in most of the countries, but the BFLPE model fit rather poorly in the majority of the countries. For the BFLPE model, the measurement model at both levels is saturated and constrained to have cross-level measurement invariance as in the multilevel CFA model. If the model fit indices are to be trusted, the poor fitting of the BFLPE model could be due to: (a) unmodeled relationships between the residuals of the self-concept indicator items and mathematics achievement, (b) the orthogonality assumption across levels, or (c) both (a) and (b). In SEM, it is typically not advised to include covariances between a predictor and the residuals of indicators of an outcome variable. The orthogonality assumption across levels is not a testable assumption for latent variables (mathematics self-concept in this project). Another possibility is that the relationship between mathematics achievement and mathematics self-concept could be reciprocal. While the BFLPE research uses achievement as the predictor and self-concept as the outcome, one's self-concept could likely affect achievement.

Despite the large sample sizes, the structure of data used in this project is "simple" and data collection was through surveys only. The data are well organized and the unit of analysis is students. Data management is necessary for statistical modeling but could be done using techniques that are designed for traditional data analysis. A related concept is "big data," which is a broader concept and the massive amount of data may be unstructured and in different formats such as texts, speeches, and photographs. From the "big data" standpoint, the data used for this study are "small data" – data that can be represented in spreadsheets on a single computer (Chen and Wojcik, 2016). In this study, I used many "small data" files, therefore, the term "large-scale."

The use of technology allows the collection of behavior data that were not possible before. For example, the 2017 NAEP was administered for the first time as a digitally based assessment. Response process data were collected that could provide insights into students' test-taking behaviors, how such behaviors relate to achievement, and even diagnostics of learning strategies. Other types of data such as videos, texts, online social network data (e.g., Twitter and Facebook) are additional examples. Researchers in psychology and other social sciences can take advantage of these more "novel" data types with the use of data science and big data tools (Chen and Wojcik, 2016).

This study shows that the analysis of many similarly structured datasets can be automated using data science tools. However, the researcher still needs to scrutinize modeling results to identify possible problems. Any result that looks suspicious should be examined more closely. For this project, I found that the initial results for Saudi Arabia in eighth grade could be problematic due to the estimates of variances of latent variables. Because the model did run and fit information could be extracted, I might have trusted the initial results. However, a closer look rendered that the initial model had a problem with starting values. The convenience of data science tools should not be substituted for content expertise.

This study has a didactic nature and focuses on analysis of LSAs. The field of data science and big data has begun to attract more researchers in social sciences (Gilmore, 2016); there is a high demand of tutorials showing "how to" use various data tools. Some tools are more general for writing purposes. For example, R Markdown is a powerful tool to create fully reproducible documents, combining code, results, explanatory texts, tables, references, etc. Other tools, such as those used in this study, are for more specific purposes. Teaching researchers how to use these tools can be a particularly useful area in its own right. We need "twofers" who can help bridge data engineering and domain knowledge to move both worlds forward.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://timssandpirls.bc.edu/timss2015/international-database.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation

and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

ZW conceived the study, performed the statistical analysis, and wrote the manuscript.

REFERENCES

Bache, S. M., and Wickham, H. (2014). *magrittr: A Forward-Pipe Operator for R. R package version 1.5*.

Bailey, P., Emad, A., Huo, H., Lee, M., Liao, Y., Lishinski, A., et al. (2020). *EdSurvey: Analysis of NCES Education Survey and Assessment Data. R package version 2.5.0*.

Barrett, T. S. (2019). *Six Reasons to Consider Using R in Psychological Research*. Available online at: https://psyarxiv.com/8mb6d/ (accessed August 14, 2020).

Bollen, K. A. (1987). Outliers and improper solutions: a confirmatory factor analysis example. *Sociol. Methods Res.* 15, 375–384. doi: 10.1177/0049124187015004002

Browne, M. W., and Cudeck, R. (1993). "Alternative ways of assessing model fit," in *Testing Structural Equation Models*, eds K. A. Bollen and J. S. Long (Thousand Oaks, CA: Sage), 136–162.

Chen, E. E., and Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychol. Methods* 21, 458–474. doi: 10.1037/met0000111

Dowle, M., and Srinivasan, A. (2019). *Data.Table: Extension of 'Data.Frame'. R Package Version 1.12.8*.

Enders, C. K. (2010). *Applied Missing Data Analysis*. Los Angeles, CA: Guilford.

Festinger, L. (1954). A theory of social comparison processes. *Hum. Relat.* 7, 117–140. doi: 10.1177/001872675400700202

Foy, P. (2017). *TIMSS 2015 User Guide for the International Database*. Boston, MA: TIMSS & PIRLS International Study Center, Lynch School of Education.

Gilmore, R. O. (2016). From big data to deep insight in developmental science. *WIREs Cogn. Sci.* 7, 112–126. doi: 10.1002/wcs.1379

Grolemund, G., and Wickham, H. (2018). *R for Data Science*. Sebastopol, CA: O'Reilly Media, Inc.

Hallquist, M. N., and Wiley, J. F. (2018). mplusautomation: an R package for facilitating large-scale latent variable analyses in Mplus. *Struct. Equ. Model.* 25, 621–638. doi: 10.1080/10705511.2017.1402334

Henry, L., and Wickham, H. (2020a). *purrr: Functional Programming Tools. R package version 0.3.4*.

Henry, L., and Wickham, H. (2020b). *rlang: Functions for Base Types and Core R and 'Tidyverse' Features. R package version 0.4.6*.

Honaker, J., King, G., and Blackwell, M. (2011). Amelia II: a program for missing data. *J. Stat. Softw.* 45:47. doi: 10.18637/jss.v045.i07

Hu, L.-T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Huguet, P., Dumas, F., Marsh, H., Régner, I., Wheeler, L., Suls, J., et al. (2009). Clarifying the role of social comparison in the big-fish–little-pond effect (BFLPE): an integrative study. *J. Pers. Soc. Psychol.* 97, 156–170. doi: 10.1037/a0015558

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., and Rosseel, Y. (2020). *semTools: Useful Tools for Structural Equation Modeling. . R package version 0.5-3*.

Kline, R. B. (2016). *Principles and Practice of Structural Equation Modeling*, 4th Edn. Los Angeles, CA: Guilford.

Kolenikov, S., and Bollen, K. A. (2012). Testing negative error variances: is a Heywood case a symptom of misspecification? *Sociol. Methods Res.* 41, 124–167. doi: 10.1177/0049124112442138

Lohr, S. (2014). *For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights*. New York, NY: The New York Times.

Marsh, H. W. (1990). A multidimensional, hierarchical model of self-concept: theoretical and empirical justification. *Educ. Psychol. Rev.* 2, 77–172. doi: 10.1007/bf01322177

Marsh, H. W., Abduljabbar, A. S., Morin, A. J. S., Parker, P., Abdelfattah, F., Nagengast, B., et al. (2015). The big-fish-little-pond effect: generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries. *J. Educ. Psychol.* 107, 258–271. doi: 10.1037/a0037485

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., and Nagengast, B. (2014). The Big-Fish-Little-Pond Effect in mathematics: a Cross-cultural comparison of U.S. and Saudi Arabian TIMSS responses. *J. Cross Cult. Psychol.* 45, 777–804. doi: 10.1177/0022022113519858

Marsh, H. W., and Hau, K.-T. (2003). Big-Fish–Little-Pond effect on academic self-concept: a cross-cultural (26-country) test of the negative effects of academically selective schools. *Am. Psychol.* 58, 364–376. doi: 10.1037/0003-066x.58.5.364

Muenchen, R. A. (n.d.). *The Popularity of Data Science Software*. Vienna: R Language Training & Data Science Market Share Analysis.

Muthén, L. K., and Muthén, B. O. (1998-2017). *Mplus User's Guide*, 8th Edn. North Yorkshire: Muthén & Muthén.

Nagengast, B., and Marsh, H. W. (2011). The negative effect of school-average ability on science self-concept in the UK, the UK countries and the world: the Big-Fish-Little-Pond-Effect for PISA 2006. *Educ. Psychol.* 31, 629–656. doi: 10.1080/01443410.2011.586416

National Center for Education Statistics (2020). *History and Innovation*. Available online at: https://nces.ed.gov/nationsreportcard/about/timeline.aspx (accessed August 14, 2020).

National Center for Education Statistics (n.d.). *International Activities Program*. Available online at: https://nces.ed.gov/surveys/international/index.asp (accessed August 14, 2020).

Oberski, D. (2014). lavaan.survey: an r package for complex survey analysis of structural equation models. *J. Stat. Softw.* 57:27. doi: 10.18637/jss.v057.i01

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Raudenbush, S. W., and Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2nd Edn. Thousand Oaks, CA: Sage.

Rogers, C. M., Smith, M. D., and Coleman, J. M. (1978). Social comparison in the classroom: the relationship between academic achievement and self-concept. *J. Educ. Psychol.* 70, 50–57. doi: 10.1037/0022-0663.70.1.50

Rosseel, Y. (2012). lavaan: an R Package for structural equation modeling. *J. Stat. Softw.* 48:36. doi: 10.18637/jss.v048.i02

Rutkowski, L., von Davier, M., and Rutkowski, D. (eds) (2014). *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Seaton, M., Marsh, H. W., and Craven, R. G. (2009). Earning its place as a pan-human theory: Universality of the big-fish-little-pond effect across 41 culturally and economically diverse countries. *J. Educ. Psychol.* 101, 403–419. doi: 10.1037/a0013838

Smith, T. J., Walker, D. A., Chen, H.-T., and Hong, Z.-R. (2020). Students' sense of school belonging and attitude towards science: a cross-cultural examination. *Int. J. Sci. Math. Educ.* 18, 855–867. doi: 10.1007/s10763-019-10002-7

Suls, J., Martin, R., and Wheeler, L. (2002). Social comparison: why, with whom, and with what effect? *Curr. Direct. Psychol. Sci.* 11, 159–163. doi: 10.1111/1467-8721.00191

van Buuren, S., and Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *J. Stat. Softw.* 45:67. doi: 10.18637/jss.v045.i03

Wagemaker, H. (2014). "International large-scale assessments: from research to policy," in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, eds L. Rutkowski, M. von Davier, and D. Rutkowski (Milton Park: Taylor & Francis), 11–36.

Wang, Z. (2015). Examining big-fish-little-pond-effects across 49 countries: a multilevel latent variable modelling approach. *Educ. Psychol.* 35, 228–251. doi: 10.1080/01443410.2013.827155

Wang, Z. (2017). Editorial: large-scale educational assessments. *Int. J. Quant. Res. Educ.* 4, 1–2. doi: 10.1007/978-94-007-4629-9_1

Wang, Z., and Bergin, D. A. (2017). Perceived relative standing and the big-fish-little-pond effect in 59 countries and regions: analysis of TIMSS 2011 data. *Learn. Individ. Diff.* 57, 141–156. doi: 10.1016/j.lindif.2017.04.003

Wang, Z., Osterlind, S. J., and Bergin, D. A. (2012). Building mathematics achievement models in four countries using TIMSS 2003. *Int. J. Sci. Math. Educ.* 10, 1215–1242. doi: 10.1007/s10763-011-9328-6

Wickham, H. (2019). *Advanced R*. Boca Raton, FL: CRC Press.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D. A., François, R., et al. (2019). Welcome to the tidyverse. *J. Open Source Softw.* 4:1686. doi: 10.21105/joss.01686

Wu, M. (2005). The role of plausible values in large-scale surveys. *Stud. Educ. Eval.* 31, 114–128. doi: 10.1016/j.stueduc.2005.05.005

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

REVIEW
published: 19 October 2020
doi: 10.3389/fpsyg.2020.580820

Challenges and Future Directions of Big Data and Artificial Intelligence in Education

*Hui Luan[1], Peter Geczy[2], Hollis Lai[3], Janice Gobert[4,5], Stephen J. H. Yang[6], Hiroaki Ogata[7], Jacky Baltes[8], Rodrigo Guerra[9], Ping Li[10] and Chin-Chung Tsai[1,11]**

[1] Institute for Research Excellence in Learning Sciences, National Taiwan Normal University, Taipei, Taiwan, [2] National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan, [3] School of Dentistry, Faculty of Medicine & Dentistry, University of Alberta, Edmonton, AB, Canada, [4] Graduate School of Education, Rutgers – The State University of New Jersey, New Brunswick, NJ, United States, [5] Apprendis, LLC, Berlin, MA, United States, [6] Department of Computer Science and Information Engineering, College of Electrical Engineering and Computer Science, National Central University, Taoyuan City, Taiwan, [7] Graduate School of Informatics, Kyoto University, Kyoto, Japan, [8] Department of Electrical Engineering, College of Technology and Engineering, National Taiwan Normal University, Taipei, Taiwan, [9] Centro de Tecnologia, Universidade Federal de Santa Maria, Santa Maria, Brazil, [10] Department of Chinese and Bilingual Studies, Faculty of Humanities, The Hong Kong Polytechnic University, Kowloon, Hong Kong, [11] Program of Learning Sciences, National Taiwan Normal University, Taipei, Taiwan

OPEN ACCESS

Edited by:
Ronnel B. King,
University of Macau, China

Reviewed by:
Hannele Niemi,
University of Helsinki, Finland
Ze Wang,
University of Missouri, United States

****Correspondence:***
Chin-Chung Tsai
tsaicc@ntnu.edu.tw

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *07 July 2020*
Accepted: *22 September 2020*
Published: *19 October 2020*

Citation:
Luan H, Geczy P, Lai H, Gobert J, Yang SJH, Ogata H, Baltes J, Guerra R, Li P and Tsai C-C (2020) Challenges and Future Directions of Big Data and Artificial Intelligence in Education. Front. Psychol. 11:580820. doi: 10.3389/fpsyg.2020.580820

We discuss the new challenges and directions facing the use of big data and artificial intelligence (AI) in education research, policy-making, and industry. In recent years, applications of big data and AI in education have made significant headways. This highlights a novel trend in leading-edge educational research. The convenience and embeddedness of data collection within educational technologies, paired with computational techniques have made the analyses of big data a reality. We are moving beyond proof-of-concept demonstrations and applications of techniques, and are beginning to see substantial adoption in many areas of education. The key research trends in the domains of big data and AI are associated with assessment, individualized learning, and precision education. Model-driven data analytics approaches will grow quickly to guide the development, interpretation, and validation of the algorithms. However, conclusions from educational analytics should, of course, be applied with caution. At the education policy level, the government should be devoted to supporting lifelong learning, offering teacher education programs, and protecting personal data. With regard to the education industry, reciprocal and mutually beneficial relationships should be developed in order to enhance academia-industry collaboration. Furthermore, it is important to make sure that technologies are guided by relevant theoretical frameworks and are empirically tested. Lastly, in this paper we advocate an in-depth dialog between supporters of "cold" technology and "warm" humanity so that it can lead to greater understanding among teachers and students about how technology, and specifically, the big data explosion and AI revolution can bring new opportunities (and challenges) that can be best leveraged for pedagogical practices and learning.

Keywords: big data, artificial intelligence, education, learning, teaching

INTRODUCTION

The purpose of this position paper is to present current status, opportunities, and challenges of big data and AI in education. The work has originated from the opinions and panel discussion minutes of an international conference on big data and AI in education (The International Learning Sciences Forum, 2019), where prominent researchers and experts from different disciplines such as education, psychology, data science, AI, and cognitive neuroscience, etc., exchanged their knowledge and ideas. This article is organized as follows: we start with an overview of recent progress of big data and AI in education. Then we present the major challenges and emerging trends. Finally, based on our discussions of big data and AI in education, conclusion and future scope are suggested.

Rapid advancements in big data and artificial intelligence (AI) technologies have had a profound impact on all areas of human society including the economy, politics, science, and education. Thanks in large part to these developments, we are able to continue many of our social activities under the COVID-19 pandemic. Digital tools, platforms, applications, and the communications among people have generated vast amounts of data ('big data') across disparate locations. Big data technologies aim at harnessing the power of extensive data in real-time or otherwise (Daniel, 2019). The characteristic attributes of big data are often referred to as the four V's. That is, volume (amount of data), variety (diversity of sources and types of data), velocity (speed of data transmission and generation), and veracity (the accuracy and trustworthiness of data) (Laney, 2001; Schroeck et al., 2012; Geczy, 2014). Recently, a 5th V was added, namely value (i.e., that data could be monetized; Dijcks, 2013). Because of intrinsic big data characteristics (the five Vs), large and complex datasets are impossible to process and utilize by using traditional data management techniques. Hence, novel and innovative computational technologies are required for the acquisition, storage, distribution, analysis, and management of big data (Lazer et al., 2014; Geczy, 2015). Big data analytics commonly encompasses the processes of gathering, analyzing, and evaluating large datasets. Extraction of actionable knowledge and viable patterns from data are often viewed as the core benefits of the big data revolution (Mayer-Schönberger and Cukier, 2013; Jagadish et al., 2014). Big data analytics employ a variety of technologies and tools, such as statistical analysis, data mining, data visualization, text analytics, social network analysis, signal processing, and machine learning (Chen and Zhang, 2014).

As a subset of AI, machine learning focuses on building computer systems that can learn from and adapt to data automatically without explicit programming (Jordan and Mitchell, 2015). Machine learning algorithms can provide new insights, predictions, and solutions to customize the needs and circumstances of each individual. With the availability of large quantity and high-quality input training data, machine learning processes can achieve accurate results and facilitate informed decision making (Manyika et al., 2011; Gobert et al., 2012, 2013; Gobert and Sao Pedro, 2017). These data-intensive, machine learning methods are positioned at the intersection of big data and AI, and are capable of improving the services and productivity of education, as well as many other fields including commerce, science, and government.

Regarding education, our main area of interest here, the application of AI technologies can be traced back to approximately 50 years ago. The first Intelligent Tutoring System "SCHOLAR" was designed to support geography learning, and was capable of generating interactive responses to student statements (Carbonell, 1970). While the amount of data was relatively small at that time, it was comparable to the amount of data collected in other traditional educational and psychological studies. Research on AI in education over the past few decades has been dedicated to advancing intelligent computing technologies such as intelligent tutoring systems (Graesser et al., 2005; Gobert et al., 2013; Nye, 2015), robotic systems (Toh et al., 2016; Anwar et al., 2019), and chatbots (Smutny and Schreiberova, 2020). With the breakthroughs in information technologies in the last decade, educational psychologists have had greater access to big data. Concretely speaking, social media (e.g., Facebook, Twitter), online learning environments [e.g., Massive Open Online Courses (MOOCs)], intelligent tutoring systems (e.g., AutoTutor), learning management systems (LMSs), sensors, and mobile devices are generating ever-growing amounts of dynamic and complex data containing students' personal records, physiological data, learning logs and activities, as well as their learning performance and outcomes (Daniel, 2015). Learning analytics, described as "the measurement, collection, analysis, and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs" (Long and Siemens, 2011, p. 34), are often implemented to analyze these huge amounts of data (Aldowah et al., 2019). Machine learning and AI techniques further expand the capabilities of learning analytics (Zawacki-Richter et al., 2019). The essential information extracted from big data could be utilized to optimize learning, teaching, and administration (Daniel, 2015). Hence, research on big data and AI is gaining increasing significance in education (Johnson et al., 2011; Becker et al., 2017; Hwang et al., 2018) and psychology (Harlow and Oswald, 2016; Yarkoni and Westfall, 2017; Adjerid and Kelley, 2018; Cheung and Jak, 2018). Recently, the adoption of big data and AI in the psychology of learning and teaching has been trending as a novel method in cutting-edge educational research (Daniel, 2015; Starcic, 2019).

THE POSITION FORMULATION

A growing body of literature has attempted to uncover the value of big data at different education levels, from preschool to higher education (Chen N.-S. et al., 2020). Several journal articles and book chapters have presented retrospective descriptions and the latest advances in the rapidly expanding research area from different angles, including systematic literature review (Zawacki-Richter et al., 2019; Quadir et al., 2020), bibliometric study (Hinojo-Lucena et al., 2019), qualitative

analysis (Malik et al., 2019; Chen L. et al., 2020), and social network analysis (Goksel and Bozkurt, 2019). More details can be found in the previously mentioned reviews. In this paper, we aim at presenting the current progress of the application of big data and AI in education. By and large, the research on the learner side is devoted to identifying students' learning and affective behavior patterns and profiles, improving methods of assessment and evaluation, predicting individual students' learning performance or dropouts, and providing adaptive systems for personalized support (Papamitsiou and Economides, 2014; Zawacki-Richter et al., 2019). On the teacher side, numerous studies have attempted to enhance course planning and curriculum development, evaluation of teaching, and teaching support (Zawacki-Richter et al., 2019; Quadir et al., 2020). Additionally, teacher dashboards, such as Inq-Blotter, driven by big data techniques are being used to inform teachers' instruction in real time while students simultaneously work in Inq-ITS (Gobert and Sao Pedro, 2017; Mislevy et al., 2020). Big data technologies employing learning analytics and machine learning have demonstrated high predictive accuracy of students' academic performance (Huang et al., 2020). Only a small number of studies have focused on the effectiveness of learning analytics programs and AI applications. However, recent findings have revealed encouraging results in terms of improving students' academic performance and retention, as well as supporting teachers in learning design and teaching strategy refinement (Viberg et al., 2018; Li et al., 2019; Sonderlund et al., 2019; Mislevy et al., 2020).

Despite the growing number of reports and methods outlining implementations of big data and AI technologies in educational environments, we see a notable gap between contemporary technological capabilities and their utilization for education. The fast-growing education industry has developed numerous data processing techniques and AI applications, which may not be guided by current theoretical frameworks and research findings from psychology of learning and teaching. The rapid pace of technological progress and relatively slow educational adoption have contributed to the widening gap between technology readiness and its application in education (Macfadyen, 2017). There is a pressing need to reduce this gap and stimulate technological adoption in education. This work presents varying viewpoints and their controversial issues, contemporary research, and prospective future developments in adoption of big data and AI in education. We advocate an interdisciplinary approach that encompasses educational, technological, and governmental spheres of influence. In the educational domain, there is a relative lack of knowledge and skills in AI and big data applications. On the technological side, few data scientists and AI developers are familiar with the advancements in education psychology, though this is changing with the advent of graduate programs at the intersection of Learning Sciences and Computer Science. Finally, in terms of government policies, the main challenges faced are the regulatory and ethical dilemmas between support of educational reforms and restrictions on adoptions of data-oriented technologies.

AN INTERDISCIPLINARY APPROACH TO EDUCATIONAL ADOPTION OF BIG DATA AND AI

In response to the new opportunities and challenges that the big data explosion and AI revolution are bringing, academics, educators, policy-makers, and professionals need to engage in productive collaboration. They must work together to cultivate our learners' necessary competencies and essential skills important for the 21st century work, driven by the knowledge economy (Bereiter, 2002). Collaboration across diverse disciplines and sectors is a demanding task—particularly when individual sides lack a clear vision of their mutually beneficial interests and the necessary knowledge and skills to realize that vision. We highlight several overlapping spheres of interest at the intersection of research, policy-making, and industry engagements. Researchers and the industry would benefit from targeted educational technology development and its efficient transfer to commercial products. Businesses and governments would benefit from legislature that stimulates technology markets while suitably protecting data and users' privacy. Academics and policy makers would benefit from prioritizing educational reforms enabling greater adoption of technology-enhanced curricula. The recent developments and evolving future trends at intersections between researchers, policy-makers, and industry stakeholders arising from advancements and deployments of big data and AI technologies in education are illustrated in **Figure 1**.

The constructive domains among stakeholders progressively evolve along with scientific and technological developments. Therefore, it is important to reflect on longer-term projections and challenges. The following sections highlight the novel challenges and future directions of big data and AI technologies at the intersection of education research, policy-making, and industry.

BIG DATA AND AI IN EDUCATION: RESEARCH

An understanding of individual differences is critical for developing pedagogical tools to target specific students and to tailor education to individual needs at different stages. Intelligent educational systems employing big data and AI techniques are capable of collecting accurate and rich personal data. Data analytics can reveal students' learning patterns and identify their specific needs (Gobert and Sao Pedro, 2017; Mislevy et al., 2020). Hence, big data and AI have the potential to realize individualized learning to achieve precision education (Lu et al., 2018). We see the following emerging trends, research gaps, and controversies in integrating big data and AI into education research so that there is a deep and rigorous understanding of individual differences that can be used to personalize learning in real time and at scale.

(1) Education is progressively moving from a one-size-fits-all approach to precision education or personalized learning

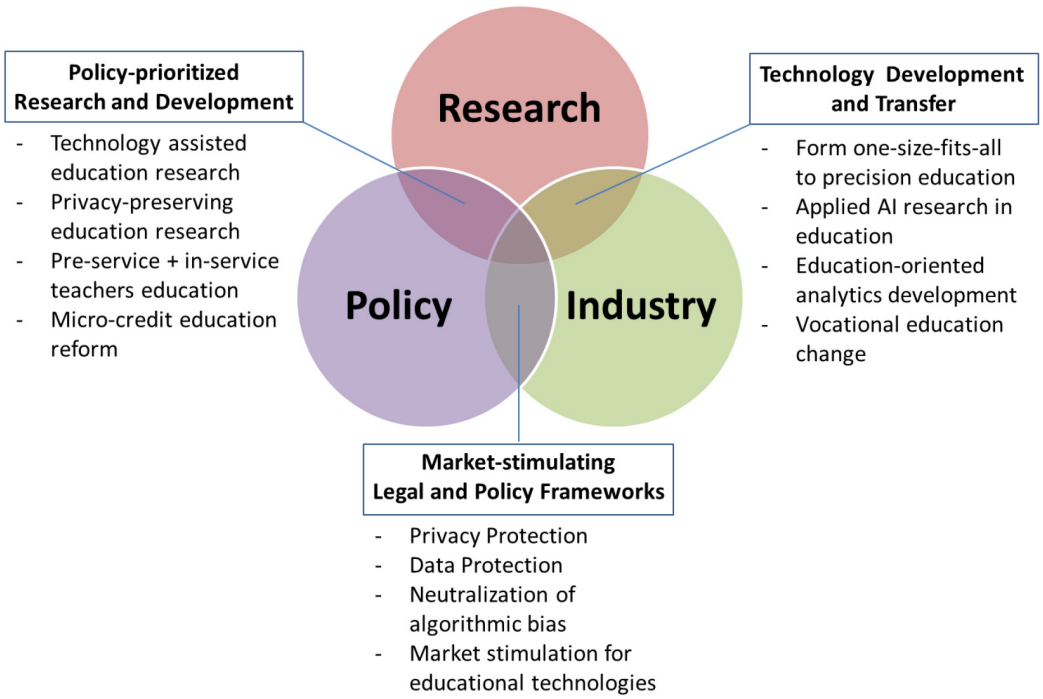

FIGURE 1 | Contemporary developments and future trends at the intersections between research, policy, and industry driven by big data and AI advances in education.

(Lu et al., 2018; Tsai et al., 2020). The one-size-fits-all approach was designed for average students, whereas precision education takes into consideration the individual differences of learners in their learning environments, along with their learning strategies. The main idea of precision education is analogous to "precision medicine," where researchers harvest big data to identify patterns relevant to specific patients such that prevention and treatment can be customized. Based on the analysis of student learning profiles and patterns, precision education predicts students' performance and provides timely interventions to optimize learning. The goal of precision education is to improve the diagnosis, prediction, treatment, and prevention of learning outcomes (Lu et al., 2018). Contemporary research gaps related to adaptive tools and personalized educational experiences are impeding the transition to precision education. Adaptive educational tools and flexible learning systems are needed to accommodate individual learners' interaction, pace, and learning progress, and to fit the specific needs of the individual learners, such as students with learning disabilities (Xie et al., 2019; Zawacki-Richter et al., 2019). Hence, as personalized learning is customized for different people, researchers are able to focus on individualized learning that is adaptive to individual needs in real time (Gobert and Sao Pedro, 2017; Lu et al., 2018).

(2) The research focus on deploying AI in education is gradually shifting from a computational focus that demonstrates use cases of new technology to cognitive focus that incorporates cognition in its design, such as perception (VanRullen, 2017), emotion (Song et al., 2016), and cognitive thinking (Bramley et al., 2017). Moreover, it is also shifting from a single domain (e.g., domain expertise, or expert systems) to a cross-disciplinary approach through collaboration (Spikol et al., 2018; Krouska et al., 2019) and domain transfers (L'heureux et al., 2017). These controversial shifts are facilitating transitions from the knowing of the unknown (gaining insights through reasoning) to the unknown of the unknown (figuring out hidden values and unknown results through algorithms) (Abed Ibrahim and Fekete, 2019; Cutumisu and Guo, 2019). In other words, deterministic learning, aimed at deductive/inductive reasoning and inference engines, predominated in traditional expert systems and old AI. Whereas, today, dynamic and stochastic learning, the outcome of which involves some randomness and uncertainty, is gradually becoming the trend in modern machine learning techniques.

(3) The format of machine-generated data and the purpose of machine learning algorithms should be carefully designed. There is a notable gap between theoretical design and its applicability. A theoretical model is needed to guide the development, interpretation, and validation of algorithms (Gobert et al., 2013; Hew et al., 2019). The outcomes of data analytics and algorithmically generated evidence must be shared with educators and applied with caution. For instance, efforts to algorithmically detect mental states such as boredom, frustration, and confusion (Baker

et al., 2010) must be supported by the operational definitions and constructs that have been prudently evaluated. Additionally, the affective data collected by AI systems should take into account the cultural differences combined with contextual factors, teachers' observations, and students' opinions (Yadegaridehkordi et al., 2019). Data need to be informatively and qualitatively balanced, in order to avoid implicit biases that may propagate into algorithms trained on such data (Staats, 2016).

(4) There are ethical and algorithmic challenges when balancing human provided learning and machine assisted learning. The significant influence of AI and contemporary technologies is a double-edged sword (Khechine and Lakhal, 2018). On the one hand, it facilitates better usability and drives progress. On the other, it might lead to the algorithmic bias and loss of certain essential skills among students who are extensively relying on technology. For instance, in creativity- or experience-based learning, technology may even become an obstacle to learning, since it may hinder students from attaining first-hand experiences and participating in the learning activities (Cuthbertson et al., 2004). Appropriately balancing the technology adoption and human involvement in various educational contexts will be a challenge in the foreseeable future. Nonetheless, the convergence of human and machine learning has the potential for highly effective teaching and learning beyond the simple "sum of the parts of human and artificial intelligence" (Topol, 2019).

(5) Algorithmic bias is another controversial issue (Obermeyer et al., 2019). Since modern AI algorithms extensively rely on data, their performance is governed solely by data. Algorithms adapt to inherent qualitative and quantitative characteristics of data. For example, if data is unbalanced and contains disproportionately better information on students from general population in comparison to minorities, the algorithms may produce systematic and repeatable errors disadvantaging minorities. These controversial issues need to be addressed before its wide implementation in education practice since every single student is precious. More rigorous studies and validation in real learning environments are required though work along these lines is being done (Sao Pedro et al., 2013).

(6) The fast expansion of technology and inequalities of learning opportunities has aroused great controversies. Due to the exponential nature of technological progress, particularly big data and AI revolution, a fresh paradigm and new learning landscape are on the horizon. For instance, the elite smartphone 10 years ago, in 2010, was BlackBerry. Today, 10 years later, even in sub-Saharan Africa, 75% of the population has mobile phones several generations more advanced (GSMA Intelligence, 2020). Hence, the entry barriers are shifting from the technical requirements to the willingness of and/or need for adoption. This has been clearly demonstrated during the COVID-19 pandemic. The need for social distancing and continuing education has led to online/e-learning deployments within months (United Nations, 2020). A huge amount of learning data is created accordingly. The extraction of meaningful patterns and the discovery of knowledge from these data is expected to be carried out through learning analytics and AI techniques. Inevitably, the current learning cultures, learning experiences, and classroom dynamics are changing as "we live algorithmic lives" (Bucher, 2018). Thus, there is a critical need to adopt proper learning theories of educational psychology and to encourage our learners to be active participants rather than passive recipients or merely tracked objects (Loftus and Madden, 2020). For example, under the constructionist framework (Tsai, 2000), the technology-enhanced or AI-powered education may empower students to know their learning activities and patterns, predict their possible learning outcomes, and strategically regulate their learning behavior (Koh et al., 2014; Loftus and Madden, 2020). On the other hand, in the era of information explosion and AI revolution, the disadvantaged students and developing countries are indeed facing a wider digital divide. To reduce the inequalities and bring more opportunities, cultivating young people's competencies is seemed like one of the most promising means (UNESCO, 2015). Meanwhile, overseas support from international organizations such as World Bank and UNESCO are imperative for developing countries in their communication infrastructure establishment (e.g., hardware, software, connectivity, electricity). Naturally, technology will not replace or hinder human learning; rather, a smart use of new technologies will facilitate transfer and acquisition of knowledge (Azevedo et al., 2019).

An overarching theme from the above trends of research is that we need theories of cognitive and educational psychology to guide our understanding of the individual learner (and individual differences), in order to develop best tools, algorithms, and practices for personalized learning. Take, for example, VR (virtual reality) or AR (augmented reality) as a fast-developing technology for education. The industry has developed many different types of VR/AR applications (e.g., *Google Expeditions* with over 100 virtual field trips), but these have typically been developed in the views of the industry (see further discussion below) and may not be informed by theories and data from educational psychology about how students actually learn. To make VR/AR effective learning tools, we must separate the technological features from the human experiences and abilities (e.g., cognitive, linguistic, spatial abilities of the learner; see Li et al., 2020). For example, VR provides a high-fidelity 3D real-life virtual environment, and the technological tools are built on the assumption that 3D realism enables the learner to gain 'perceptual grounding' during learning (e.g., having access to visual, auditory, tactile experiences as in real world). Following the 'embodied cognition' theory (Barsalou, 2008), we should expect VR learning to yield better learning outcomes compared with traditional classroom learning. However, empirical data suggest that there are significant individual differences in that some students benefit more than

others from VR learning. It may be that the individuals with higher cognitive and perceptual abilities need no additional visuospatial information (provided in VR) to succeed in learning. In any case, we need to understand how embodied experiences (provided by the technology) interact with different learners' inherent abilities (as well as their prior knowledge and background) for the best application of the relevant technology in education.

BIG DATA AND AI IN EDUCATION: POLICY-MAKING

Following the revolution triggered by breakthroughs in big data and AI technology, policy-makers have attempted to formulate strategies and policies regarding how to incorporate AI and emerging technologies into primary, secondary, and tertiary education (Pedró et al., 2019). Major challenges must be overcome in order to suitably integrate big data and AI into educational practice. The following three segments highlight pertinent policy-oriented challenges, gaps, and evolving trends.

(1) In digitally-driven knowledge economies, traditional formal education systems are undergoing drastic changes or even a paradigm shift (Peters, 2018). Lifelong learning is quickly being adopted and implemented through online or project-based learning schemes that incorporate multiple ways of teaching (Lenschow, 1998; Sharples, 2000; Field, 2001; Koper and Tattersall, 2004). This new concept of continual education will require micro-credits or micro-degrees to sustain learners' efforts (Manuel Moreno-Marcos et al., 2019). The need to change the scope and role of education will become evident in the near future (Williams, 2019). For example, in the next few years, new instruction methods, engagement, and assessment will need to be developed in formal education to support lifelong education. The system should be based on micro-credits or micro-degrees.

(2) Solutions for integrating cutting-edge research findings, innovative theory-driven curricula, and emerging technologies into students' learning are evidently beneficial, and perhaps even ready for adoption. However, there is an apparent divergence between a large number of pre-service and in-service teachers and their willingness to support and adopt these emerging technologies (Pedró et al., 2019). Pre-service teachers have greater exposure to modern technologies and, in general, are more willing to adopt them. In-service teachers have greater practical experience and tend to more rely on it. To bridge the gap, effective teacher education programs and continuing education programs have to be developed and offered to support the adoption of these new technologies so that they can be implemented with fidelity (O'Donnell, 2008). This issue could become even more pressing to tackle in light of the extended period of the COVID-19 pandemic.

(3) A suitable legislative framework is needed to protect personal data from unscrupulous collection, unauthorized disclosure, commercial exploitation, and other abuses (Boyd and Crawford, 2012; Pardo and Siemens, 2014). Education records and personal data are highly sensitive. There are significant risks associated with students' educational profiles, records, and other personal data. Appropriate security measures must be adopted by educational institutions. Commercial educational system providers are actively exploiting both legislative gaps and concealed data acquisition channels. Increasing numbers of industry players are implementing data-oriented business models (Geczy, 2018). There is a vital role to play for legislative, regulatory, and enforcing bodies at both the national and local levels. It is pertinent that governments enact, implement, and enforce privacy and personal data protection legislation and measures. In doing so, there is a need to strike a proper balance between desirable use of personal data for educational purposes and undesirable commercial monetization and abuse of personal data.

BIG DATA AND AI IN EDUCATION: INDUSTRY

As scientific and academic aspects of big data and AI in education have their unique challenges, so does the commercialization of educational tools and systems (Renz et al., 2020). Numerous countries have attempted to stimulate innovation-based growth through enhancing technology transfer and fostering academia-industry collaboration (Huggins and Thompson, 2015). In the United States, this was initiated by the Bayh-Dole Act (Mowery et al., 2001). Building a reciprocal and sustained partnership is strongly encouraged. It facilitates technology transfers and strengthens the links between academia and the education industry. There are several points to be considered when approaching academia-industry collaboration. It is important that collaboration is mutually beneficial. The following points highlight the overlapping spheres of benefits for both educational commerce and academia. They also expose existing gaps and future prospects.

(1) Commercializing intelligent educational tools and systems that include the latest scientific and technological advances can provide educators with tools for developing more effective curricula, pedagogical frameworks, assessments, and programs. Timely release of educational research advances onto commercial platforms is desirable by vendors from development, marketing, and revenue perspectives (Renz and Hilbig, 2020). Implementation of the latest research enables progressive development of commercial products and distinctive differentiation for marketing purposes. This could also potentially solve the significant gap between what the industry knows and develops and what the academic research says with regard to student learning. Novel features may also be suitably monetized—hence, expanding revenue streams. The gaps between availability of the latest research and its practical adoption are slowing progress and negatively impacting commercial vendors. A viable solution is a

closer alignment and/or direct collaboration between academia and industry.

(2) A greater spectrum of commercially and freely available tools helps maintain healthy market competition. It also helps to avoid monopolies and oligopolies that stifle innovation, limit choices, and damage markets for educational tools. Some well-stablished or free-of-charge platforms (e.g., Moodle, LMS) might show such potential of oligopolies during the COVID-19 pandemic. With more tools available on the market, educators and academics may explore novel avenues for improving education and research. New and more effective forms of education may be devised. For instance, multimodal virtual educational environments have high potential future prospects. These are environments that would otherwise be impossible in conventional physical settings (see previous discussion of VR/AR). Expanding educational markets and commerce should inevitably lead to expanding resources for research and development funding (Popenici and Kerr, 2017). Collaborative research projects sponsored by the industry should provide support and opportunities for academics to advance educational research. Controversially, in numerous geographies there is a decreasing trend in collaborative research. To reverse the trend, it is desirable that academic researchers and industry practitioners increase their engagements via mutual presentations, educations, and even government initiatives. All three stakeholders (i.e., academia, industry, and government) should play more active roles.

(3) Vocational and practical education provides numerous opportunities for fruitful academia-industry collaboration. With the changing nature of work and growing technology adoption, there is an increasing demand for radical changes in vocational education—for both teachers and students (World Development and Report, 2019). Domain knowledge provided by teachers is beneficially supplemented by AI-assisted learning environments in academia. Practical skills are enhanced in industrial environments with hands-on experience and feedback from both trainers and technology tools. Hence, students benefit from acquiring domain knowledge and enhancing their skills via interactions with human teachers and trainers. Equally, they benefit from gaining the practical skills via interactions with simulated and real-world technological environments. Effective vocational training demands teachers and trainers on the human-learning side, and AI environments and actual technology tools on machine-learning side. Collaboration between academia and industry, as well as balanced human and machine learning approaches are pertinent for vocational education.

DISCUSSION AND CONCLUSION

Big data and AI have enormous potential to realize highly effective learning and teaching. They stimulate new research questions and designs, exploit innovative technologies and tools in data collection and analysis, and ultimately become

TABLE 1 | Major challenges and possible solutions for integrating big data and AI into education.

Aspect	Major challenges	Possible solutions
Research	• The mode of education is progressively moving from a one-size-fits-all approach to precision education and individualized learning. • AI research in education is currently focused on intelligent computing technologies in a single domain. • The format, purpose, and meaning of machine-generated data should be carefully designed. • The significant influence of AI and big data technologies is a double-edged sword.	• Adaptive educational tools and flexible learning systems will be needed to accommodate individual learners' needs. • The research focus on deploying AI in education needs to incorporate theories of cognition and knowledge about individual differences in student learning. • A theoretical model is needed to guide the development, interpretation, and validation of algorithms. The data analytics must be applied with caution. • Future studies should be aimed at using educational technologies in the appropriate context tailored to the characteristics of individual learners.
Policy-making	• In digitally-driven knowledge economies, traditional formal education systems are undergoing drastic changes or even a paradigm shift. • A large number of pre-service and in-service teachers are not ready to support and adopt new technologies. • There is a pressing need for privacy and personal data protections against unauthorized disclosure, commercial exploitation, and other abuses.	• New methods of instruction, engagement, and assessment will need to be developed in formal education to support lifelong education systems based on micro-credits or micro-degrees. • Effective teacher education and continuing education programs have to be designed and offered to support the adoption of these new technologies. • The government must seek an optimal balance between personal data collection and personal data protection in policy-making, implementation, and enforcement.
Industry	• The commercialization of intelligent educational tools and systems presents a set of difficult challenges. • Expanding spectrum of commercially and freely available tools is necessary to maintain healthy market competition. • Vocational and practical trainings need radical changes to remain relevant and prudent.	• Building a reciprocal and sustained partnership between academia and the education industry is strongly encouraged. • Collaborative research projects sponsored by the industry should provide support for academics to advance applied research and its commercialization. • Closer academia-industry collaboration with balanced human-oriented and machine-assisted learning.

a mainstream research paradigm (Daniel, 2019). Nonetheless, they are still fairly novel and unfamiliar to many researchers and educators. In this paper, we have described the general background, core concepts, and recent progress of this rapidly growing domain. Along with the arising opportunities, we have highlighted the crucial challenges and emerging trends of big data and AI in education, which are reflected in educational research, policy-making, and industry. **Table 1** concisely summarizes the major challenges and possible solutions of big data and AI in education. In summary, future studies should be aimed at theory-based precision education, incorporating cross-disciplinary application, and appropriately using educational technologies. The government should be devoted to supporting lifelong learning, offering teacher education programs, and protecting personal data. With regard to the education industry, reciprocal and mutually beneficial relationships should be developed in order to enhance academia-industry collaboration.

Regarding the future development of big data and AI, we advocate an in-depth dialog between the supporters of "cold" technology and "warm" humanity so that users of technology can benefit from its capacity and not see it as a threat to their livelihood. An equally important issue is that overreliance on technology may lead to an underestimation of the role of humans in education. Remember the fundamental role of schooling: the school is a great equalizer as well as a central socialization agent. We need to better understand the role of social and affective processing (e.g., emotion, motivation) in addition to cognitive processing in student learning successes (or failures). After all, human learning is a social behavior, and a number of key regions in our brains are wired to be socially engaged (see Li and Jeong, 2020 for a discussion).

It has been estimated that approximately half of the current routine jobs might be automated in the near future (Frey and Osborne, 2017; World Development and Report, 2019). However, the teacher's job could not be replaced. The teacher-student relationship is indispensable in students' learning, and inspirational in students' personal growth (Roorda et al., 2011; Cheng and Tsai, 2019). On the other hand, new developments in technologies will enable us to collect and analyze large-scale, multimodal, and continuous real-time data. Such data-intensive and technology-driven analysis of human behavior, in real-world and simulated environments, may assist teachers in identifying students' learning trajectories and patterns, developing corresponding lesson plans, and adopting effective teaching strategies (Klašnja-Milicevic et al., 2017; Gierl and Lai, 2018). It may also support teachers in tackling students' more complex problems and cultivating students' higher-order thinking skills by freeing the teachers from their monotonous and routine tasks (Li, 2007; Belpaeme et al., 2018). Hence, it is now imperative for us to embrace AI and technology and prepare our teachers and students for the future of AI-enhanced and technology-supported education.

The adoption of big data and AI in learning and teaching is still in its infancy and limited by technological and mindset challenges for now; however, the convergence of developments in psychology, data science, and computer science shows great promise in revolutionizing educational research, practice, and industry. We hope that the latest achievements and future directions presented in this paper will advance our shared goal of helping learners and teachers pursue sustainable development.

AUTHOR CONTRIBUTIONS

HLu wrote the initial draft of the manuscript. PG, HLa, JG, and PL revised the drafts and provided theoretical background. SY, HO, JB, and RG contributed content for the original draft preparation of the manuscript. C-CT provided theoretical focus, design, draft feedback, and supervised throughout the research. All authors contributed to the article and approved the submitted version.

FUNDING

This work was financially supported by the Institute for Research Excellence in Learning Sciences of National Taiwan Normal University (NTNU) from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

REFERENCES

Abed Ibrahim, L., and Fekete, I. (2019). What machine learning can tell us about the role of language dominance in the diagnostic accuracy of german litmus non-word and sentence repetition tasks. *Front. Psychol.* 9:2757. doi: 10.3389/fpsyg.2018.02757

Adjerid, I., and Kelley, K. (2018). Big data in psychology: a framework for research advancement. *Am. Psychol.* 73, 899–917. doi: 10.1037/amp0000190

Aldowah, H., Al-Samarraie, H., and Fauzy, W. M. (2019). Educational data mining and learning analytics for 21st century higher education: a review and synthesis. *Telemat. Inform.* 37, 13–49. doi: 10.1016/j.tele.2019.01.007

Anwar, S., Bascou, N. A., Menekse, M., and Kardgar, A. (2019). A systematic review of studies on educational robotics. *J. Pre-College Eng. Educ. Res. (J-PEER)* 9, 19–42. doi: 10.7771/2157-9288.1223

Azevedo, J. P. W. D., Crawford, M. F., Nayar, R., Rogers, F. H., Barron Rodriguez, M. R., Ding, E. Y. Z., et al. (2019). *Ending Learning Poverty: What Will It Take?*. Washington, D.C: The World Bank.

Baker, R. S. J. D., D'Mello, S. K., Rodrigo, M. M. T., and Graesser, A. C. (2010). Better to be frustrated than bored: the incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments. *Int. J. Human-Comp. Stud.* 68, 223–241. doi: 10.1016/j.ijhcs.2009.12.003

Barsalou, L. W. (2008). "Grounding symbolic operations in the brain's modal systems," in *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches*, eds G. R. Semin and E. R. Smith (Cambridge: Cambridge University Press), 9–42. doi: 10.1017/cbo9780511805837.002

Becker, S. A., Cummins, M., Davis, A., Freeman, A., Hall, C. G., and Ananthanarayanan, V. (2017). *NMC Horizon Report: 2017 Higher Education Edition*. Austin, TX: The New Media Consortium.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: a review. *Sci. Robot.* 3:eaat5954. doi: 10.1126/scirobotics.aat5954

Bereiter, C. (2002). *Education and MIND in the Knowledge Age*. Mahwah, NJ: LEA.

Boyd, D., and Crawford, K. (2012). Critical questions for big data: provocations for a cultural, technological, and scholarly phenomenon. *Inform. Commun. Soc.* 15, 662–679. doi: 10.1080/1369118x.2012.678878

Bramley, N. R., Dayan, P., Griffiths, T. L., and Lagnado, D. A. (2017). Formalizing Neurath's ship: approximate algorithms for online causal learning. *Psychol. Rev.* 124, 301–338. doi: 10.1037/rev0000061

Bucher, T. (2018). *If Then: Algorithmic Power and Politics*. New York, NY: Oxford University Press.

Carbonell, J. R. (1970). AI in CAI: an artificial-intelligence approach to computer-assisted instruction. *IEEE Trans. Man-Machine Sys.* 11, 190–202. doi: 10.1109/TMMS.1970.299942

Chen, C. P., and Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: a survey on big data. *Inform. Sci.* 275, 314–347. doi: 10.1016/j.ins.2014.01.015

Chen, L., Chen, P., and Lin, Z. (2020). Artificial intelligence in education: a review. *IEEE Access* 8, 75264–75278. doi: 10.1109/ACCESS.2020.2988510

Chen, N.-S., Yin, C., Isaias, P., and Psotka, J. (2020). Educational big data: extracting meaning from data for smart education. *Interact. Learn. Environ.* 28, 142–147. doi: 10.1080/10494820.2019.1635395

Cheng, K.-H., and Tsai, C.-C. (2019). A case study of immersive virtual field trips in an elementary classroom: students' learning experience and teacher-student interaction behaviors. *Comp. Educ.* 140:103600. doi: 10.1016/j.compedu.2019.103600

Cheung, M. W.-L., and Jak, S. (2018). Challenges of big data analyses and applications in psychology. *Zeitschrift Fur Psychol. J. Psychol.* 226, 209–211. doi: 10.1027/2151-2604/a000348

Cuthbertson, B., Socha, T. L., and Potter, T. G. (2004). The double-edged sword: critical reflections on traditional and modern technology in outdoor education. *J. Adv. Educ. Outdoor Learn.* 4, 133–144. doi: 10.1080/14729670485200491

Cutumisu, M., and Guo, Q. (2019). Using topic modeling to extract pre-service teachers' understandings of computational thinking from their coding reflections. *IEEE Trans. Educ.* 62, 325–332. doi: 10.1109/te.2019.2925253

Daniel, B. (2015). Big data and analytics in higher education: opportunities and challenges. *Br. J. Educ. Technol.* 46, 904–920. doi: 10.1111/bjet.12230

Daniel, B. K. (2019). Big data and data science: a critical review of issues for educational research. *Br. J. Educ. Technol.* 50, 101–113. doi: 10.1111/bjet.12595

Dijcks, J. (2013). *Oracle: Big data for the enterprise. Oracle White Paper*. Redwood Shores, CA: Oracle Corporation.

Field, J. (2001). Lifelong education. *Int. J. Lifelong Educ.* 20, 3–15. doi: 10.1080/09638280010008291

Frey, C. B., and Osborne, M. A. (2017). The future of employment: how susceptible are jobs to computerisation? *Technol. Forecast. Soc. Change* 114, 254–280. doi: 10.1016/j.techfore.2016.08.019

Geczy, P. (2014). Big data characteristics. *Macrotheme Rev.* 3, 94–104.

Geczy, P. (2015). Big data management: relational framework. *Rev. Bus. Finance Stud.* 6, 21–30.

Geczy, P. (2018). Data-Oriented business models: gaining competitive advantage. *Global J. Bus. Res.* 12, 25–36.

Gierl, M. J., and Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Appl. Psychol. Measurement* 42, 42–57. doi: 10.1177/0146621617726788

Gobert, J., Sao Pedro, M., Raziuddin, J., and Baker, R. S. (2013). From log files to assessment metrics for science inquiry using educational data mining. *J. Learn. Sci.* 22, 521–563. doi: 10.1080/10508406.2013.837391

Gobert, J. D., and Sao Pedro, M. A. (2017). "Digital assessment environments for scientific inquiry practices," in *The Wiley Handbook of Cognition and Assessment*, eds A. A. Rupp and J. P. Leighton (West Sussex: Frameworks, Methodologies, and Applications), 508–534. doi: 10.1002/9781118956588.ch21

Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., and Montalvo, O. (2012). Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *J. Educ. Data Min.* 4, 104–143. doi: 10.5281/zenodo.3554645

Goksel, N., and Bozkurt, A. (2019). "Artificial intelligence in education: current insights and future perspectives," in *Handbook of Research on Learning in the Age of Transhumanism*, eds S. Sisman-Ugur and G. Kurubacak (Hershey, PA: IGI Global), 224–236 doi: 10.4018/978-1-5225-8431-5.ch014

Graesser, A. C., Chipman, P., Haynes, B. C., and Olney, A. (2005). AutoTutor: an intelligent tutoring system with mixed-initiative dialogue. *IEEE Trans. Educ.* 48, 612–618. doi: 10.1109/te.2005.856149

GSMA Intelligence (2020). *The Mobile Economy 2020*. London: GSM Association.

Harlow, L. L., and Oswald, F. L. (2016). Big data in psychology: introduction to the special issue. *Psychol. Methods* 21, 447–457. doi: 10.1037/met0000120

Hew, K. F., Lan, M., Tang, Y., Jia, C., and Lo, C. K. (2019). Where is the "theory" within the field of educational technology research? *Br. J. Educ. Technol.* 50, 956–971. doi: 10.1111/bjet.12770

Hinojo-Lucena, F. J., Aznar-Díaz, I., Cáceres-Reche, M. P., and Romero-Rodríguez, J. M. (2019). Artificial intelligence in higher education: a bibliometric study on its impact in the scientific literature. *Educ. Sci.* 9:51. doi: 10.3390/educsci9010051

Huang, A. Y., Lu, O. H., Huang, J. C., Yin, C., and Yang, S. J. (2020). Predicting students' academic performance by using educational big data and learning analytics: evaluation of classification methods and learning logs. *Int. Learn. Environ.* 28, 206–230. doi: 10.1080/10494820.2019.1636086

Huggins, R., and Thompson, P. (2015). Entrepreneurship, innovation and regional growth: a network theory. *Small Bus. Econ.* 45, 103–128. doi: 10.1007/s11187-015-9643-3

Hwang, G.-J., Spikol, D., and Li, K.-C. (2018). Guest editorial: trends and research issues of learning analytics and educational big data. *Educ. Technol. Soc.* 21, 134–136.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., et al. (2014). Big data and its technical challenges. *Commun. ACM.* 57, 86–94. doi: 10.1145/2611567

Johnson, L., Smith, R., Willis, H., Levine, A., and Haywood, K. (2011). *The 2011 Horizon Report*. Austin, TX: The New Media Consortium.

Jordan, M. I., and Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science* 349, 255–260. doi: 10.1126/science.aaa8415

Khechine, H., and Lakhal, S. (2018). Technology as a double-edged sword: from behavior prediction with UTAUT to students' outcomes considering personal characteristics. *J. Inform. Technol. Educ. Res.* 17, 63–102. doi: 10.28945/4022

Klašnja-Milicevic, A., Ivanovic, M., and Budimac, Z. (2017). Data science in education: big data and learning analytics. *Comput. Applicat. Eng. Educ.* 25, 1066–1078. doi: 10.1002/cae.21844

Koh, J. H. L., Chai, C. S., and Tsai, C. C. (2014). Demographic factors, TPACK constructs, and teachers' perceptions of constructivist-oriented TPACK. *J. Educ. Technol. Soc.* 17, 185–196.

Koper, R., and Tattersall, C. (2004). New directions for lifelong learning using network technologies. *Br. J. Educ. Technol.* 35, 689–700. doi: 10.1111/j.1467-8535.2004.00427.x

Krouska, A., Troussas, C., and Virvou, M. (2019). SN-Learning: an exploratory study beyond e-learning and evaluation of its applications using EV-SNL framework. *J. Comp. Ass. Learn.* 35, 168–177. doi: 10.1111/jcal.12330

Laney, D. (2001). 3D data management: controlling data volume, velocity and variety. *META Group Res. Note* 6, 70–73.

Lazer, D., Kennedy, R., King, G., and Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science* 343, 1203–1205. doi: 10.1126/science.1248506

Lenschow, R. J. (1998). From teaching to learning: a paradigm shift in engineering education and lifelong learning. *Eur. J. Eng. Educ.* 23, 155–161. doi: 10.1080/03043799808923494

L'heureux, A., Grolinger, K., Elyamany, H. F., and Capretz, M. A. (2017). Machine learning with big data: challenges and approaches. *IEEE Access* 5, 7776–7797. doi: 10.1109/ACCESS.2017.2696365

Li, H., Gobert, J., and Dickler, R. (2019). "Evaluating the transfer of scaffolded inquiry: what sticks and does it last?," in *Artificial Intelligence in Education*, eds S. Isotani, E. Millán, A. Ogan, P. Hastings, B. McLaren, and R. Luckin (Cham: Springer), 163–168. doi: 10.1007/978-3-030-23207-8_31

Li, P., and Jeong, H. (2020). The social brain of language: grounding second language learning in social interaction. *npj Sci. Learn.* 5:8. doi: 10.1038/s41539-020-0068-7

Li, P., Legault, J., Klippel, A., and Zhao, J. (2020). Virtual reality for student learning: understanding individual differences. *Hum. Behav. Brain* 1, 28–36. doi: 10.37716/HBAB.2020010105

Li, X. (2007). Intelligent agent–supported online education. *Dec. Sci. J. Innovat. Educ.* 5, 311–331. doi: 10.1111/j.1540-4609.2007.00143.x

Loftus, M., and Madden, M. G. (2020). A pedagogy of data and Artificial intelligence for student subjectification. *Teach. Higher Educ.* 25, 456–475. doi: 10.1080/13562517.2020.1748593

Long, P., and Siemens, G. (2011). Penetrating the fog: analytics in learning and education. *Educ. Rev.* 46, 31–40. doi: 10.1007/978-3-319-38956-1_4

Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Lin, A. J. Q., Ogata, H., and Yang, S. J. H. (2018). Applying learning analytics for the early prediction of students' academic performance in blended learning. *Educ. Technol. Soc.* 21, 220–232.

Macfadyen, L. P. (2017). Overcoming barriers to educational analytics: how systems thinking and pragmatism can help. *Educ. Technol.* 57, 31–39.

Malik, G., Tayal, D. K., and Vij, S. (2019). "An analysis of the role of artificial intelligence in education and teaching," in *Recent Findings in Intelligent Computing Techniques. Advances in Intelligent Systems and Computing*, eds P. Sa, S. Bakshi, I. Hatzilygeroudis, and M. Sahoo (Singapore: Springer), 407–417.

Manuel Moreno-Marcos, P., Alario-Hoyos, C., Munoz-Merino, P. J., and Delgado Kloos, C. (2019). Prediction in MOOCs: a review and future research directions. *IEEE Trans. Learn. Technol.* 12, 384–401. doi: 10.1109/TLT.2018.2856808

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., et al. (2011). *Big data: The Next Frontier for Innovation, Competition and Productivity*. New York, NY: McKinsey Global Institute.

Mayer-Schönberger, V., and Cukier, K. (2013). *Big data: A Revolution That Will Transform How we live, Work, and Think*. Boston, MA: Houghton Mifflin Harcourt.

Mislevy, R. J., Yan, D., Gobert, J., and Sao Pedro, M. (2020). "Automated scoring in intelligent tutoring systems," in *Handbook of Automated Scoring*, eds D. Yan, A. A. Rupp, and P. W. Foltz (London: Chapman and Hall/CRC), 403–422. doi: 10.1201/9781351264808-22

Mowery, D. C., Nelson, R. R., Sampat, B. N., and Ziedonis, A. A. (2001). The growth of patenting and licensing by US universities: an assessment of the effects of the Bayh–Dole act of 1980. *Res. Pol.* 30, 99–119. doi: 10.1515/9780804796361-008

Nye, B. D. (2015). Intelligent tutoring systems by and for the developing world: a review of trends and approaches for educational technology in a global context. *Int. J. Art. Intell. Educ.* 25, 177–203. doi: 10.1007/s40593-014-0028-6

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

O'Donnell, C. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K-12 curriculum intervention research. *Rev. Educ. Res.* 78, 33–84. doi: 10.3102/0034654307313793

Papamitsiou, Z., and Economides, A. A. (2014). Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. *Educ. Technol. Soc.* 17, 49–64.

Pardo, A., and Siemens, G. (2014). Ethical and privacy principles for learning analytics. *Br. J. Educ. Technol.* 45, 438–450. doi: 10.1111/bjet.12152

Pedró, F., Subosa, M., Rivas, A., and Valverde, P. (2019). *Artificial Intelligence in Education: Challenges and Opportunities for Sustainable Development*. Paris: UNESCO.

Peters, M. A. (2018). Deep learning, education and the final stage of automation. *Educ. Phil. Theory* 50, 549–553. doi: 10.1080/00131857.2017.1348928

Popenici, S. A., and Kerr, S. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Res. Pract. Technol. Enhanced Learn.* 12:22. doi: 10.1186/s41039-017-0062-8

Quadir, B., Chen, N.-S., and Isaias, P. (2020). Analyzing the educational goals, problems and techniques used in educational big data research from 2010 to 2018. *Int. Learn. Environ.* 1–17. doi: 10.1080/10494820.2020.1712427

Renz, A., and Hilbig, R. (2020). Prerequisites for artificial intelligence in further education: identification of drivers, barriers, and business models of educational technology companies. *Int. J. Educ. Technol. Higher Educ.* 17:14. doi: 10.1186/s41239-020-00193-3

Renz, A., Krishnaraja, S., and Gronau, E. (2020). Demystification of artificial intelligence in education–how much ai is really in the educational technology? *Int. J. Learn. Anal. Art. Intell. Educ. (IJAI).* 2, 4–30. doi: 10.3991/ijai.v2i1.12675

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., and Oort, F. J. (2011). The influence of affective teacher-student relationships on students' school engagement and achievement: a meta-analytic approach. *Rev. Educ. Res.* 81, 493–529. doi: 10.3102/0034654311421793

Sao Pedro, M., Baker, R., and Gobert, J. (2013). "What different kinds of stratification can reveal about the generalizability of data-mined skill assessment models," in *Proceedings of the 3rd Conference on Learning Analytics and Knowledge* (Leuven), 190–194.

Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., and Tufano, P. (2012). Analytics: the real-world use of big data. *IBM Global Bus. Serv.* 12, 1–20. doi: 10.1002/9781119204183.ch1

Sharples, M. (2000). The design of personal mobile technologies for lifelong learning. *Comp. Educ.* 34, 177–193. doi: 10.1016/s0360-1315(99)00044-5

Smutny, P., and Schreiberova, P. (2020). Chatbots for learning: a review of educational chatbots for the facebook messenger. *Comp. Educ.* 151:103862. doi: 10.1016/j.compedu.2020.103862

Sonderlund, A. L., Hughes, E., and Smith, J. (2019). The efficacy of learning analytics interventions in higher education: a systematic review. *Br. J. Educ. Technol.* 50, 2594–2618. doi: 10.1111/bjet.12720

Song, Y., Dai, X.-Y., and Wang, J. (2016). Not all emotions are created equal: expressive behavior of the networked public on China's social media site. *Comp. Hum. Behav.* 60, 525–533. doi: 10.1016/j.chb.2016.02.086

Spikol, D., Ruffaldi, E., Dabisias, G., and Cukurova, M. (2018). Supervised machine learning in multimodal learning analytics for estimating success in project-based learning. *J. Comp. Ass. Learn.* 34, 366–377. doi: 10.1111/jcal.12263

Staats, C. (2016). Understanding implicit bias: what educators should know. *Am. Educ.* 39, 29–33. doi: 10.2307/3396655

Starcic, A. I. (2019). Human learning and learning analytics in the age of artificial intelligence. *Br. J. Educ. Technol.* 50, 2974–2976. doi: 10.1111/bjet.12879

The International Learning Sciences Forum (2019). *The International Learning Sciences Forum: International Trends for Ai and Big Data in Learning Sciences*. Taipei: National Taiwan Normal University.

Toh, L. P. E., Causo, A., Tzuo, P. W., Chen, I. M., and Yeo, S. H. (2016). A review on the use of robots in education and young children. *J. Educ. Technol. Soc.* 19, 148–163.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7

Tsai, C. C. (2000). Relationships between student scientific epistemological beliefs and perceptions of constructivist learning environments. *Educ. Res.* 42, 193–205. doi: 10.1080/001318800363836

Tsai, S. C., Chen, C. H., Shiao, Y. T., Ciou, J. S., and Wu, T. N. (2020). Precision education with statistical learning and deep learning: a case study in Taiwan. *Int. J. Educ. Technol. Higher Educ.* 17, 1–13. doi: 10.1186/s41239-020-00186-2

UNESCO (2015). *SDG4-Education 2030, Incheon Declaration (ID) and Framework for Action. For the Implementation of Sustainable Development Goal 4, Ensure Inclusive and Equitable Quality Education and Promote Lifelong Learning Opportunities for All, ED-2016/WS/28*. London: UNESCO

United Nations (2020). *Policy Brief: Education During Covid-19 and Beyond*. New York, NY: United Nations

VanRullen, R. (2017). Perception science in the age of deep neural networks. *Front. Psychol.* 8:142. doi: 10.3389/fpsyg.2017.00142

Viberg, O., Hatakka, M., Bälter, O., and Mavroudi, A. (2018). The current landscape of learning analytics in higher education. *Comput. Human Behav.* 89, 98–110. doi: 10.1016/j.chb.2018.07.027

Williams, P. (2019). Does competency-based education with blockchain signal a new mission for universities? *J. Higher Educ. Pol. Manag.* 41, 104–117. doi: 10.1080/1360080x.2018.1520491

World Development and Report (2019). *The Changing Nature of Work*. Washington, DC: The World Bank/International Bank for Reconstruction and Development.

Xie, H., Chu, H.-C., Hwang, G.-J., and Wang, C.-C. (2019). Trends and development in technology-enhanced adaptive/personalized learning: a

systematic review of journal publications from 2007 to 2017. *Comp. Educ.* 140:103599. doi: 10.1016/j.compedu.2019.103599

Yadegaridehkordi, E., Noor, N. F. B. M., Ayub, M. N. B., Affal, H. B., and Hussin, N. B. (2019). Affective computing in education: a systematic review and future research. *Comp. Educ.* 142:103649. doi: 10.1016/j.compedu.2019.103649

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Zawacki-Richter, O., Marín, V. I., Bond, M., and Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education–where are the educators? *Int. J. Educ. Technol. Higher Educ.* 16:39. doi: 10.1186/s41239-019-0171-0

Conflict of Interest: JG was employed by company Apprendis, LLC, Berlin.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 25 November 2020
doi: 10.3389/fpsyg.2020.578092

Gender Differences in the Interest in Mathematics Schoolwork Across 50 Countries

Kimmo Eriksson[1,2]*

[1] School of Education, Culture and Communication, Mälardalen University, Västerås, Sweden, [2] Centre for Cultural Evolution, Stockholm University, Stockholm, Sweden

Although much research has found girls to be less interested in mathematics than boys are, there are many countries in which the opposite holds. I hypothesize that variation in gender differences in interest are driven by a complex process in which national culture promoting high math achievement drives down interest in math schoolwork, with the effect being amplified among girls due to their higher conformity to peer influence. Predictions from this theory were tested in a study of data on more than 500,000 grade 8 students in 50 countries from the 2011 and 2015 waves of TIMSS. Consistent with predictions, national achievement levels were strongly negatively correlated with national levels of math schoolwork interest and this variation was larger among girls: girls in low-achievement, high-interest countries had especially high interest in math schoolwork, whereas girls in high-achievement, low-interest countries had especially low interest in math schoolwork. Gender differences in math schoolwork interest were also found to be related to gender differences in math achievement, emphasizing the importance of understanding them better.

Keywords: learning attitudes, gender differences, mathematics achievement, peer influence, female amplification

OPEN ACCESS

Edited by:
Ronnel B. King,
University of Macau, China

Reviewed by:
Manuel Soriano-Ferrer,
University of Valencia, Spain
Hanke Korpershoek,
University of Groningen, Netherlands

***Correspondence:**
Kimmo Eriksson
kimmo.eriksson@mdh.se;
kimmoe@gmail.com

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *30 June 2020*
Accepted: *02 November 2020*
Published: *25 November 2020*

Citation:
Eriksson K (2020) Gender Differences in the Interest in Mathematics Schoolwork Across 50 Countries. Front. Psychol. 11:578092. doi: 10.3389/fpsyg.2020.578092

INTRODUCTION

Children and young adolescents are typically obliged to go to school and must take part in schoolwork even if they do not find it interesting. Nonetheless, it is preferable that students are interested in their schoolwork, both because they are likely to experience more satisfaction in school and because they are likely to achieve better (e.g., Artelt et al., 2003). Given these benefits of having high levels of interest, it is problematic that a large body of research has found that girls tend to have less interest in mathematics than boys do (Hyde et al., 1990; Lippa, 1998; Preckel et al., 2008; Su et al., 2009; Frenzel et al., 2010). However, this gender gap in mathematics interest does not seem to be universal. Recent research using cross-national data from the Trends in Mathematics and Science Survey (TIMSS) has uncovered that in many countries the gender gap in mathematics attitudes, including interest in schoolwork, is reversed (Ghasemi and Burley, 2019; Reilly et al., 2019). These findings suggest that the correct question to ask is not why girls are less interested in math than boys are, because often the opposite holds. From an egalitarian perspective, however, a gender gap in interest in a key subject in school may be regarded as equally problematic regardless of whether it favors boys or girls. Therefore, the present research aims to better understand why the level of interest in mathematics schoolwork may differ between the average boys and the average girls in a society develop, and why the difference may go either way depending on the society.

The theoretical idea I propose is that gender differences in interest in schoolwork may be influenced by a societal factor—the achievement culture, which tends to drive interest in schoolwork down—in combination with a gender difference in conformity, with girls tending to conform more than boys. The outcome, I argue, would be a specific, complex pattern. In high-achievement cultures, it would be common for students to have a low level of interest in math schoolwork and, due to conformity, a low level of interest would be especially common among girls. In low-achievement cultures, by contrast, it would be common for students to have a high level of interest in math schoolwork and, again due to conformity, a high level of interest would be especially common among girls. Thus, high-achievement cultures would exhibit gender gaps in math schoolwork interest that favor boys, while gender gaps would be reversed in low-achievement cultures. Below I develop this novel hypothesis in greater detail, grounding its assumptions in previous literature.

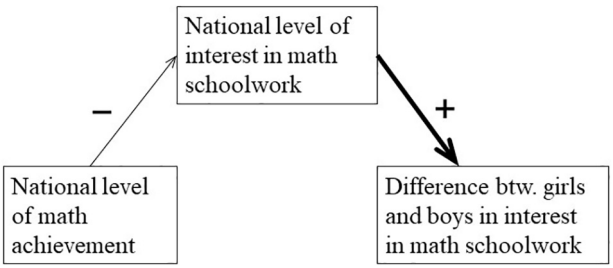

FIGURE 1 | The hypothesized path from high-achievement culture to gender differences in interest in math schoolwork. The first arrow represents the hypothesis that high-achievement nations will have lower average student interest in math schoolwork. The second (bold) arrow represents the hypothesis that differences in average national interest will be amplified among girls.

The Impact of High-Achievement Culture on Students' Math Schoolwork Interest

The achievement culture of a society may be an important factor behind how interested students are in mathematics schoolwork. When comparing across countries, it is well-known that a high average level of student achievement in mathematics and science is related to a range of negative outcomes, including more negative attitudes to math and science (Artelt et al., 2003; Shen and Tam, 2008; Leung, 2014; Täht et al., 2014), worse student-teacher relations (Mikk et al., 2016), and worse academic self-concept among students (Shen and Tam, 2008; Van de gaer et al., 2012; Leung, 2014). For instance, on a global scale, math achievement is very low in Egypt and very high in Japan; nonetheless, a high positive affect toward mathematics was found to be very common among grade 8 students in Egypt but very uncommon in the same grade in Japan (see Table 21 in Leung, 2014). This phenomenon has been described as "paradoxical" (Shen and Tam, 2008). At the individual level, positive attitudes, good student-teacher relations, and a positive academic self-concept are generally regarded as conducive to learning (Artelt et al., 2003). Yet, a national culture that focuses on high achievement may bring about more negative attitudes, worse student-teacher relations, and a more negative academic self-concept. To explain this phenomenon, researchers have pointed to particularly high educational norms and standards in high-achievement countries (Van de gaer et al., 2012). Many students may struggle to fully meet these high standards, even though their achievement is high on a global scale, and this could account for the surprisingly low academic self-concept in high-achievement countries (Van de gaer et al., 2012). The level of interest in schoolwork could similarly be driven down by high educational norms and standards. Students who are struggling to keep up with a progressively difficult curriculum in math could lose interest in doing the schoolwork to progress to even more advanced math, instead preferring to consolidate their knowledge. Consistent with this body of research, I therefore expect that a higher average level of math achievement in a society is linked to a lower average level of interest in math schoolwork. The first (fine) arrow in the diagram in **Figure 1** illustrates this hypothesized negative relation.

Peer Influence Among Girls and Boys

It is well-known that students' motivation may be heavily influenced by their peers, both positively and negatively, and both intentionally and unintentionally (Kindermann, 2016; Wentzel and Muenks, 2016). Thus, if high-achievement culture makes some students lose interest in math schoolwork, it may negatively influence their peers' level of interest as well. Moreover, this effect of peer influence could differ between boys and girls. Studies of peer influence in school that have addressed the gender aspect have found that, compared to boys, girls' interest in schoolwork is more susceptible to peer influence (Berndt and Keefe, 1995; Riegle-Crumb et al., 2006). This effect has been attributed to girls' friendships being more supportive and discussion-oriented, compared to boys' friendships that tend to be more competitive and center more on specific activities (Riegle-Crumb et al., 2006; see also Beutel and Marini, 1995). Gender differences in the susceptibility to peer influence in school settings are consistent with findings from conformity research is general. Several meta-analyses have found conformity to be stronger among women than among men (Eagly and Carli, 1981; Bond and Smith, 1996). Although most of these studies have been carried out in Western countries, stronger conformity among women than among men was also observed in a recent study in Sudan when participants evaluated appropriate behavior (Efferson and Vogt, 2018). The gender effect on conformity has been explained in terms of women having less confidence and being more risk-averse than men (Cross et al., 2017; Brand et al., 2018).

Based on this previous literature I expect that within-society conformity with respect to interest in math schoolwork will be more accentuated among women than among men (through mechanisms such as gender differences in confidence, risk aversion, and friendships). However, the focus of the present research is on *between-society* variation, and greater conformity within societies is likely to lead to larger variation between societies. To understand this theoretical point, consider boys and girls in a high-achievement culture where some students

lose their interest in math schoolwork due to high educational norms and standards. To some degree they will exert a negative influence on their peers' interest, and this peer influence is expected to be more pervasive among girls than among boys. Thus, the direct negative effect of high-achievement culture on students' interest is expected to be amplified by social dynamics, and this amplification is expected to be stronger among girls than among boys. The result would be greater variation across societies in the average interest levels of girls than in the average interest levels of boys.

This hypothesis is illustrated in the diagram in **Figure 1**, where the second (bold) arrow signifies the gender-specific amplification of national differences in interest in math schoolwork. Together, the two arrows describe a hypothetical indirect negative effect of a high national level of math achievement on gender differences in math schoolwork interest, mediated by the national level of math schoolwork interest.

The Impact of Gender Differences in Math Schoolwork Interest on Math Achievement

Gender differences in math schoolwork interest are important not least because they are likely to impact on the math achievement of boys and girls. At the individual level, interest in math schoolwork is thought to be conducive to learning (Artelt et al., 2003). In societies where girls tend to be less interested in math schoolwork than the boys are, it could contribute to a corresponding gender difference in math achievement. Why the gender gap in math achievement varies across societies has been the topic of extensive research for decades (e.g., Guiso et al., 2008; Else-Quest et al., 2010). A recent study, using data from all waves of the TIMSS and PISA assessments between 2000 and 2015, found the societal level of gender egalitarian values to be the strongest and most robust predictor of gender differences in achievement in math, science, and reading, but there was still a large amount of unexplained variation (Eriksson et al., 2020). I hypothesize that some of this variation is accounted for by gender differences in interest in schoolwork. This would underscore the importance of the study's main aim of understanding why these gender differences vary across countries.

Research Questions

Above I have outlined a theory about antecedents and consequences of societal levels of math schoolwork interest among girls and boys. To fully test claims of causality would require experimental or, at least, longitudinal data, neither of which are available. Instead, I here make do with analyzing cross-sectional data provided by TIMSS. The theory predicts certain statistical patterns to arise in such data and the aim of the empirical part of this study is to examine whether these patterns can indeed be observed. It is an important first test of the theory to see whether it correctly predicts several non-trivial features of a complex dataset, even though alternative causal accounts cannot be excluded.

RQ1. The hypothesis of high-achievement culture impacting on students' math schoolwork interest yields the first prediction to be examined: Is there a negative correlation between national levels of achievement and math schoolwork interest?

RQ2. The hypothesis of a difference between boys and girls in peer influence on math schoolwork interest yields a suite of testable predictions: (a) Is within-society variation in math schoolwork interest smaller among girls than among boys? (b) Is between-society variation in math schoolwork interest larger among girls than among boys? (c) Is there a positive correlation between national levels of math schoolwork interest and gender gaps in math schoolwork interest favoring girls? (d) Do national levels of math schoolwork interest mediate a negative correlation between national levels of achievement and gender gaps in math schoolwork interest favoring girls?

RQ3. The hypothesis that gender differences in math schoolwork interest has an independent impact on the gender gap in math achievement also yields a testable prediction: Does the gender gap in math schoolwork interest account for some of the variance in the gender gap in math achievement, over and beyond the variation already accounted for by gender egalitarian values?

MATERIALS AND METHODS

To answer the research questions, the current study analyzes TIMSS data. TIMSS is an excellent resource for comparative research as it uses large representative national samples of students from many countries. Details on the design are provided by the International Association for the Evaluation of Educational Achievement (Martin et al., 2016). In brief, TIMSS assesses math achievement of students in the eighth grade, in which most participants are about 14 years old. In addition to the achievement test, participating students also complete a background questionnaire. This questionnaire is not fixed across waves. In the 2011 and 2015 waves of TIMSS, the questionnaire included items on students' interest in what the teacher says and students' interest in what the teacher tells them to do. No such questions were included in previous waves of TIMSS, nor have they been included in other large-scale international student assessments like PISA. For this reason, this study will use data from the 2011 and 2015 waves of TIMSS.

Data from the 2011 and 2015 waves of TIMSS were downloaded from IEA[1]. Data were available for a total of 50 countries, out of which 35 countries had participated in both waves, 10 countries had participated only in the 2011 wave, and 5 countries only in the 2015 wave. See **Table 1** for countries and samples sizes in each wave. All populated world continents were represented, including 24 countries in Asia from Israel and Saudi Arabia in the west to Korea and Japan in the east, 6 countries in Africa from Morocco and Egypt in the north to Botswana and South Africa in the south, 4 countries in the Americas from Canada to Chile, 14 countries in Europe from Sweden to Malta, as well as Australia and New Zealand.

The TIMSS datasets come with appropriate sampling weights, which were used when calculating the below measures. Missing data (less than 3% of data) were ignored. Preliminary analyses

[1]https://timssandpirls.bc.edu/

TABLE 1 | TIMSS sample sizes and key measures.

Country	2011	2015	Math schoolwork interest				Mean math		Gender
	Sample size	Sample size	Girls		Boys		Achievement		Egalitarian values
			M	SD	M	SD	Girls	Boys	
Armenia	7,556	10,338	3.52	0.63	3.45	0.69	472.8	464.8	−0.91
Australia	4,640	4,918	2.70	0.79	2.79	0.78	502.6	508.2	1.09
Bahrain	5,846	5,060	3.08	0.76	3.05	0.84	446.4	417.3	
Botswana	5,400	5,964	3.36	0.68	3.31	0.71	401.8	385.4	
Canada		8,757	2.97	0.74	3.01	0.74	525.8	530.8	1.31
Chile	5,835	4,849	3.07	0.78	3.12	0.79	414.0	430.1	0.31
Chinese Taipei	5,042	5,711	2.41	0.74	2.51	0.81	605.8	602.7	
Egypt		4,035	3.50	0.69	3.43	0.74	396.6	387.4	−1.87
Finland	4,266		2.42	0.72	2.40	0.73	516.6	512.1	0.62
Georgia	4,563	4,155	3.39	0.65	3.33	0.70	441.7	442.7	−0.85
Ghana	7,812		3.53	0.58	3.58	0.55	318.7	342.2	−0.65
Honduras	7,323		3.53	0.65	3.51	0.68	327.8	351.0	
Hong Kong	4,418	4,893	2.58	0.75	2.73	0.81	590.2	590.0	
Hungary	4,015	6,130	2.74	0.77	2.74	0.81	505.8	513.3	0.52
Indonesia	5,178		3.14	0.44	3.10	0.48	392.4	379.5	−0.56
Iran	5,795	4,704	3.16	0.75	3.24	0.76	424.6	426.5	−1.10
Ireland		5,512	2.75	0.81	2.80	0.82	520.8	526.3	1.43
Israel	6,029	4,481	2.91	0.82	2.88	0.85	515.1	512.6	0.68
Italy	4,699	4,745	2.82	0.70	2.85	0.74	491.9	500.5	0.87
Japan	3,979	4,887	2.26	0.66	2.39	0.72	576.8	578.4	−0.13
Jordan	4,414	7,865	3.47	0.65	3.43	0.71	407.6	384.3	−1.55
Kazakhstan	4,390	5,309	3.42	0.55	3.31	0.58	508.6	506.2	0.13
Korea	7,694	4,503	2.28	0.65	2.38	0.71	607.4	611.3	−0.23
Kuwait		3,873	3.01	0.79	3.23	0.76	396.0	388.7	−1.39
Lebanon	5,166	4,347	3.30	0.76	3.30	0.76	442.3	449.9	−0.32
Lithuania	3,974	9,726	2.91	0.73	2.92	0.75	508.5	505.3	−0.09
Macedonia	4,747		3.23	0.79	3.23	0.78	429.6	422.6	0.09
Malaysia	5,733	3,817	3.18	0.63	3.04	0.67	459.5	445.5	−0.98
Malta		13,035	2.80	0.81	2.93	0.83	495.1	492.7	
Morocco	8,986	8,883	3.49	0.64	3.46	0.66	378.5	377.8	−0.97
New Zealand	9,542	8,142	2.72	0.78	2.85	0.77	486.8	494.6	0.49
Norway	5,336	4,697	2.70	0.77	2.78	0.78	493.6	493.1	1.90
Oman	3,862	5,403	3.52	0.56	3.31	0.71	408.4	361.0	
Palestine	4,422		3.57	0.57	3.38	0.72	415.3	392.2	
Qatar	5,523	4,780	3.01	0.81	3.08	0.83	427.8	419.1	−1.48
Romania	4,893		3.05	0.79	2.98	0.81	463.7	452.8	−0.13
Russia	4,344	3,759	3.09	0.69	3.09	0.70	535.9	541.0	−0.30
Saudi Arabia	5,927	6,116	3.21	0.74	3.16	0.81	388.1	373.5	−1.79
Singapore	4,415	4,257	2.89	0.68	2.94	0.72	620.5	611.7	0.20
Slovenia	11,969	12,514	2.55	0.69	2.55	0.75	508.8	512.4	0.86
South Africa	5,573	4,090	3.41	0.65	3.38	0.68	364.8	359.5	0.04
Sweden	4,413	6,482	2.61	0.71	2.73	0.74	492.5	493.8	1.66
Syria	6,124		3.55	0.60	3.48	0.69	374.9	384.7	
Thailand	14,089	18,012	3.29	0.55	3.24	0.60	437.3	419.7	−0.20
Tunisia	5,128		3.47	0.65	3.43	0.68	416.8	433.6	
Turkey	6,928	6,079	3.11	0.65	3.06	0.71	459.0	451.6	−0.17
UAE	3,378	7,822	3.15	0.73	3.15	0.77	467.7	453.1	
United Kingdom	4,062	10,221	2.71	0.76	2.83	0.77	514.7	511.5	1.25
Ukraine	10,477		3.30	0.66	3.27	0.69	477.6	481.0	−0.25
United States	3,842	4,814	2.80	0.83	2.83	0.83	512.9	515.5	1.18

All measures are based on TIMSS data except the measure of gender egalitarian values, which is based on data from World Values Survey and GLOBE.

revealed that country measures were highly consistent across the two waves. For the below analysis we therefore pooled the individual data from the two waves.

Students' Interest in Math Schoolwork

The student questionnaire in the 2011 and 2015 waves of TIMSS included three items bearing explicitly on interest in math schoolwork: "I am interested in what my teacher says," "My teacher gives me interesting things to do," and "I learn many interesting things in mathematics." For each item, students gave their response on a four-point scale: *Disagree a lot* (coded 1), *Disagree a little* (coded 2), *Agree a little* (coded 3), *Agree a lot* (coded 4). These three items were averaged to an internally consistent measure of students' interest in math schoolwork ($\alpha = 0.78$). Mean value and standard deviation, separately among girls and among boys, are reported per country in **Table 1**.

Student Achievement in Mathematics

TIMSS provides ready calculated national average scores for girls' and boys' math achievement, which were downloaded using the International Data Explorer of the National Center for Education Statistics[2] (see **Table 1**).

Gender Egalitarian Values

Following Eriksson et al. (2020), gender egalitarian values were measured using the Equality index from the World Values Survey (Welzel, 2013) and the Gender Egalitarian Cultural Values index from the GLOBE project (House et al., 2004). Both measures are based on survey responses to items on how society should be with respect to gender equality in education, leadership, and jobs. From Eriksson et al. (2020), measures of gender egalitarian values were obtained for 39 out of the 50 countries in the study: 26 countries had measures from both WVS and GLOBE, 11 countries only from WVS, and 2 countries only from GLOBE. On the set of 26 countries for which both measures were available, they were very strongly correlated, $r = 0.85$, indicating that they indeed measure the same construct so that the measures can be combined. After transforming both WVS and GLOBE measures to z-scores (i.e., standardizing both measures to have the same mean value, zero, and the same standard deviation, one), I combined them into a single measure, using their average for any country where both measures were available.

RESULTS

Country levels of math schoolwork interest, general affect toward math, and math achievement, were calculated as the averages of the corresponding mean values for girls and boys. Gender differences for the same variables were similarly calculated as the mean value for girls minus the mean value for boys. **Table 2** presents descriptive statistics of these national levels and gender differences.

[2]https://nces.ed.gov/timss/idetimss/

TABLE 2 | Descriptive statistics for country levels and gender differences of the measures in **Table 1**.

Variable	*M*	*SD*	Min	Max
Country level of math schoolwork interest	3.06	0.34	2.33	3.55
Country level of math achievement	464.3	70.3	330	616
Gender difference in math schoolwork interest	−0.01	0.09	−0.22	0.20
Gender difference in math achievement	2.9	12.7	−24	47

*Based on n = 50 countries. Country levels are calculated as the average of the mean values for girls and boys in **Table 1**, while gender differences are calculated as the girls' mean value minus the boys' mean value.*

RQ1: The Predicted Relation Between Country Levels of Achievement and Math Schoolwork Interest

The first research question concerns the prediction of a negative correlation between country levels of achievement and math schoolwork interest. In line with the prediction, a very strong negative correlation was observed, $r(48) = -0.81$, 95% CI [−0.70, −0.89], $p < 0.001$. Here and throughout, I report bias corrected accelerated confidence intervals based on 1,000 bootstrap samples generated by SPSS v. 26.

RQ2: Predictions Based on an Assumed Difference in Peer Influence on Girls' vs. Boys' Math Schoolwork Interest

RQ2a: Is within-society variation in math schoolwork interest smaller among girls than among boys?

In **Table 1**, the standard deviation in math schoolwork interest was smaller among girls than among boys in 43 out of 50 countries, in line with the prediction. After transforming standard deviations to variances, the mean difference between girls ($M = 0.497$, $SD = 0.115$) and boys ($M = 0.545$, $SD = 0.104$) was −0.048, 95% CI [−0.062, −0.034], $t(49) = -6.93$, $p < 0.001$, $d = 0.98$, paired samples t-test.

RQ2b: Is Between-Society Variation in Math Schoolwork Interest Larger Among Girls Than Among Boys?

Figure 2 presents a scatterplot of girls' and boys' levels of interest in math schoolwork plotted against the average level. As indicated by the regression lines, girls in high-interest countries are even more interested than the boys are, while girls in low-interest countries are even less interested than the boys. Thus, between-society variation was larger for girls, in line with the prediction. To quantify the difference, the country variance of the math schoolwork interest level among girls was $\sigma^2 = 0.133$, which is 36% higher than the corresponding country variance among boys, $\sigma^2 = 0.098$. To estimate the statistical significance, we use the Morgan-Pitman test for difference in variance in paired data. This test assumes normally distributed data, and Kolmogorov-Smirnov tests indicated that the country level data on math schoolwork interest did not deviate from a normal distribution either for girls or boys, $ps > 0.20$. The Morgan-Pitman test says that testing for a difference in variance in paired data is equivalent to testing

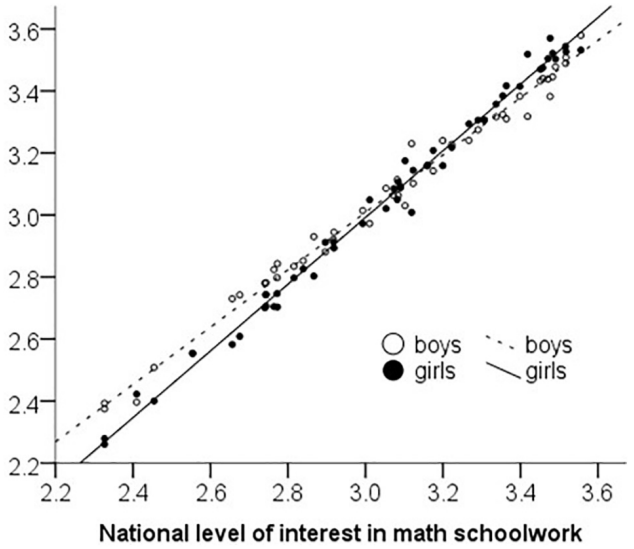

FIGURE 2 | Country variation in math schoolwork interest is amplified among girls. In high-interest countries, the interest levels of girls (black dots, solid line) tend to be higher than the interest levels of boys (circles, dashed regression line), while in low-interest countries girls tend to have lower interest levels than boys do.

for a correlation between the mean of the paired variables and the difference between the paired values (Wilcox, 2015). In our case, this means testing for a correlation between the total country level of math schoolwork interest and the gender difference in math schoolwork interest. In other words, research questions RQ2b and RQ2c are statistically equivalent. We conduct the test below.

RQ2c: Is There a Positive Correlation Between National Levels of Math Schoolwork Interest and Gender Gaps in Math Schoolwork Interest Favoring Girls?

In line with the prediction, there was a strong positive correlation between the total country level of math schoolwork interest and the gender difference in math schoolwork interest, $r(48) = 0.60$, 95% CI [0.39, 0.78], $p < 0.001$. However, Kuwait was diagnosed as an outlier (standardized residual > 3) (see **Figure 3**). If the outlier is excluded, the correlation is even higher, $r(47) = 0.66$.

RQ2d: Do National Levels of Math Schoolwork Interest Mediate a Negative Correlation Between National Levels of Achievement and Gender Gaps in Math Schoolwork Interest Favoring Girls?

In line with the prediction, there was a negative correlation between national levels of math achievement and gender differences in math schoolwork interest, $r(48) = -0.45$, 95% CI [−0.64, −0.22], $p = 0.001$. To examine mediation, I employed the PROCESS macro, model 4, for SPSS (Hayes, 2017), after standardizing all variables to have unit standard deviation. Results are reported in the mediation diagram in **Figure 4**, showing that the abovementioned correlation was fully mediated by the national interest in math schoolwork.

RQ3. Does the Gender Gap in Math Schoolwork Interest Account for Variance in the Gender Gap in Math Achievement Unexplained by Gender Egalitarian Values?

There was a positive correlation between gender differences in math schoolwork interest and gender differences in math achievement, $r(48) = 0.46$, 95% CI [0.15, 0.68], $p < 0.001$. Results are virtually unchanged when the analysis is restricted to the subset of 39 countries for which measures of gender egalitarian are available, $r(37) = 0.41$, 95% CI [0.16, 0.63], $p = 0.001$.

Consistent with prior research (Eriksson et al., 2020), gender egalitarian values were negatively correlated gender differences in achievement, $r(37) = -0.43$, 95% CI [−0.66, −0.15], $p < 0.001$. On their own, gender egalitarian values accounted for 19% of the country variation in gender differences in math achievement. In line with the prediction, this proportion increased to 27% when the regression additionally included gender differences in math schoolwork interest, $\beta = 0.31$, $p = 0.049$.

DISCUSSION

The present paper studied the difference between girls and boys in their interest in math schoolwork and how it varies across countries. A theory was proposed according to which national culture promoting high math achievement drives down interest in math schoolwork, but more among girls than among boys due to conformity to peer influence being stronger among girls. Moreover, I argued that gender differences in math schoolwork interest are important because they will contribute to gender differences in math achievement. In the absence of experimental data, I tested the predictions this theory makes about statistical observations in cross-sectional data, provided by TIMSS. Results were consistent with predictions, as detailed below.

First, an extremely strong negative correlation between national levels of achievement and math schoolwork interest was observed. This finding, which is well in line with prior research on the relation between national achievement levels and attitudes to math and science (Shen and Tam, 2008; Täht et al., 2014), is consistent with the hypothesis that students' interest in schoolwork is negatively influenced by the high educational norms and standards in high-achievement cultures (Van de gaer et al., 2012). That high-achievement culture may be killing students' interest is arguably a serious problem. Comparisons between high-achieving countries indicate that the problem might be solvable, however. In a study of TIMSS data from 1999 to 2003, Shen and Tam (2008) pointed out that students in Singapore, an extremely high-achieving country, nonetheless had relatively positive attitudes toward math and science. Leung (2002) made the same observation. Singapore was a positive exception also in the current study, having the highest

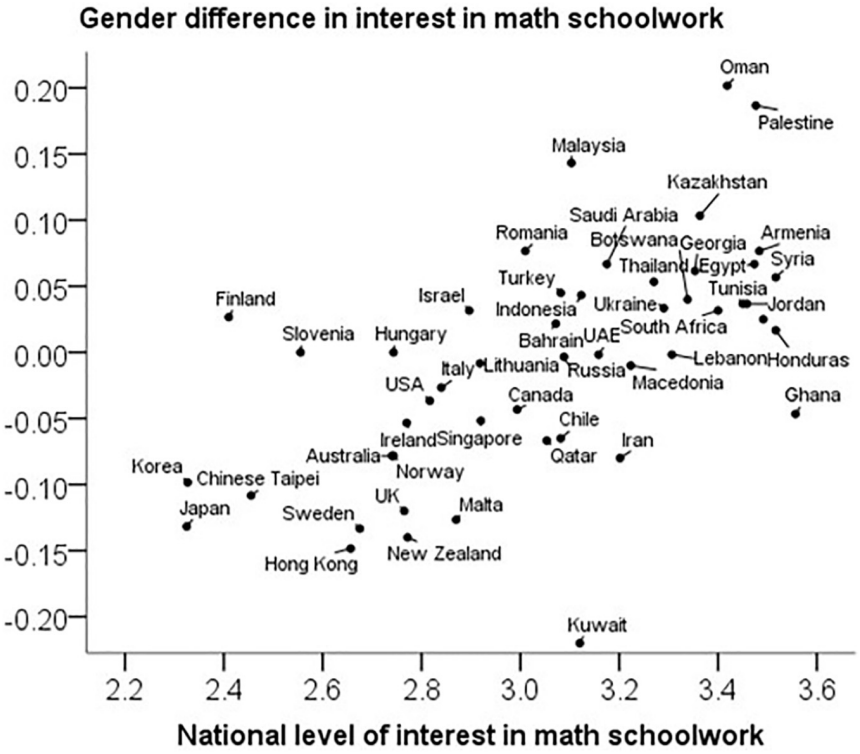

FIGURE 3 | Gender differences (favoring girls) in math schoolwork interest correlate with the national level of interest. The Pearson correlation is $r = 0.60$ (or $r = 0.66$ if the outlier Kuwait is excluded).

achievement level of all countries in the study, yet exhibiting a much higher national level of interest in math schoolwork than similarly high-achieving Korea and Japan did (**Table 1**). It would be valuable to understand whether there is some specific feature of Singapore's school system that mitigates the negative side effects of a high-achievement culture.

Second, several findings were consistent with the hypothesis that conformity to peer influence on math schoolwork interest is higher among girls than among boys. In almost all countries in the study, within-society variation in math schoolwork interest was smaller among girls than among boys, thus indicating greater female conformity. Because societies vary in their average level of interest in math schoolwork, the effect of peer pressure will vary too. Consistent with greater susceptibility to peer influence among girls, between-society variation in math schoolwork interest was larger among girls than among boys (**Figure 2**): In countries where the interest in math schoolwork was low, it tended to be especially low among girls. Similarly, in countries where the interest in math schoolwork was high, it tended to be especially high among girls. Thus, the country variation in students' interest in mathematics schoolwork was amplified among girls. The same phenomenon could be observed in terms of a positive correlation between national levels of math schoolwork interest and gender differences in math schoolwork interest favoring girls.

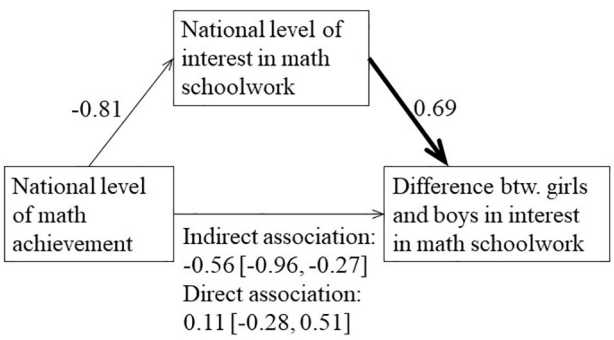

FIGURE 4 | Mediation diagram. The negative correlation between the national math achievement level and the gender difference in math schoolwork interest was mediated by the national level of math schoolwork interest. All variables were standardized prior to the mediation analysis. Bootstrapped 95% confidence intervals in brackets.

Taken together, my theory proposes a pathway in which high-achievement culture drives down schoolwork interest, which through differential peer influence creates gender gaps in interest disfavoring girls. Consistent with this pathway, I found a negative correlation between national levels of achievement and gender gaps in math schoolwork interest favoring girls, and this correlation was mediated by the national level of math schoolwork interest.

Why is it important how girls' and boys' interest in math schoolwork vary across countries? For one thing, it is

theoretically important to realize that the variation is substantial. In countries like Japan, Hong Kong, Sweden, and New Zealand, the interest level of the average girl was about 0.2 standard deviations lower than the interest of the average boy. These findings are consistent with research arguing for a fundamental gender difference in subject interest (e.g., Su et al., 2009). But this view appears to be contradicted by the finding of other societies, such as Oman, Malaysia, Palestine, and Kazakhstan, in which the gender gap is at least as wide but reversed.

Gender differences in math interest may also have real-life implications by influencing how girls achieve in mathematics relative to boys in the same country. Consistent with this hypothesis, I found that variation in the gender gap in math schoolwork interest accounts for part of the proportion of variance in the gender gap in math achievement that is not explained by variation in gender egalitarian values (Eriksson et al., 2020).

This study is an example of the benefits of using big data from large-scale assessments of student achievement to examine phenomena in educational psychology. A limitation, inherent in the reliance on cross-sectional data, is that directions of causality are not established. The findings are consistent with the proposed theory, but they could also have arisen from other mechanisms. It is helpful to consider what these alternative mechanisms could be. With respect to the strong negative correlation between national levels of achievement and schoolwork interest, it seems implausible that it would arise from low interest levels having a positive effect on achievement levels. Following Van de gaer et al. (2012), I proposed that high educational norms and standards have lower interest as an undesired side effect. However, there might be something else going on and perhaps more detailed insights into the abovementioned differences between Singapore and its East Asian neighbors could shed more light on this.

Similar reasoning applies to the amplification among girls of national variation in schoolwork interest. I proposed that this arises from differential conformity to peer influence, but it cannot be excluded that there is some alternative societal factor that causes girls' interest levels to be more extreme than the interest levels of boys. An interesting possibility for future research would be for large-scale assessments to provide some direct measures of peer influence (see also Eriksson et al., 2020).

The idea of conceiving of high-achievement culture as a factor behind gender differences has a precedent. In a study of PISA data, Mann and DiPrete (2016) found that the national achievement level correlated with gender differences in academic self-concept and STEM aspirations. However, they did not examine the female amplification account, that is, whether these effects were mediated by national levels of academic self-concept and STEM aspirations. Future research should examine the scope of female amplification as a mechanism behind gender differences in various beliefs and attitudes.

To conclude, the present study has contributed to scientific understanding of gender differences in interest in mathematics schoolwork by, first, proposing a theory of why such gender differences would arise and vary across countries, and second, testing several theoretical predictions in a large cross-national dataset. Results were consistent with both key components of the theory: high-achievement culture may be detrimental to interest in schoolwork and this effect may be amplified among girls due to their higher conformity to peer influence. These positive findings motivate further study of the validity and scope of the proposed mechanisms.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://osf.io/dwk8h/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

KE performed the statistical analysis and wrote the manuscript.

FUNDING

This research was supported by the Swedish Research Council (Grant No. 2014-2008) and the Knut and Alice Wallenberg Foundation (Grant No. 2015.0005).

ACKNOWLEDGMENTS

Ksenia Startseva-Lora provided helpful preliminary analyses.

REFERENCES

Artelt, C., Baumert, J., Julius-McElvany, N., and Peschar, J. (2003). *Learners for life: Student approaches to learning. Results from PISA 2000*. Paris: Organisation for Economic Co-operation and Development.

Berndt, T. J., and Keefe, K. (1995). Friends' influence on adolescents' adjustment to school. *Child Dev.* 66, 1312–1329. doi: 10.1111/j.1467-8624.1995.tb00937.x

Beutel, A. M., and Marini, M. M. (1995). Gender and values. *Am. Sociol. Rev.* 60, 436–448. doi: 10.2307/2096423

Bond, R., and Smith, P. B. (1996). Culture and conformity: a meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychol. Bull.* 119, 111–137. doi: 10.1037/0033-2909.119.1.111

Brand, C. O., Brown, G. R., and Cross, C. P. (2018). Sex differences in the use of social information emerge under conditions of risk. *PeerJ* 6:e4190. doi: 10.7717/peerj.4190

Cross, C. P., Brown, G. R., Morgan, T. J., and Laland, K. N. (2017). Sex differences in confidence influence patterns of conformity. *Br. J. Psychol.* 108, 655–667. doi: 10.1111/bjop.12232

Eagly, A. H., and Carli, L. L. (1981). Sex of researchers and sex-typed communications as determinants of sex differences in influenceability: a meta-analysis of social influence studies. *Psychol. Bull.* 90, 1–20. doi: 10.1037/0033-2909.90.1.1

Efferson, C., and Vogt, S. (2018). Behavioural homogenization with spillovers in a normative domain. *Proc. R. Soc. B Biol. Sci.* 285:20180492. doi: 10.1098/rspb.2018.0492

Else-Quest, N. M., Hyde, J. S., and Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychol. Bull.* 136, 103–127. doi: 10.1037/a0018053

Eriksson, K., Björnstjerna, M., and Vartanova, I. (2020). The relation between gender egalitarian values and gender differences in academic achievement. *Front. Psychol.* 11:236. doi: 10.3389/fpsyg.2020.00236

Frenzel, A. C., Goetz, T., Pekrun, R., and Watt, H. M. (2010). Development of mathematics interest in adolescence: influences of gender, family, and school context. *J. Res. Adolesc.* 20, 507–537. doi: 10.1111/j.1532-7795.2010.00645.x

Ghasemi, E., and Burley, H. (2019). Gender, affect, and math: a cross-national meta-analysis of Trends in International Mathematics and Science Study 2015 outcomes. *Large Scale Assess. Educ.* 7:10. doi: 10.1186/s40536-019-0078-1

Guiso, L., Monte, F., Sapienza, P., and Zingales, L. (2008). Culture, gender, and math. *Science* 320, 1164–1165. doi: 10.1126/science.1154094

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York, NY: Guilford publications.

House, R. J., Hanges, P. J., Javidan, M., Dorfman, P. W., and Gupta, V. (2004). *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. Thousand Oaks: SAGE Publications.

Hyde, J. S., Fennema, E., Ryan, M., Frost, L. A., and Hopp, C. (1990). Gender comparisons of mathematics attitudes and affect: a meta-analysis. *Psychol. Women Q.* 14, 299–324. doi: 10.1111/j.1471-6402.1990.tb00022.x

Kindermann, T. A. (2016). "Peer group influences on students' academic motivation," in *Handbook of Social Influences in School Contexts: Social-Emotional, Motivation, and Cognitive Outcomes*, eds G. B. Ramani, and K. R. Wentzel (Abingdon: Routledge), 31–47.

Leung, F. K. (2002). Behind the high achievement of East Asian students. *Educ. Res. Eval.* 8, 87–108. doi: 10.1076/edre.8.1.87.6920

Leung, F. K. (2014). What can and should we learn from international studies of mathematics achievement? *Math. Educ. Res. J.* 26, 579–605. doi: 10.1007/s13394-013-0109-0

Lippa, R. A. (1998). Gender-related individual differences and the structure of vocational interests: the importance of the people–things dimension. *J. Pers. Soc. Psychol.* 74, 996–1009. doi: 10.1037/0022-3514.74.4.996

Mann, A., and DiPrete, T. A. (2016). The consequences of the national math and science performance environment for gender differences in STEM aspiration. *Sociol. Sci.* 3, 568–603. doi: 10.15195/v3.a25

Martin, M. O., Mullis, I. V. S., and Hooper, M. (eds) (2016). *Methods and Procedures in TIMSS 2015*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Mikk, J., Krips, H., Säälik, Ü, and Kalk, K. (2016). Relationships between student perception of teacher-student relations and PISA results in mathematics and science. *Int. J. Sci. Math. Educ.* 14, 1437–1454. doi: 10.1007/s10763-015-9669-7

Preckel, F., Goetz, T., Pekrun, R., and Kleine, M. (2008). Gender differences in gifted and average-ability students: comparing girls' and boys' achievement, self-concept, interest, and motivation in mathematics. *Gifted Child Q.* 52, 146–159. doi: 10.1177/0016986208315834

Reilly, D., Neumann, D. L., and Andrews, G. (2019). Investigating gender differences in mathematics and science: results from the 2011 trends in mathematics and science survey. *Res. Sci. Educ.* 49, 25–50. doi: 10.1007/s11165-017-9630-6

Riegle-Crumb, C., Farkas, G., and Muller, C. (2006). The role of gender and friendship in advanced course taking. *Sociol. Educ.* 79, 206–228. doi: 10.1177/003804070607900302

Shen, C., and Tam, H. P. (2008). The paradoxical relationship between student achievement and self-perception: a cross-national analysis based on three waves of TIMSS data. *Educ. Res. Eval.* 14, 87–100. doi: 10.1080/13803610801896653

Su, R., Rounds, J., and Armstrong, P. I. (2009). Men and things, women and people: a meta-analysis of sex differences in interests. *Psychol. Bull.* 135, 859–884. doi: 10.1037/a0017364

Täht, K., Must, O., Peets, K., and Kattel, R. (2014). Learning motivation from a cross-cultural perspective: a moving target? *Educ. Res. Eval.* 20, 255–274. doi: 10.1080/13803611.2014.929009

Van de gaer, E., Grisay, A., Schulz, W., and Gebhardt, E. (2012). The reference group effect: an explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *J. Cross Cult. Psychol.* 43, 1205–1228. doi: 10.1177/0022022111428083

Welzel, C. (2013). *Freedom Rising*. Cambridge, MA: Cambridge University Press.

Wentzel, K. R., and Muenks, K. (2016). "Peer influence on students' motivation, academic achievement, and social behavior," in *Handbook of Social Influences in School Contexts: Social-Emotional, Motivation, and Cognitive Outcomes*, eds G. B. Ramani, and K. R. Wentzel (Abingdon: Routledge), 13–30.

Wilcox, R. (2015). Comparing the variances of two dependent variables. *J. Stat. Distrib. Applicat.* 2:7. doi: 10.1186/s40488-015-0030-z

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

ORIGINAL RESEARCH
published: 27 November 2020
doi: 10.3389/fpsyg.2020.575167

An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques

*Adriana Gamazo and Fernando Martínez-Abad**

Research Institute on Educational Sciences, University of Salamanca, Salamanca, Spain

International large-scale assessments, such as PISA, provide structured and static data. However, due to its extensive databases, several researchers place it as a reference in Big Data in Education. With the goal of exploring which factors at country, school and student level have a higher relevance in predicting student performance, this paper proposes an Educational Data Mining approach to detect and analyze factors linked to academic performance. To this end, we conducted a secondary data analysis and built decision trees (C4.5 algorithm) to obtain a predictive model of school performance. Specifically, we selected as predictor variables a set of socioeconomic, process and outcome variables from PISA 2018 and other sources (World Bank, 2020). Since the unit of analysis were schools from all the countries included in PISA 2018 ($n = 21{,}903$), student and teacher predictor variables were imputed to the school database. Based on the available student performance scores in Reading, Math, and Science, we applied k-means clustering to obtain a categorized (three categories) target variable of global school performance. Results show the existence of two main branches in the decision tree, split according to the schools' mean socioeconomic status (SES). While performance in high-SES schools is influenced by educational factors such as metacognitive strategies or achievement motivation, performance in low-SES schools is affected in greater measure by country-level socioeconomic indicators such as GDP, and individual educational indicators are relegated to a secondary level. Since these evidences are in line and delve into previous research, this work concludes by analyzing its potential contribution to support the decision making processes regarding educational policies.

Keywords: educational data mining, school performance, large-scale assessment, non-cognitive outcomes, socioeconomic status, decision tree, academic achievement

OPEN ACCESS

Edited by:
Ching Sing Chai,
The Chinese University of Hong Kong, China

Reviewed by:
Wim Van Den Noortgate,
KU Leuven Kulak, Belgium
Bo Ning,
Shanghai Normal University, China

***Correspondence:**
Fernando Martínez-Abad
fma@usal.es

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *22 June 2020*
Accepted: *09 November 2020*
Published: *27 November 2020*

Citation:
Gamazo A and Martínez-Abad F (2020) An Exploration of Factors Linked to Academic Performance in PISA 2018 Through Data Mining Techniques. Front. Psychol. 11:575167. doi: 10.3389/fpsyg.2020.575167

INTRODUCTION

The emergence of international large-scale assessments (ILSA) in the past two decades, together with their cyclic nature, have consistently provided educational researchers with large databases containing diverse types of variables (student performance and background, school practices and processes, etc.). Assessment schemes such as the Programme for International Student Assessment (PISA) from the Organisation for Cooperation and Economic Development (OECD), or the Trends in International Mathematics and Science Study (TIMSS) and the Progress in International Reading Literacy Study (PIRLS), both conducted by the International Association for the Evaluation of

Educational Achievement (IEA), have had a noticeable impact on the development of educational research in past years (Gamazo et al., 2016). But the great relevance of large-scale assessments is not circumscribed to educational research; some authors also highlight the great impact that PISA results have on national policies and practices in the field of education (Lingard et al., 2013). However, it has been observed that educational policies are usually influenced by the reports and analyses elaborated directly by the OECD, because these are the first ones presented to the public after a given PISA wave (Wiseman, 2013) and since these analyses can be somewhat limited considering the vast array of variables that PISA offers (Jornet, 2016), there is a certain responsibility for educational researchers to delve deeper into the databases and find relationships among variables and conclusions that might not be offered by the OECD reports in order to enrich the political debate around the topic.

Secondary analyses of PISA data can be performed through the use of different methodologies. One of the most common ones is multilevel regression analysis, given that it allows researchers to account for the variability at the level of students and schools at the same time (Willms, 2010; Gamazo et al., 2018). Other authors have opted for different methods, such as Structural Equation Modeling (Acosta and Hsu, 2014; Barnard-Brak et al., 2018) or ANCOVA (Smith et al., 2018; Zhu and Kaiser, 2019). Additionally, thanks to the emergence of big data, new possibilities in the statistical analysis of all types of databases have appeared in recent years. Namely, data mining has appeared in the past few years as one of the emerging techniques to analyse PISA data (Liu and Whitford, 2011; Tourón et al., 2018; Martínez-Abad, 2019; She et al., 2019), although it is a less-explored analysis method.

The data mining approach seeks to detect key information in huge amounts of data (Witten et al., 2016). Thus, data mining algorithms are specifically defined to be used in extensive databases, like those from large-scale assessments. These kinds of techniques build and validate models directly from the empirical data, without the use of either theoretical distributions or hypothesis tests (Xu, 2005), and allow the joint inclusion of both categorical and numerical variables. That is why, unlike the inferential and multivariate approaches, the models obtained through data mining algorithms are inductive, that is, computed exclusively from the information contained in the database. This way, data mining techniques can help to identify the main factors linked to academic performance and its interactions under a new framework, allowing researchers to reassess and refine existing theoretical models.

However, it is worth noting that the power of data mining resides in the production of exploratory studies to identify potentially significant relationships within large amounts of data, but follow-up confirmatory studies would be necessary in order to consolidate findings (Liu and Ruiz, 2008). Additionally, data mining presents other weaknesses that researchers must take into consideration, such as possible misinterpretations due to human judgment on the findings, an information overload leading to the construction of highly complex relationship systems, or the difficulty to interpret data mining results on the part of educational professionals (Papamitsiou and Economides, 2014).

Thus, the main aim of this paper is to take advantage of the benefits offered by data mining techniques in order to explore the influence of different types of student, school, and country variables on student performance in reading, science and mathematics in PISA 2018.

Research on Factors Associated With Student Performance

Although the study of variables associated with student performance has historically been a concern in educational research, the publication of the Coleman (1966), together with the following discussion about the central role of socioeconomic variables and the relevance of school practices and policies, started a research line whose relevance has spanned more than five decades and is still highly relevant today. While there are many different sources of data to conduct studies on the variables related to student performance, large-scale assessments have established themselves as a valuable source due to the large volume of variables and observations that they offer to researchers.

Educational variables have traditionally been classified as input and output, later expanded to context-input-process-output, likening the educational process to economic models (Scheerens and Bosker, 1997). However, more recently some authors have suggested to rearrange these categories to better fit ILSA structures, instead choosing to focus around content areas such as school and student background, teaching and learning processes, school policies and education governance, and education outcomes (Kuger and Klieme, 2016). Thus, this section will provide an overview of the scientific evidence of the relevance of PISA variables in relation to secondary education student achievement, following the latter categorisation.

Student context factors are among the most widely studied variables in achievement research. Factors such as socioeconomic status (SES), immigration status, age/grade, attendance to early childhood education (ISCED 0), or grade repetition have been consistently proven to be highly related to student performance (Karakolidis et al., 2016; Pholphirul, 2016; Gamazo et al., 2018). Gender constitutes a special case within this category, since its influence can favor male or female students depending on the competence under study (generally boys outperform girls in math and science, and the opposite is true for reading), also with varying degrees of intensity (Gamazo et al., 2018). At school level, one of the only factors that seem to generate consensus about its positive relationship with performance is mean SES (Asensio-Muñoz et al., 2018; Gamazo et al., 2018). Other school variables, like ownership, resources, student–teacher ratio, or size, have yielded diverse results. There are studies that find no significant relationship between these variables and student performance (Gamazo et al., 2018), some that find positive relationships between school size, resources or ratio and performance (Kim and Law, 2012; Tourón et al., 2018) and others with contradictory results, depending on the country and the PISA wave analyzed. Ownership, for example, has yielded significant results both in favor of public (Kim and Law, 2012; Chaparro Caso López and Gamazo, 2020) and private schools (Acosta and Hsu, 2014).

Although the aggregation bias is a widely studied effect (Fertig and Wright, 2005), several studies based on a multivariate methodological framework have aggregated student data to estimate school indices (Brunello and Rocco, 2013; Gamazo et al., 2018; Avvisati, 2020).

Lastly, although ILSAs do not gather information at system level, some studies incorporate these kinds of factors when comparing several countries, and it has been found that background variables related to a country's affluence and quality of life, like GDP per capita or the Human Development Index (HDI), are closely related to student performance (Täht et al., 2014; Rodríguez-Santero and Gil-Flores, 2018). However, the inclusion of country level variables is relatively uncommon in the literature.

The category of teaching and learning processes encompasses both student and school level variables related to school climate, teaching methodologies, learning time in and out of the school or teacher support (Kuger and Klieme, 2016). While there seems to be some consensus on the positive relationship between student performance and process variables such as climate, learning time or teacher support (Lazarević and Orlić, 2018; Tourón et al., 2018; She et al., 2019), the study of other factors, like inquiry-based teaching practices, yields mixed results (Gil-Flores and García-Gómez, 2017; Tourón et al., 2018).

School policies and educational governance are a less-studied field within large-scale assessment research, although there is some evidence on the positive effect on student achievement of variables like educational leadership, teacher participation in decision-making processes, parental involvement or school autonomy (Drent et al., 2013; Cordero-Ferrera, 2015; Rodríguez-Santero and Gil-Flores, 2018; Tourón et al., 2018).

The last category of variables according to their content area is education outcomes. Student achievement is not the only outcome that education systems should be striving to improve; on the contrary, non-cognitive outcomes like motivation, metacognitive strategies, self-efficacy or domain-related beliefs (Farrington et al., 2012; Khine and Areepattamannil, 2016) constitute a fundamental element when assessing the quality of education systems (Kyriakides and Creemers, 2008; OECD, 2019). Non-cognitive outcomes are usually studied alongside cognitive results, with authors intending to discover the possible relationships between the two kinds of variables. Some of these factors, such as self-efficacy, motivation toward achievement or task mastery, expected occupational status or domain enjoyment have been found to be positively related to student performance (Aksu and Güzeller, 2016; Tourón et al., 2018; She et al., 2019). Metacognitive strategies like summarizing, understanding and remembering, or assessing information have also been positively associated with the students' reading skills (Cheung et al., 2014, 2016), and they are, in fact, an integral part in some theoretical models that aim to explain student performance through its associated factors, such as the one proposed by Farrington et al. (2012). There are some other variables that have been proven to have a negative effect on student achievement. Such is the case of truancy, which is linked to low levels of achievement, and this relationship is especially relevant in students with low SES (Rutkowski et al., 2017).

Given that the studies reviewed in this section use diverse research methods and include a great variety of different variables, it is not possible to confidently gauge which variables are more relevant overall or have more impact on student performance; instead, only the statistical significance and sign of the relationship (positive of negative) can be reported here.

Educational Data Mining

Educational data mining (EDM) constitutes an analytical process that enables researchers to turn large amounts of raw data into useful information about different aspects of educational policies and practices (Romero and Ventura, 2010). Although some previous works exist, the main development of this discipline occurred in the first decade of the 21st century, when most of the international conferences and workshops on the subject were first celebrated, and its use has kept on growing in popularity over the past decade (Romero et al., 2010; Tufan and Yildirim, 2018).

Educational data mining is not a method in itself, but rather a group of techniques that share some similarities in terms of procedures and goals. Although there are many different approaches that fall within the scope of data-driven educational research, the main ones, according to their goal, are prediction, relationship mining, and structure discovery (Baker and Inventado, 2014).

The main aim of the prediction approach is to help researchers infer information about a certain variable of interest from a set of other variables (predictors), and also to explore which constructs in a dataset have a relevant role in predicting another (Baker and Inventado, 2014). Prediction can be achieved through two types of techniques: classification and regression, depending on the nature of the predicted variable (categorical or continuous, respectively).

Relationship mining aims to find the strongest relationships among variables in datasets with large amounts of data without a prior designation of criterion or predictor variables. This can be done through different techniques such as association rules or correlation mining (Baker, 2010).

Lastly, structure discovery methods are employed to find natural groupings between data points or variables without a *priori* assumptions of what the analysis should find. The main techniques within this approach are clustering and factor analysis, which look to group together data points/variables that are more similar to those on their group than to those on other groups (Baker and Inventado, 2014).

As we already pointed out initially, EDM-based approaches present some differences from the use of more traditional statistical analysis methods which can be useful in the study of factors linked with performance in large-scale assessments. In this sense, EDM algorithms are being considered by some authors as a more effective and reliable alternative in many aspects than classic inferential and multivariate statistics for the analysis of massive databases (Martínez-Abad, 2019). Moreover, data mining enables the collection of non-trivial information in massive data sets without starting from pre-established models, with minimal human intervention and without raising previous assumptions about the distribution of the data (Xu, 2005).

EDM and Large-Scale Assessments

Educational data mining can be used to study many characteristics of the teaching-learning process, such as student behavior and/or performance, dropout and retention rates, feedback provided to students, or teacher and student reflection and awareness of the learning process (Papamitsiou and Economides, 2014). Within the field of performance prediction, most of the studies found are conducted at a higher education level, and in virtual learning environments, MOOCs or computer-based learning (Papamitsiou and Economides, 2014). A plausible reason for this is that it is easier to gather large amounts of data from online or computer-based courses given that they allow for the registration of all kinds of participation and interaction data, and these courses are more frequent in Higher Education than in School Education levels.

However, large-scale assessments conducted at a Secondary Education level, such as PISA or TIMSS eighth grade, provide a great opportunity to apply EDM approaches with a less-explored student population. Although the PISA assessment contains static, limited, and structured data, Andreas Schleicher (2013), Director of Education of the OECD and coordinator of PISA, did not hesitate in considering these assessments as big data in education. Other authors have made similar statements, considering the OECD as one of the main providers of system-level big data in the field of education (Sahlberg, 2017).

The use of EDM approaches with large scale assessments is usually focused on predicting student performance in one or more competences (math, reading, and science) by using a set of predicting variables such as student and school background, educational practices or non-cognitive student outcomes, in order to find out which of these variables are more strongly related to performance and thus can serve as better predictors. The past decade has seen the publication of many research works that use EDM techniques for performance prediction. Although there is some diversity in terms of the particular techniques used, the most popular seem to be decision trees and their different algorithms, such as Classification and Regression Trees (CART) (Asensio-Muñoz et al., 2018; Gabriel et al., 2018; She et al., 2019), Chi-squared Automatic Interaction Detection (CHAID) (Aksu and Güzeller, 2016; Asensio-Muñoz et al., 2018; Tourón et al., 2018) or other algorithms like C4.5 (Liu and Ruiz, 2008; Oskouei and Askari, 2014; Martínez-Abad, 2019) or J48, which is another form of the C4.5 algorithm (Aksu and Güzeller, 2016; Martínez-Abad and Chaparro-Caso-López, 2016; Kılıç Depren et al., 2017; Martínez-Abad et al., 2020). Some of these studies aggregate student variables to school level (e.g., Martínez-Abad, 2019), however, there are not, to our knowledge, any basic studies on the effects of the aggregation bias on the computation of data mining models. Another common technique when dealing with student performance data is clustering. This process is usually used to find out which is the best way to group students, schools, or countries according to the similarities in their performance levels, often aiming to conduct a subsequent prediction analysis with said clusters as a criterion variable (Rodríguez-Santero and Gil-Flores, 2018; Soh, 2019). It is worth noting that all the aforementioned studies are focused on single-country analyses.

In this paper, clustering techniques (*k*-means) are used first in order to group schools from 78 countries according to their mean performance level in PISA 2018, and then a prediction analysis is performed in order to discover which country, school, and student variables better predict school performance.

MATERIALS AND METHODS

From a purely quantitative approach (Johnson et al., 2007), the main objective of this study is to analyze factors linked to academic performance in large-scale assessments mainly using data mining techniques (Witten et al., 2016), specifically decision trees. To address this goal, secondary data analyses were conducted with PISA 2018 databases (OECD, 2019), where decision trees (C4.5 algorithm) were built to obtain a predictive model of school performance. In this sense, in order to get an integrated and comprehensive model, student and teacher data were aggregated in the schools' database. In addition, some socio-economic and educational variables at the country level were added to the final database.

Thus, this study follows a non-experimental design based on transversal data (secondary panel data from the PISA 2018 assessment).

Research Questions

In line with the stated goal, this study seeks to answer four main research questions:

- Is it possible to model school performance by using decision trees and obtain acceptable levels of fit? Which type of factors presents the highest explanatory levels: socio-economic country variables, school indicators and factors, or non-cognitive educational outcomes?
- Do country-level socioeconomic indicators have a relevant impact on performance? Which wealth indicators are more relevant: gross or adjusted?
- Which school indicators have a greater contribution to explain performance? Is their impact conditioned by country-level variables?
- What are the non-cognitive educational outcomes with the greatest contribution to explain performance? Is their impact conditioned by country-level variables?

Participants

The population of this study were 15-year-old students, teachers, and schools from the countries participating in PISA 2018. Thereby, the initial sample of this research was the entire set of schools, teachers and students included in PISA 2018. An initial review of the data revealed that the Spanish and Vietnamese samples did not include the full scores of the 3 main domains assessed in PISA (science, mathematics, and reading), therefore both countries were removed from the final database.

Thus, the sample was composed of 20,663 schools from 78 countries, and the aggregated data of 570,684 15-year-old students and 85,746 secondary education teachers.

Variables and Instruments

The full study was carried out using the instruments developed by the OECD for the 2018 PISA wave, which can be grouped in two categories according to their content:

- Context questionnaires: In PISA 2018, different context questionnaires were answered by school principals, teachers, students, and their families. Context questionnaires include a set of items with a wide range of sociodemographic, economic, and educational information related to student outcomes (OECD, 2019). Most of the included items are grouped into constructs referring to different issues: school organization and governance, Teaching and learning factors, student and family background, and non-cognitive/metacognitive factors. The scales obtained from these constructs were calculated using two parameter item-response model. Specifically, PISA uses the Generalized Partial Credit Model, appropriate for working with ordinal items (Martínez-Abad et al., 2020).
- Performance tests in reading, mathematics, and science domains: performance tests include an item bank, and each student is presented with only a fraction of those items. To account for this item disparity, Item Response Theory (IRT) techniques were applied to estimate the ability of the students in each domain. Consequently, the PISA 2018 data does not include a point estimate of a student's ability in each competence, but rather 10 plausible values that account for the variability in scores depending on the different sets of items available.

Therefore, to define a single criterion variable in this study, it was necessary to apply grouping techniques. Specifically, k-means clustering was used to group schools according to their average performance levels in science, mathematics, and reading. Following previous studies (Shulruf et al., 2008; Zhang and Jiao, 2013; Yao et al., 2015), 3 clusters were obtained: low performance, medium performance, and high performance.

We computed 10 different models, one for each of the 10 plausible values (PV) available, obtaining a final criterion variable with 3 groups:

- Low performance: set of schools classified within the low performance cluster in each of the 10 models.
- High performance: set of schools classified within the high performance cluster in each of the 10 models.
- Medium performance: all other schools.

All variables with high levels of missing values (more than 80%) were removed. In this sense, even though PISA 2018 databases included a sample of teachers only in 19 of the 80 participating countries (including Spain), teacher variables were maintained. This decision was made due to the high level of response of the teacher variables in these countries (in all the teacher variables the general level of missing values is less than 80%), to the construction procedure of the decision trees (based on the consecutive division of the sample to build the model) and the handling of missing values in the C4.5 algorithm (which is not based on data imputation of point values, as noted below). Thus, the predictor variables included in the final database were:

- All the derived variables (scales) available in PISA 2018 from the student, teacher, and school questionnaires.
- All the school-level indicators: school and class size; Ownership; % of students with special needs, with low SES and immigrants; % of girls and repeating students; job and academic expectations of students; Language at home; Additional instruction; Students' SES; Learning time at school; Attendance to ISCED 0; Average teachers' age; % of female teachers; Teacher training and development; Teacher employment time; Student–Teacher ratio; Computer-Student ratio.

In addition, the following socioeconomic country indicators were included: Gross Domestic Product (GDP), GDP adjusted by Purchasing Power Parity (PPP), GDP per capita, and GDP (PPP) per capita (International Monetary Fund, 2019); Human Development Index (HDI) (United Nations Development Programme, 2019); and expenditure on education as a percentage of GDP (World Bank, 2020). All the variables included in the study, along with a brief description, can be found in the Appendix.

Procedure and Data Analysis

According to the technical recommendations (OECD, 2017), school base weights provided in the PISA 2018 database were used in all statistical analyses (student weights were also used when aggregating student variables to school level). After filtering the database, obtaining the criterion variable by using the indicated clustering procedures, and implementing an initial examination of the sample distribution, the decision trees were calculated.

A decision tree includes a set of nested rules, whose graphic representation forms an inverted tree. Decision trees are made up of nodes (which contain the selected predictor variables), branches (which indicate the rules) and leaves (terminal nodes). Thus, trees start with an initial node, which includes the predictor variable with a higher information gain score, and end with a leaf or terminal node, which includes the subsample that complies with all the rules formulated from the initial node to that leaf. Finally, it is important to note that a predictor variable can be included in several tree nodes simultaneously.

The algorithm implemented in the estimation of the final model was C4.5 (Quinlan, 1992). Specifically, we used an extension of C4.5 implemented in the software Weka 3.8 called J48 (Witten et al., 2016). Given its simplicity and characteristics, the use of this algorithm is widespread in Educational Data Mining (Martínez-Abad, 2019). C4.5 and its derived algorithms allow the use of both categorical and numerical predictor variables, and the use of the information gain score (index of the relevance of the predictor variables in a sample that goes through a single branch) to select the predictor variable included in each cut of the tree.

The C4.5 algorithm includes a specific procedure to manage missing data with a probabilistic approach. This approach, which

is different from the main imputation methods (e.g., Mean, hot/cold deck, regression, and interpolation), seems to perform better in large databases with a great percentage of missing values (Grzymala-Busse and Hu, 2001), as it is common in large-scale assessments. J48 manages missing data in any predictor variable selected in a node by assigning to each derived branch "a weight proportional to the number of training instances going down that branch, normalized by the total number of training instances" (Witten et al., 2016, p. 230). If another predictor variable with missing values is included in any following nodes of the tree, this procedure is replicated. These instances contribute to the terminal nodes in the same way as the other instances, with their estimated proportional weight.

Initially, we calculated the baseline model, which is quite similar to the null models used in multivariate analysis, since it calculates the fit of a model without predictor variables. Specifically, the baseline model provides the base accuracy level, which is used as a reference to assess the fit of the final model (Witten et al., 2016).

The baseline model was followed by the estimation of the final decision tree that included the predictor variables. In accordance with previous studies (Martínez-Abad and Chaparro-Caso-López, 2016; Martínez-Abad, 2019), the size of the tree was restricted to a maximum of 20 terminal nodes to facilitate the interpretation and to limit the possibility of overfitting of the final model. Although specialized literature recommends the use of a validation procedure in the estimation of the final model (Witten et al., 2016), we included information obtained from both the training set and the 10-folds cross-validation procedure, which facilitates the analysis of overfitting problems. This method implements these consecutive steps (Witten et al., 2016; Martínez-Abad et al., 2020):

- First, the full sample (of size n) is divided in 10 approximately equal groups.
- These divisions are used now to obtain pairs of sub-samples. Each pair of sub-samples is composed of both a sub-sample of size n/k and other sub-sample with the remaining sample, of size $n - (n/k)$. In this process, the 10 possible pairs of different sub-samples are calculated.
- For each pair, the biggest sub-sample will be used as training set (to build the initial model) and the sized n/k sample will be used as test set (to check the accuracy of the training set model). This procedure will be executed 10 times independently in any of the 10 obtained pairs.
- Finally, the error estimates obtained in all 10 models are averaged to obtain the fit indices and an overall error estimate.

To assess the model fit, the following fit indices were considered (Witten et al., 2016):

- Overall model Accuracy: proportion of the total instances predicted as positive that are correctly classified.
- True Positive rate (TP): proportion of the total number of positive instances that are correctly identified.
- Area under the Receiver Operating Characteristic curve (AUROC): reports on the ability of the model to distinguish between classes. Formally, it can be defined as the probability that the model ranks a randomly chosen positive instance above a randomly chosen negative instance.
- Kappa index: level of agreement between the classification proposed by the model and the true instance classes.
- Root Relative Squared Error (RRSE): proportion of the differences between classes predicted by the model and the true instance classes.

RESULTS

K-Means Clustering

Table 1 shows the final cluster centers (variable means) in all the computed models. Regardless of the model or the predictor variable, results consistently show high scores in cluster 2, medium scores in cluster 1 and low scores in cluster 3. The contribution of the 3 variables used is highly significant ($p < 0.001$) in all models.

After obtaining the groups of schools based on clustering, schools were allocated in the following groups: *high performance* (school grouped in cluster 2 in all 10 models), *low performance*

TABLE 1 | Final cluster centers in 10 *K*-means cluster models.

	Mathematics			Reading			Science		
	Cl 1	Cl 2	Cl 3	Cl 1	Cl 2	Cl 3	Cl 1	Cl 2	Cl 3
PV 1	418.35	517.40	330.93	419.12	518.51	328.85	423.86	518.42	347.23
PV 2	418.29	517.01	332.39	419.50	517.91	329.47	424.34	519.13	347.53
PV 3	418.34	515.84	330.85	419.85	518.43	330.28	424.76	518.55	346.30
PV 4	423.39	518.86	332.96	423.12	520.66	332.60	428.40	521.26	349.69
PV 5	412.66	512.98	329.25	413.46	513.72	325.89	418.86	514.21	345.06
PV 6	415.40	514.83	330.57	415.92	515.48	329.08	421.96	516.25	347.73
PV 7	417.37	517.91	332.48	419.03	518.36	328.87	424.58	519.40	346.85
PV 8	410.72	511.30	328.14	411.14	512.71	325.35	416.60	513.16	345.27
PV 9	416.86	514.54	331.26	415.35	515.97	327.99	420.91	517.30	345.69
PV 10	417.68	516.02	330.60	418.10	516.86	329.55	423.70	518.16	348.15

TABLE 2 | Final distribution of schools based on clustering models.

	Not weighted		Weighted	
	Freq.	**%**	**Freq.**	**%**
High	7,888	38.17	106,610.51	21.25
Medium	3,087	14.94	157,005.77	31.29
Low	9,688	46.89	238,146.47	47.46
Total	20,663	100.00	501,762.75	100.00

(school grouped in cluster 3 in all 10 models) *medium performance* (schools not included in the above groups). The final distribution of schools (**Table 2**), accounting for the school sample weights, shows approximately 10% more low performance schools than high performance schools.

Input Variables: Country and School Characteristics

All of the country level variables explored showed significant differences when comparing school groups according to performance (**Table 3**). High performance schools tend to be located in countries with greater levels of GDP (both nominal and adjusted by purchasing power parity and per capita), with greater expenditure on education (% GDP) and greater levels of HDI. The eta-squared (η^2) effect size scores indicate that HDI and GDP per capita (PPP) are the variables that provide the greatest explanation of the level of performance (in terms of percentage of variance explained). Thus, results show that higher levels of socio-economic wealth, equality and social development promote better levels of academic performance in schools and society.

Similarly, school characteristics have a highly significant relationship with school performance (**Table 4**). While schools with greater average SES and percentage of migrant students are related with high performance, higher proportions of repeating students, together with larger school sizes and teacher-student ratios are related with low performance, with school SES and percentage of repeating students showing the largest effect sizes.

Table 5 shows the bivariate distribution by school ownership and performance level. While the distribution of public schools is quite similar in high, medium and low performance schools, private independent and government dependent schools are distributed differently. Although both variables can be considered dependent ($\chi^2 = 16{,}998.42$; $p<0.001$), the relationship is weak (Cramér's $V = 0.134$).

Decision Tree

The size of the computed decision tree was 36 branches and 20 final leaves. Compared with the baseline model, the average fit obtained in the Training set and Cross-Validation models reached good levels (**Table 6**): increases in both correctly classified instances (20%) and model accuracy (50%) and an almost 20% reduction in relative error. Moreover, considering that the baseline model classified all the schools as medium performance, levels of accuracy of classified instances in high and low performance clusters could be considered highly satisfactory.

Table 7 shows the confusion matrix obtained in both the full training set and Cross-Validated models. It should be noted that, among the incorrectly classified instances in high and low

TABLE 3 | Country statistics by school performance level.

	GDP*	**GDP (PPP)***	**GDP pc****	**GDP (PPP) pc****	**% GDP Ed.**	**HDI**
High	5.110 (7.45)	6.596 (8.47)	30.695 (22.42)	39.951 (17.60)	4.544 (1.21)	0.861 (0.07)
Medium	1.921 (3.94)	3.121 (4.03)	16.147 (16.65)	26.958 (15.39)	4.537 (1.04)	0.800 (0.07)
Low	1.144 (1.79)	2.627 (2.06)	9.027 (8.94)	19.235 (10.96)	4.349 (1.13)	0.755 (0.05)
F (p.)***	26,826 (<0.001)	23,734 (<0.001)	57,205 (<0.00)	63,215 (<0.001)	1,576 (<0.001)	86,610 (<0.001)
D.f.	501,761	501,761	501,761	501,761	498,064	499,703
η^2	9.660%	8.643%	18.568%	20.126%	0.629%	25.741%

Descriptive statistics and One-Way ANOVA.
Rows "High," "Medium," and "Low" show mean values, with standard deviation in brackets.
**In trillions of dollars; **In thousands of dollars; ***F-statistic (p-value); Total degrees of freedom.*

TABLE 4 | School statistics by performance level.

	SES	**SCH size**	**St-Tch ratio**	**% immig.**	**% Repeat.**
High	0.097 (0.604)	664.004 (641.448)	12.714 (6.747)	0.095 (0.182)	0.052 (0.089)
Medium	−0.990 (0.935)	431.752 (489.215)	15.463 (12.354)	0.056 (0.149)	0.225 (0.304)
Low	−1.834 (0.746)	333.336 (436.765)	17.465 (13.291)	0.036 (0.124)	0.492 (0.337)
F (p.)*	177,969 (<0.001)	12,078 (<0.001)	4,600 (<0.001)	4,854 (<0.001)	79,058 (<0.001)
D.f.	499,488	442,471	436,922	485,053	492,036
η^2	41.610%	5.177%	2.062%	1.935%	24.584%

Descriptive statistics and One-Way ANOVA.
Rows "High," "Medium," and "Low" show mean values, with standard deviation in brackets.
**F-statistic (p-value); Total degrees of freedom.*

TABLE 5 | Number and percentage of schools in each performance cluster, by type of ownership.

		Private independent	Priv. Gov. Depend.	Public	Total
High	Freq	18,288	7,775	73,748	99,811
	%	18.3%	7.8%	73.9%	100%
Medium	Freq	35,916	16,691	174,132	226,739
	%	15.8%	7.4%	76.8%	100%
Low	Freq	8,436	23,762	115,466	147,664
	%	5.7%	16.1%	78.2%	100%
Total	Freq	62,640	48,228	363,346	474,214
	%	13.2%	10.2%	76.6%	100%

Data are weighted by school weight.

performance schools, a negligible percentage was assigned by the model to schools grouped in the cluster with the opposite performance. While both the training set and cross-validated models showed less than 1% of schools classified as high performance belonged to the low performance group, less than 1.5% of schools classified as low performance belonged to the high performance group. These results reinforce the previous evidence of the goodness of fit of the predictive model.

The scheme of the model obtained in the decision tree is shown in **Figure 1**, presenting the following information:

- Oval nodes indicate group segmentation variables. The initial node (ST – SES) performs the first segmentation of the sample, and the different sub-samples go through different branches of the tree, going down it and performing segmentations until reaching a terminal node (leaves).
- The information included in the arrows shows the segmentation score of the sample from the variable of the previous node. For example, for the initial node (ST – SES), the main sample is divided in two sub-samples, one on the left, which includes the instances with scores between (−inf, −0.19], and one on the right with the instances with scores between (−0.19, +inf). The value under parenthesis indicates the percentage of cases of the (sub)sample included in the previous node that progress through that branch.
- The rectangular terminal nodes (final leaves of the tree) include multiple information: first, a capital letter to indicate the group assigned o classification in the predictive model for that sub-sample (L = low performance; M = medium performance; H = high performance); Second, the percentage of correctly classified instances in the sub-sample, highlighting in black the better accuracy (>0.7), in garnet the acceptable ones (0.6–0.7) and in red those of less fit (<0.6); Finally, the numbers in parentheses show the number of instances included in one specific rule or sub-sample.

The first remarkable question that we can observe in the decision tree is the initial node, that is, the first variable of segmentation. The average SES in schools is the variable with a greater predictive power in the model. Taking into account the terminal nodes of the left side of the tree, it can be noted that schools with lower levels of SES are more related to low performance levels. Specifically, in schools with lower levels of SES most of the consequent nodes include socio-economic variables. In this sense, the model indicates that in schools with disadvantaged socio-economic levels the contextual conditions of the country and the school reach a greater importance than in schools with better socio-economic environments.

In the left side of the tree, which tends to show low performance levels, almost only schools located in countries with very high GDP are associated with high performance levels. In this left side, schools with very high grade repetition rates or located in countries with very low per capita GDP are clearly associated with low performance. However, the model includes some non-cognitive educational outcomes that improve the prediction of school performance (ST – Workmast and ST- Expected SEI) in countries with low GDP and per capita income levels. Thus, the job expectations and the culture of effort of the students can be considered factors that promote better academic performance in these disadvantaged schools and contexts. Finally, in countries with better levels of GDP and per capita income, higher levels of student competence to assess the credibility of the information (ST – Metaspam) are related with better school performance levels. Due to the great differences between countries regarding the characteristics

TABLE 6 | Decision tree fit indices.

		TP	Accuracy	AUROC	Kappa	RRSE
Baseline model (ZeroR). average fit		0.479	0.230	0.500	0	100.00%
Training set	High perform.	0.689	0.802	0.904	–	–
	Medium perform.	0.801	0.666	0.759	–	–
	Low perform.	0.502	0.688	0.877	–	–
	Average fit	0.708	0.715	0.830	0.515	80.97%
Cross-Validation	High perform.	0.688	0.786	0.898	–	–
	Medium perform.	0.789	0.656	0.752	–	–
	Low perform.	0.477	0.679	0.870	–	–
	Average fit	0.697	0.704	0.823	0.496	82.06%

Baseline model, training set, and cross-validation.
TP, true positives rate; AUROC, area under the ROC curve; RRSE, root relative squared error.

TABLE 7 | Confusion matrices in full training set and cross-validated models.

		Classification (decision tree – J48)					
		Training set			Cross-validation		
		High	Medium	Low	High	Medium	Low
Cluster(*k*-means)	High	89,416.58	39,526.6	772.62	89,271.90	39,815.25	628.65
	Medium	21,242.62	149,805.32	15,907.47	23,520.68	147,551.42	15,883.32
	Low	814.04	35,696.62	36,818.12	760.82	37,586.33	34,981.63

TABLE 8 | Distribution of non-cognitive outcomes by School SES and school performance.

		Low SES		High SES		Full sample	
		Mean (*SD*)	$F^*;\eta^2$	Mean (*SD*)	$F^*;\eta^2$;D.f.	Mean (*SD*)	$F^*;\eta^2$;D.f.
Workmast	High P.	0.17 (0.37)	5,438	0.08 (0.37)	1,820	0.10 (0.37)	5,531
	Med. P.	0.16 (0.52)	2.96%	0.08 (0.50)	2.79%	0.14 (0.52)	2.24%
	Low P.	−0.02 (0.51)	420,554	−0.50 (0.95)	62,443	−0.03 (0.52)	483,026
Expect. SEI	High P.	63.7 (11.2)	753	68.6 (7.9)	1,562	67.3 (9.1)	6,737
	Med. P.	63.5 (12.4)	0.42%	65.8 (10.5)	2.39%	64.0 (12.1)	2.71%
	Low P.	62.0 (12.0)	420,595	64.5 (12.8)	62,446	62.0 (12.0)	483,078
Metaspam	High P.	−0.03 (0.42)	33,636	0.17 (0.39)	23,377	0.12 (0.41)	127,357
	Med. P.	−0.50 (0.43)	15.82%	−0.32 (0.41)	26.90%	−0.47 (0.43)	34.43%
	Low P.	−0.70 (0.37)	422,482	−0.55 (0.46)	62,989	−0.70 (0.37)	485,473
Truancy	High P.	0.25 (0.33)	8,561	0.29 (0.27)	5,661	0.28 (0.29)	19,790
	Med. P.	0.49 (0.43)	4.92%	0.50 (0.45)	8.22%	0.49 (0.43)	7.96%
	Low P.	0.61 (0.44)	395,551	0.63 (0.67)	61,833	0.61 (0.44)	457,387

*$p < 0.001$.

of public and private schools, the variable school ownership is hardly interpretable in a single sense.

The right side of the tree is composed of schools with higher average socio-economic levels. The most notable issue in this sub-sample is that the non-cognitive educational outcomes have a greater predictive influence. In this sense, better levels of information credibility assessment in students are clearly related with higher levels of school performance. In fact, the model achieves high accuracy in prediction high performing schools when this factor is combined with not excessively low levels of attendance at early childhood education (ST – Childhood Ed.: more than 1.02 years of attendance to early childhood education on average, a rate reached by more than 99% of schools) and not excessively high levels of self-perceived effort in school tasks (ST – Workmast >1.04, range where more than 99% of schools are located). In schools with lower levels of ST-Metaspam, truancy is the factor with the greatest impact on performance: Schools with non-extreme grade repetition rates in which students, on average, have missed less than 0.49 classes during the last 2 weeks (65.5% of the schools in this sub-sample) are more related to high performance levels.

Non-cognitive Educational Outcomes

At the educational policy level, the variables of greatest interest are the main non-cognitive educational outcomes included. In this sense, **Table 8** shows the distribution of these variables taking into account the main two branches of the tree divided according to school SES. The scores obtained with the full sample indicate low levels of effect sizes in variables Workmast and Expected SEI, moderate effects in Truancy and very high effects in Metaspam. Taking into account the mean scores, schools with low performance have significantly lower levels of students' Workmast, expected SEI and Metaspam and higher levels of truancy. These descriptive results are quite similar when we divide the sample of schools based on SES. However, this relationship is more intense in the upper SES group of schools, mainly in variables Expected SEI, Metaspam and Truancy.

DISCUSSION AND CONCLUSION

The main goal of this research was to study the different factors comprised in the PISA context questionnaires regarding their ability to predict student performance. The study proposes a systematic process for high and low performing schools through the use of clustering techniques, followed by a predictive approach that yielded results with no interference from previous theoretical models, allowing for the emergence of relationships that might be overlooked or less researched in traditional multivariate literature. In this sense, taking into account the main advantages of Data Mining Techniques (Witten et al., 2016; Martínez-Abad, 2019), it was possible to obtain an explanatory

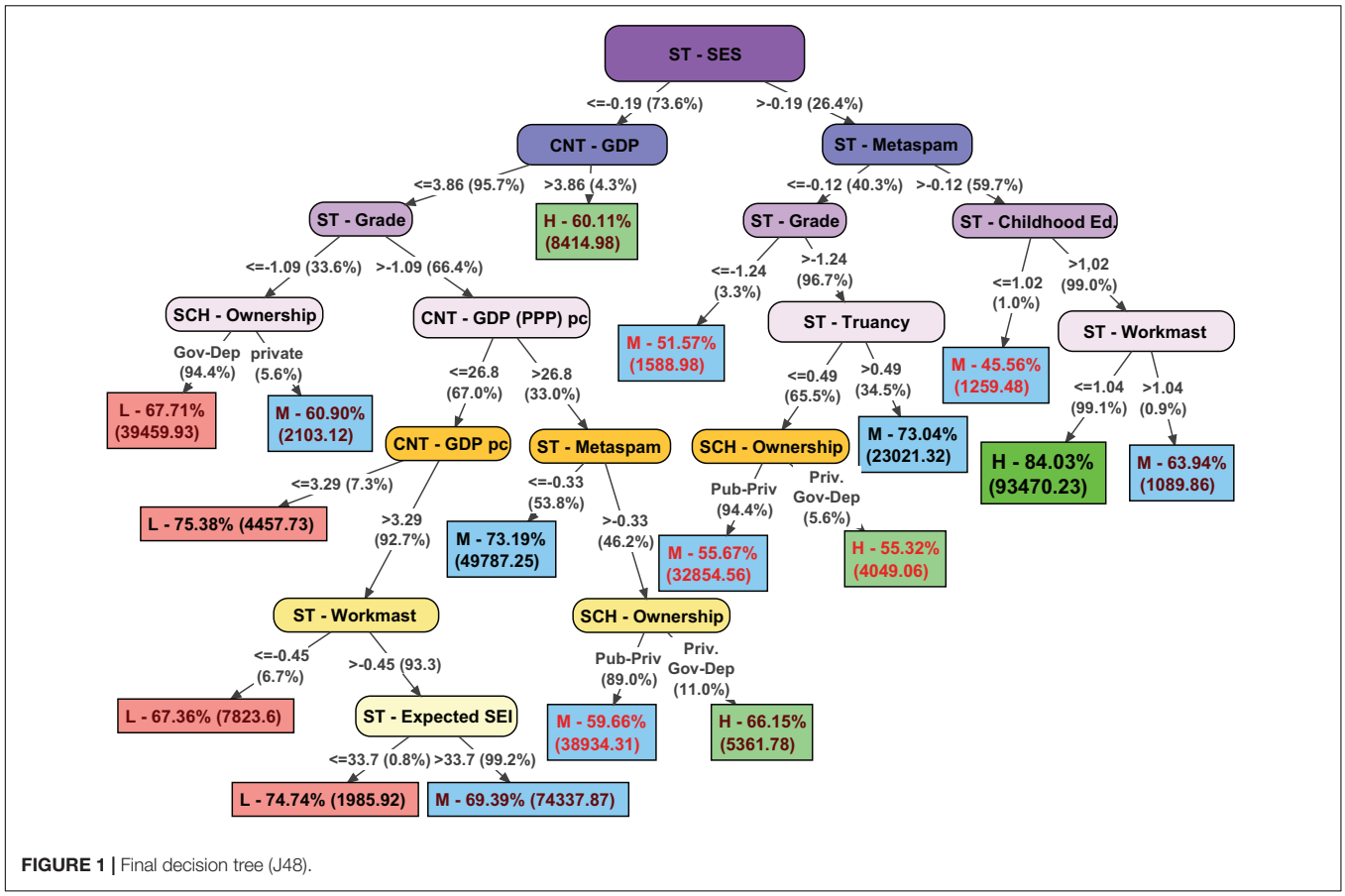

FIGURE 1 | Final decision tree (J48).

model of school performance based on decision trees with acceptable levels of fit. In contrast to the usual practice in EDM research with large scale assessments (Liu and Whitford, 2011; Oskouei and Askari, 2014; Costa et al., 2017; Kılıç Depren et al., 2017; Asensio-Muñoz et al., 2018; Tourón et al., 2018), and according to current studies (She et al., 2019; Martínez-Abad et al., 2020), we limited the size of the final decision tree. This decision made possible a detailed analysis of the main predictive factors linked with school performance and their interactions. To assess the significance and effect size in the main variables of interest, the information obtained from the decision tree was complemented with descriptive and inferential analyses.

Despite of the small size of the decision tree computed, we achieved levels of fit close to previous studies with less parsimonious models (Liu and Whitford, 2011; Oskouei and Askari, 2014; Costa et al., 2017; Kılıç Depren et al., 2017). These results provide a clear answer to the first research question. In this sense, in line with the findings from previous studies based on multivariate analyses (Täht et al., 2014; Karakolidis et al., 2016; Gamazo et al., 2018), the variables with the greatest impact on the model, located in the initial nodes of the tree, were socio-economic factors both at school and country level.

The most relevant variable to predict school performance is SES, which creates the two main branches in the tree: schools with a mean SES above or below −0.19 (these groups of schools will be referred to as "affluent" and "non-affluent," respectively). An overall glance at the characteristics of both branches reveals notable differences in the types of variables that appear in each one. The most relevant variables in the affluent schools branch are largely related to educational and individual characteristics, such as metacognitive strategies, ISCED 0 attendance, or truancy. Although these variables have been highlighted by previous multivariate studies (Cheung et al., 2014, 2016; Rutkowski et al., 2017; Gamazo et al., 2018, respectively) the fact that these variables seem to be relevant to student performance only in affluent school settings has not been explored in any of them.

On the other hand, the non-affluent schools branch contains many economic variables such as nominal GDP and its adjusted variants (per capita and Purchasing Power Parity), or school characteristics such as ownership, which appears twice in this branch; student-level educational indicators, such as metacognitive strategies or motivation to master tasks, seem to be less relevant, as they appear nearer the bottom of the tree. All this seems to indicate that, while affluent schools need to turn their focus on improving student-level educational indicators in order to thrive, non-affluent schools' scores depend in greater measure on economic characteristics that are out of their scope, since they are country-level indicators.

Out of all the country-level economic variables introduced in the model (all of which are located on the non-affluent schools branch of the decision tree), the most relevant one is the country's GDP without any adjustments per population or

purchasing power. This variable generates one of the only two leaves containing high-performance schools in the non-affluent branch, which means that one of the few ways a low-SES school can belong to the high-achievement cluster is by being located on a country with a high nominal GDP; therefore, high levels of GDP (above approximately 4) function as a "protecting factor" for schools with low-SES students. A low level on the other two country variables included (GDP pc and GDP PPP pc) generates a terminal node for low performance schools, leaving little to no space for the consideration of educational variables. Thus, a school from a country with poor economic indicators, both nominal and adjusted, has a meager chance to produce a medium or high level of performance in the PISA test, which attests to the high relevance that economic indicators have as hindering factors for performance (Rodríguez-Santero and Gil-Flores, 2018).

The third research question deals with the school-level variables, of which only school ownership has been included in the model, appearing three times as a node without any derived internal nodes, with medium levels of predictive accuracy. This variable has three possible values, two of which (public and private) have a clear definition in all participating countries. However, the concept of "government-dependent" or "publicly-funded" privately managed schools varies greatly among countries, both in the percentage of public funding allotted and the type of organization managing the school (OECD, 2012). Although this hampers a common interpretation of the meaning and implications of the results of this variable, the positive impact of government-dependent schools on student performance, evidenced by two of the three instances in which it appears in the model, seem to be in line with previous research based on multivariate analyses (Dronkers and Robert, 2008). In any case, these results point to a valuable future line of research that examines the different characteristics and models of government-dependent private schools and their impact on student performance and other outcome variables.

The last research question turns the focus on the educational factors and non-cognitive outcomes included in the decision tree. On the one hand, the relevance achieved by some education indicators in both branches of the tree should be highlighted. In line with previous multivariate studies (Pholphirul, 2016; Gamazo et al., 2018), the model indicates that extremely low average scores in variables Grade and early childhood education attendance prevent schools from belonging to the high-performance cluster. On the other hand, we have previously shown that these variables, mainly non-cognitive outcomes, reach a greater impact in the school performance explanation on the affluent schools branch. We must emphasize that the affluent schools branch includes schools with a high average SES (26.4% of schools sample with higher SES). Thus, in environments with a favorable SES, some educational issues gain relevance. This differential impact depending on the presence of country and school SES has valuable implications for planning educational policies at national levels (Lingard et al., 2013). In this sense, we must study in detail the non-cognitive outcomes included in the model, their contribution and their interactions.

The non-cognitive educational outcome with the greatest contribution to explain school performance has been the students' competence to assess the credibility of the information. Schools, regardless of having high student SES, can only achieve high performance levels in the model with acceptable levels of fit if their students, on average, reach medium or high skills in information assessment. In fact, the effect size of this variable in the general explanation of the school performance is high, an evidence backed up by other works based on multivariate analyses (Cheung et al., 2014, 2016), adding that these effects are even higher in schools with high SES. Although its effects on the decision tree are weak, school truancy also has a major effect size, mainly in schools with high SES. Bearing in mind that previous studies suggest that the prevalence and effects of truancy are mostly related with impoverished settings (Rutkowski et al., 2017), this result merits further research.

The other non-cognitive factor included in the two main branches of the model is the self-perceived effort in school tasks. Considering that this variable is one of the components of achievement motivation in PISA 2018 (OECD, 2019), it is only logical that a high motivation to master tasks should be related to higher levels or school performance, which is the case in this study and others that have examined achievement motivation and its relationship with performance through data mining techniques (Tourón et al., 2018; She et al., 2019). Finally, in accordance with the previous findings (Tourón et al., 2018), the effects of the students' expected occupational status are significant, acting as a promoting factor of school performance, especially in low SES schools from low GDP countries, which is a relevant evidence of the importance of fostering high job and academic expectations among all students.

It is worth noting that, although many of these individual findings find support in studies based both on EDM and multivariate statistics, the use of decision trees allows for an in depth study of the relationships that each of the predictor variables have, not only with the criterion variable, but also with each other (Xu, 2005). This feature generates conclusions such as the importance of country-level economic variables only for low-SES schools, or the higher relevance of truancy or early childhood education in more affluent schools, which are not often found in multivariate studies that focus mainly on the relationship established between each predictor variable and the criterion variable (e.g., Acosta and Hsu, 2014; Karakolidis et al., 2016; Gamazo et al., 2018).

Despite this evidence, which seems robust, it is important to note some important limitations linked both with the use of PISA databases and the methodological approach of this study. On the one hand, the use of cross-sectional data makes it difficult to establish causal relationships (Martínez-Abad et al., 2020). Another notable issue is the variability in the indicators and scales used in different PISA waves (González-Such et al., 2016), which are gradually adapting to socio-educative requirements and trends (López-Rupérez et al., 2019). Thus, the replicability and the development of longitudinal studies are hindered. Another key issue related to the processing of the databases is the categorization of the variable academic performance. Despite the fact that we used clustering techniques to avoid human intervention in the process, and that the decision trees are not based on the covariance matrix to build its models, this

categorization implies a loss of information in the criterion variable. In future studies, it would be advisable to test the fit of models with a greater number of categories of the criterion variable.

On the other hand, we have used an EDM approach trying to find patterns in big data and to transfer that knowledge to support the decision making of educational policies. It is important to note that we have aggregated student variables to the school database to build the decision tree. In this sense, previous research shows better model fits in decision trees computed with aggregated data in the school level compared to the use of student level as the unit of analysis (Martínez-Abad, 2019).

Apart from that, although the study of the gross academic performance in educational research is widespread (Kiray et al., 2015; Aksu and Güzeller, 2016; Karakolidis et al., 2016; Martínez-Abad and Chaparro-Caso-López, 2016), this practice has led to an overrepresentation of the socioeconomic factors in the predictive model. In fact, despite the presence of the socioeconomic factors in the initial nodes of the model has allowed to differentiate some contexts, we also cannot forget that the educational ecologies are complex and multiple (Bronfenbrenner, 1979; Martin and Lazendic, 2018), which makes it difficult to generalize the results obtained.

Finally, there are some future lines of work that derive from the results and reflections of this study. First, in order to collect more solid evidence on the factors linked with school performance in diverse educational environments, future works should delve into the study of differential performance, testing different predictive models depending on the different socio-economic and contextual conditions (Cordero-Ferrera, 2015; Tourón et al., 2018). Second, considering the vast amount of studies that perform secondary analyses of PISA data, it would be convenient to produce a thorough systematic review in order to explore the different methodologies employed, research questions posed and evidences on the impact of diverse variables on student performance.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.oecd.org/pisa/data/2018database/.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin. This study is based on the public databases of the PISA 2018 assessment (OECD). Data collection for OECD-PISA studies is under the responsibility of the governments from the participating countries.

AUTHOR CONTRIBUTIONS

FM-A: problem statement, methods and statistical models, and interpretation and discussion of results. AG: conceptual framework, discussion and conclusions, and style and structure review. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Ministry of Economy and Competitiveness (government of Spain) and FEDER Funds under Grant PGC2018-099174-B-I00.

REFERENCES

Acosta, S. T., and Hsu, H. Y. (2014). Negotiating diversity: an empirical investigation into family, school and student factors influencing New Zealand adolescents' science literacy. *Educ. Stud.* 40, 98–115. doi: 10.1080/03055698.2013.830243

Aksu, G., and Güzeller, C. O. (2016). Classification of PISA 2012 mathematical literacy scores using. decision-tree method: Turkey sampling. *Egitim Bilim* 41, 101–122. doi: 10.15390/EB.2016.4766

Asensio-Muñoz, I., Carpintero-Molina, E., Exposito-Casas, E., and Lopez-Martin, E. (2018). How much gold is in the sand? Data mining with Spain's PISA 2015 results. *Revist. Española Pedagogía* 76, 225–246. doi: 10.22550/REP76-2-2018-02

Avvisati, F. (2020). The measure of socio-economic status in PISA: a review and some suggested improvements. *Large Scale Assess. Educ.* 8:8. doi: 10.1186/s40536-020-00086-x

Baker, R. S. (2010). "Data mining for education," in *International Encyclopedia of Education*, 3 Edn, eds P. Peterson, E. Baker, and B. McGaw (Amsterdam: Elsevier).

Baker, R. S., and Inventado, P. S. (2014). "Educational data mining and learning analytics," in *Learning Analytics: From Research to Practice*, eds J. A. Larusson and B. White (Cham: Springer). doi: 10.1007/978-1-4614-3305-7_4

Barnard-Brak, L., Lan, W. Y., and Yang, Z. (2018). Differences in mathematics achievement according to opportunity to learn: a 4pL item response theory examination. *Stud. Educ. Eval.* 56, 1–7. doi: 10.1016/j.stueduc.2017.11.002

Bronfenbrenner, U. (1979). *The Ecology of Human Development: Experiments by Nature and Design*. Cambridge, MA: Harvard University Press.

Brunello, G., and Rocco, L. (2013). The effect of immigration on the school performance of natives: cross country evidence using PISA test scores. *Econ. Educ. Rev.* 32, 234–246. doi: 10.1016/j.econedurev.2012.10.006

Chaparro Caso López, A. A., and Gamazo, A. (2020). Estudio multinivel sobre las variables explicativas de los resultados de México en PISA 2015. *Arch. Analíticos Polít. Educ.* 28:26. doi: 10.14507/epaa.28.4620

Cheung, K. C., Mak, S. K., Sit, P. S., and Soh, K. C. (2016). A typology of student reading engagement: preparing for response to intervention in the school curriculum. *Stud. Educ. Eval.* 48, 32–42. doi: 10.1016/j.stueduc.2015.12.001

Cheung, K. C., Sit, P. S., Soh, K. C., Ieong, M. K., and Mak, S. K. (2014). Predicting academic resilience with reading engagement and demographic variables: comparing shanghai, Hong Kong, Korea, and Singapore from the PISA perspective. *Asia Pac. Educ. Res.* 23, 895–909. doi: 10.1007/s40299-013-0143-4

Coleman, J. S. (1966). *Equality of Educational Opportunity*. Washington, DC: National Center for Education Statistics.

Cordero-Ferrera, J. M. (2015). Factors promoting educational attainment in unfavorable socioeconomic conditions. *Revist. Educ.* 370, 172–198. doi: 10.4438/1988-592X-RE-2015-370-302

Costa, E. B., Fonseca, B., Santana, M. A., de Araújo, F. F., and Rego, J. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Comput. Hum. Behav.* 73, 247–256. doi: 10.1016/j.chb.2017.01.047

Drent, M., Meelissen, M. R. M., and van der Kleij, F. M. (2013). The contribution of TIMSS to the link between school and classroom factors and student achievement. *J. Curricul. Stud.* 45, 198–224. doi: 10.1080/00220272.2012.727872

Dronkers, J., and Robert, P. (2008). Differences in scholastic achievement of public, private government-dependent, and private independent schools: a cross-national analysis. *Educ. Policy* 22, 541–577.

Farrington, C. A., Roderick, M., Allensworth, E., Nagaoka, J., Keyes, T. S., Johnson, D. W., et al. (2012). *Teaching Adolescents to Become Learners: The Role of Noncognitive Factors in Shaping School Performance–A Critical Literature Review*. Chicago: Consortium on Chicago School Research. doi: 10.1177/0895904807307065

Fertig, M., and Wright, R. E. (2005). School quality, educational attainment and aggregation bias. *Econ. Lett.* 88, 109–114. doi: 10.1016/j.econlet.2004.12.028

Gabriel, F., Signolet, J., and Westwell, M. (2018). A machine learning approach to investigating the effects of mathematics dispositions on mathematical literacy. *Int. J. Res. Method Educ.* 41, 306–327. doi: 10.1080/1743727X.2017.1301916

Gamazo, A., Martínez-Abad, F., Olmos-Migueláñez, S., and Rodríguez-Conde, M. J. (2018). Assessment of factors related to school effectiveness in PISA 2015. A Multilevel Analysis. *Revist. Educ.* 379, 56–84. doi: 10.4438/1988-592X-RE-2017-379-369

Gamazo, A., Olmos-Migueláñez, S., and Martínez-Abad, F. (2016). Multilevel models for the assessment of school effectiveness using PISA scores. *Proc. Fourth Int. Conf. Technol. Ecosyst. Enhancing Multicult.* 16, 1161–1166. doi: 10.1145/3012430.3012663

Gil-Flores, J., and García-Gómez, S. (2017). The importance of teaching practices in relation to regional educational policies in explaining PISA achievement. *Rev. de Educ.* 2017, 52–74. doi: 10.4438/1988-592X-RE-2017-378-361

González-Such, J., Sancho-Álvarez, C., and Sánchez-Delgado, P. (2016). Background questionnaires of PISA: a study of the assessment indicators. *Revist. Electr. Invest. Eval. Educ.* 22:M7. doi: 10.7203/relieve.22.1.8274

Grzymala-Busse, J. W., and Hu, M. (2001). "A comparison of several approaches to missing attribute values in data mining," in *Rough Sets and Current Trends in Computing*, eds W. Ziarko and Y. Yao (Cham: Springer), 378–385. doi: 10.1007/3-540-45554-X_46

International Monetary Fund (2019). *World Economic Outlook, October 2019: Global Manufacturing Downturn, Rising Trade Barriers*. International Monetary Fund. doi: 10.5089/9781513520537.081

Johnson, R. B., Onwuegbuzie, A. J., and Turner, L. A. (2007). Toward a definition of mixed methods research. *J. Mixed Methods Res.* 1, 112–133. doi: 10.1177/1558689806298224

Jornet, M. (2016). Methodological Analysis of the PISA Project as International Assessment. *Rev. Electron. de Investig. y Evaluacion Educ.* 22. doi: 10.7203/relieve22.1.8293

Karakolidis, A., Pitsia, V., and Emvalotis, A. (2016). Examining students' achievement in mathematics: a multilevel analysis of the programme for international student assessment (PISA) 2012 data for Greece. *Int. J. Educ. Res.* 79, 106–115. doi: 10.1016/j.ijer.2016.05.013

Khine, M. S., and Areepattamannil, S. (2016). *Non-Cognitive Skills and Factors in Educational Attainment*. Cham: Sense Publishers. doi: 10.1007/978-94-6300-591-3

Kılıç Depren, S., Aşkın, ÖE., and Öz, E. (2017). Identifying the classification performances of educational data mining methods: a case study for TIMSS. *Educ. Sci. Theory Pract.* 17, 1605–1623. doi: 10.12738/estp.2017.5.0634

Kim, D. H., and Law, H. (2012). Gender gap in maths test scores in South Korea and Hong Kong: role of family background and single-sex schooling. *Int. J. Educ. Dev.* 32, 92–103. doi: 10.1016/j.ijedudev.2011.02.009

Kiray, S. A., Gok, B., and Bozkir, A. S. (2015). Identifying the factors affecting science and mathematics achievement using data mining methods. *J. Educ. Sci. Environ. Health* 1, 28–48. doi: 10.21891/jeseh.41216

Kuger, S., and Klieme, E. (2016). "Dimensions of context assessment," in *Assessing Contexts of Learning: An International Perspective*, eds S. Kuger, E. Klieme, N. Jude, and D. Kaplan (Cham: Springer). doi: 10.1007/978-3-319-45357-6

Kyriakides, L., and Creemers, B. P. (2008). Using a multidimensional approach to measure the impact of classroom-level factors upon student achievement: a study testing the validity of the dynamic model. *Sch. Effect. Sch. Improv.* 19, 183–205. doi: 10.1080/09243450802047873

Lazarević, L. B., and Orlić, A. (2018). PISA 2012 mathematics literacy in serbia: A multilevel analysis of students and schools. *Psihologija* 51, 413–432. doi: 10.2298/PSI170817017L

Lingard, B., Martino, W., and Rezai-Rashti, G. (2013). Testing regimes, accountabilities and education policy: commensurate global and national developments. *J. Educ. Policy* 28, 539–556. doi: 10.1080/02680939.2013.820042

Liu, X., and Ruiz, M. E. (2008). Using data mining to predict K–12 students' performance on large-scale assessment items related to energy. *J. Res. Sci. Teach.* 45, 554–573. doi: 10.1002/tea.20232

Liu, X., and Whitford, M. (2011). Opportunities-to-Learn at home: profiles of students with and without reaching science proficiency. *J. Sci. Edu. Technol.* 20, 375–387. doi: 10.1007/s10956-010-9259-y

López-Rupérez, F., Expósito-Casas, E., and García-García, I. (2019). Equal opportunities and educational inclusion in Spain. *Revist. Electr. Invest. Eval. Educ.* 25, doi: 10.7203/relieve.25.2.14351

Martin, A. J., and Lazendic, G. (2018). Achievement in large-scale national numeracy assessment: an ecological study of motivation and student, home, and school predictors. *J. Educ. Psychol.* 110, 465–482. doi: 10.1037/edu0000231

Martínez-Abad, F. (2019). Identification of factors associated with school effectiveness with data mining techniques: testing a new approach. *Front. Psychol.* 10:2583. doi: 10.3389/fpsyg.2019.02583

Martínez-Abad, F., and Chaparro-Caso-López, A. A. (2016). Data-mining techniques in detecting factors linked to academic achievement. *Sch. Effect. Sch. Improv.* 28, 39–55. doi: 10.1080/09243453.2016.1235591

Martínez-Abad, F., Gamazo, A., and Rodríguez-Conde, M.-J. (2020). Educational data mining: identification of factors associated with school effectiveness in PISA assessment. *Stud. Educ. Eval.* 66:100875. doi: 10.1016/j.stueduc.2020.100875

OECD (2012). *Public and Private Schools: How Management and Funding Relate to their Socio-economic Profile*. Paris: OECD Publishing. doi: 10.1787/9789264175006-en

OECD (2017). *PISA 2015. Technical Report*. Paris: OECD Publishing.

OECD (2019). *PISA 2018 Assessment and Analytical Framework*. Paris: OECD Publishing. doi: 10.1787/b25efab8-en

Oskouei, R. J., and Askari, M. (2014). Predicting academic performance with applying data mining techniques (generalizing the results of two different case studies). *Comput. Eng. Appl. J.* 3, 79–88. doi: 10.18495/comengapp.v3i2.81

Papamitsiou, Z., and Economides, A. (2014). Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence. *Educ. Technol. Soc.* 17, 49–64.

Pholphirul, P. (2016). Pre-primary education and long-term education performance: evidence from programme for international student assessment (PISA) Thailand. *J. Early Child. Res.* 15, 410–432. doi: 10.1177/1476718x15616834

Quinlan, R. (1992). *C4.5: Programs for Machine Learning*. Burlington, MA: Morgan Kaufmann Publishers Inc.

Rodríguez-Santero, J., and Gil-Flores, J. (2018). Contextual variables associated with differences in educational performance between european union countries. *Cult. Educ.* 30, 605–632. doi: 10.1080/11356405.2018.1522024

Romero, C., and Ventura, S. (2010). Educational data mining: a review of the state of the art. *IEEE Trans. Syst. Man Cybern. C* 40, 601–618. doi: 10.1109/TSMCC.2010.2053532

Romero, C., Ventura, S., Pechenizkiy, M., and Baker, R. S. (2010). *Handbook of Educational Data Mining*. Boca Raton, FL: CRC Press. doi: 10.1201/b10274

Rutkowski, D., Rutkowski, L., Wild, J., and Burroughs, N. (2017). Poverty and educational achievement in the US: a less-biased estimate using PISA 2012 data. *J. Child. Poverty* 24, 47–67. doi: 10.1080/10796126.2017.1401898

Sahlberg, P. (2017). *FinnishED Leadership: Four Big, Inexpensive Ideas to Transform Education*, 1st Edn. Corwin.

Scheerens, J., and Bosker, R. (1997). *The Foundations of Educational Efectiveness*. Pergamon: Emerald Group Publishing Limited.

Schleicher, A. (2013). *Big data and PISA*. Available online at: https://oecdedutoday.com/big-data-and-pisa/ (accessed July 26, 2013).

She, H. C., Lin, H. S., and Huang, L. Y. (2019). Reflections on and implications of the programme for international student assessment 2015 (PISA 2015) performance of students in Taiwan: the role of epistemic beliefs about science in scientific literacy. *J. Res. Sci. Teach.* 56, 1309–1340. doi: 10.1002/tea.21553

Shulruf, B., Hattie, J., and Tumen, S. (2008). Individual and school factors affecting students' participation and success in higher education. *High. Educ.* 56, 613–632. doi: 10.1007/s10734-008-9114-8

Smith, P., Cheema, J., Kumi-Yeboah, A., Warrican, S. J., and Alleyne, M. L. (2018). Language-based differences in the literacy performance of bidialectal youth. *Teach. College Rec.* 120, 1–36. doi: 10.1097/tld.0000000000000143

Soh, K. C. (2019). *PISA And PIRLS: The Effects Of Culture And School Environment*. Singapore: World Scientific Publishing Company. doi: 10.1142/11163

United Nations Development Programme (2019). *Human Development Report 2019. Beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century. United Nations*. Available online at: http://www.hdr.undp.org/sites/default/files/hdr2019.pdf

Täht, K., Must, O., Peets, K., and Kattel, R. (2014). Learning motivation from a cCoss-cultural perspective: a moving target? *Educ. Res. Eval.* 20, 255–274. doi: 10.1080/13803611.2014.929009

Tourón, J., López-González, E., Lizasoain-Hernández, L., García-San Pedro, M. J., and Navarro-Asencio, E. (2018). Spanish high and low achievers in science in PISA 2015: impact analysis of some contextual variables. *Revist. Educ.* 380, 156–184. doi: 10.4438/1988-592X-RE-2017-380-376

Tufan, D., and Yildirim, S. (2018). "Historical and theoretical perspectives of data analytics and data mining in education," in *Responsible Analytics and Data Mining in Education: Global Perspectives on Quality, Support, and Decision Making*, eds B. H. Khan, J. R. Corbeil, and M. E. Corbeil (Abingdon: Routledge), 120–140. doi: 10.4324/9780203728703

Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teach. College Rec.* 112, 1008–1037.

Wiseman, A. W. (2013). "Policy responses to PISA in comparative perspective," in *PISA, Power, and Policy: The Emergence of Global Educational Governance*, eds H. D. Meyer and A. Benavot (Providence, RI: Symposium Books), 303–322.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical Machine Learning Tools and Techniques*, 4 Edn. Burlington, MA: Morgan Kaufmann.

World Bank (2020). *Government Expenditure on Education, Total (% of GDP) | Data*. Washington, DC: The World Bank.

Xu, Y. J. (2005). An exploration of using data mining in educational research. *J. Mod. Appl. Stat. Methods* 4, 251–274. doi: 10.22237/jmasm/1114906980

Yao, G., Zhimin, L., and Peng, F. (2015). The effect of family capital on the academic performance of college students–a survey at 20 higher education institutions in Jiangsu Province. *Chin. Educ. Soc.* 48, 81–91. doi: 10.1080/10611932.2015.1014713

Zhang, L., and Jiao, J. (2013). A study on effective hybrid math teaching strategies. *Int. J. Innov. Learn.* 13, 451–466. doi: 10.1504/IJIL.2013.054239

Zhu, Y., and Kaiser, G. (2019). Do east asian migrant students perform equally well in mathematics? *Int. J. Sci. Math. Educ.* 18, 1127–1147. doi: 10.1007/s10763-019-10014-3

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

TABLE A1 | Description and Composition of Variables Used in the Study.

Variable Name	Description
COUNTRY	
CNT_GDP	Gross Domestic Product
CNT_GDP_PPP	Gross Domestic Product (Purchasing Power Parity)
CNT_GDP_pc	Gross Domestic Product, per capita
CNT_GDP_PPP_pc	Gross Domestic Product (Purchasing Power Parity), per capita
CNT_Porc_GDP_Ed	Percentage of GPD spent on education
CNT_Porc_GDP_Sec	Percentage of GPD spent on secondary education
CNT_HDI	Human Development Index
SCHOOL	
SCH_SC001Q01TA	Size of the town/city where the school is located
SCH_SCHLTYPE	School Ownership
SCH_SC048Q02NA	Percentage of students with special needs
SCH_ SC048Q03NA	Percentage of students from socioeconomically disadvantaged homes
SCH_SCHSIZE	School Size
SCH_SC025Q01NA	Percentage of teaching staff attending professional development courses during the last 3 months
SCH_053 (Q01-04, Q09-16)	School offer of extracurricular activities (band, orchestra or choir; school play or musical; school yearbook or newspaper; volunteering; book club; debating club; art club or activities; sporting teams; lectures and/or seminars; collaboration with local libraries; collaboration with local newspapers)
SCH_STRATIO	Student/Teacher ratio
SCH_RATCMP1	Number of available computers per student at modal grade
SCH_RATCMP2	Proportion of available computers that are connected to the Internet
SCH_TOTAT	Total number of all teachers at school
SCH_PROATCE	Index proportion of all teachers fully certified
SCH_PROAT5AB	Index proportion of all teachers ISCED LEVEL 5A Bachelor
SCH_PROAT5AM	Index proportion of all teachers ISCED LEVEL 5A Master
SCH_PROAT6	Index proportion of all teachers ISCED LEVEL 6
SCH_CLSIZE	Class size
SCH_CREACTIV	Creative extra-curricular activities (SC053)
SCH_EDUSHORT	Shortage of educational material (SC017: Q05–Q08)
SCH_STAFFSHORT	Shortage of educational staff (SC017: Q01NA–Q04NA)
SCH_STUBEHA	Student behavior hindering learning (SC061: Q01–Q05, Q11)
SCH_TEACHBEHA	Teacher behavior hindering learning (SC061: Q06–Q10)
STUDENT	
ST_GRADE	Grade compared to modal grade in country
ST_ ST004D01T	Gender
ST_AGE	Age
ST_LANGN	Language spoken at home
ST_IMMIG	Immigration status
ST_DURECEC	Duration in early childhood education and care (ISCED 0)
ST_ST062Q01TA	In the last two full weeks of school, how often: I a whole school day.
ST_ST062Q02TA	In the last two full weeks of school, how often: I some classes.
ST_ ST062Q03TA	In the last two full weeks of school, how often: I arrived late for school.
ST_ EC154 (Q01-Q09)	Additional instruction: enrichment or remedial lessons for test language, mathematics, science or foreign language
ST_ECEC	Duration in early childhood education and care
ST_REPEAT	Grade Repetition
ST_BSMJ	Student's expected occupational status (SEI)
ST_MMINS, LMINS, SMINS	Learning time in mathematics, test language and science
ST_CHANGE	Number of changes in educational biography (Sum)
ST_SES	Index of economic, social and cultural status

(Continued)

TABLE A1 | Continued

Variable Name	Description
ST_UNDREM	Meta-cognition: understanding and remembering (ST164)
ST_METASUM	Meta-cognition: summarizing (ST165)
ST_METASPAM	Meta-cognition: assessing credibility (ST166)
ST_DISCLIMA	Disciplinary climate in test language lessons (ST097)
ST_TEACHSUP	Teacher support in test language lessons (ST100)
ST_DIRINS	Teacher-directed instruction (ST102)
ST_PERFEED	Perceived feedback (ST 104)
ST_EMOSUPS	Parents' emotional support perceived by student (ST123)
ST_STIMREAD	Teacher's stimulation of reading engagement perceived by student (ST152)
ST_ADAPTIVITY	Adaptation of instruction (ST212)
ST_TEACHINT	Perceived teacher's interest (ST213)
ST_JOYREAD	Joy/Like reading (ST160)
ST_SCREADCOMP	Self-concept of reading: perception of competence (ST161: Q01–Q03)
ST_SCREADDIFF	Self-concept of reading: perception of difficulty (ST161: Q06–Q08)
ST_PERCOMP	Perception of competitiveness at school (ST205)
ST_PERCOOP	Perception of cooperation at school (ST206)
ST_ATTLNACT	Attitude toward school: learning activities (ST036)
ST_COMPETE	Competitiveness: dispositional desire to outperform others (ST181)
ST_WORKMAST	Work mastery: dispositional desire to work hard to master tasks (ST182)
ST_GFOFAIL	General fear of failure (ST183)
ST_EUDMO	Eudaemonia: sense of meaning and purpose in life (ST185)
ST_SWBP	Subjective well-being: positive affect (st186)
ST_RESILIENCE	Resilience (ST188)
ST_MASTGOAL	Mastery goal orientation (ST208)
ST_BELONG	Subjective well-being: sense of belonging to school (ST034)
ST_BEINGBULLIED	Student's experience of being bullied (ST038)
ST_ENTUSE	ICT use outside of school (leisure) (IC008)
ST_HOMESCH	Use of ICT outside of school (for school work activities) (IC010)
ST_USESCH	Use of ICT at school in general (IC011)
ST_INTICT	Interest in ICT (IC013)
ST_COMPICT	Perceived ICT competence (IC014)
ST_AUTICT	Perceived autonomy related to ICT use (IC015)
ST_SOIAICT	ICT as a topic in social interaction (IC016)
ST_ICTCLASS	Subject-related ICT use during lessons (IC150)
ST_ICTOUTSIDE	Subject-related ICT use outside of lessons (IC151)
ST_INFOCAR	Information about careers (EC150)
ST_INFOJOB1	Information about the labor market provided by the school (EC151)
ST_INFOJOB2	Information about the labor market provided outside of school (EC151)
TEACHER	
TCH_AGE	Age
TCH_GENDER	Gender
TCH_TC007Q01NA	Year(s) working as a teacher at this school
TCH_TC007Q02NA	Year(s) working as a teacher in total
TCH_TC014Q01HA	Completion of a teacher education or training program
TCH_EMPLTIM	Teacher employment time
TCH_OTT1	Originally trained teacher (strict definition): standard teacher training
TCH_OTT2	Originally trained teacher (wide definition): standard, in-service, or work-based teacher training
TCH_TCSTAFFSHORT	Teacher's view on staff shortage (TC028)
TCH_TCEDUSHORT	Teacher's view on educational material shortage (TC028)
TCH_COLT	Test language teacher collaboration (TC031)
TCH_EXCHT	Exchange and co-ordination for teaching (TC046)
TCH_SATJOB	Teacher's satisfaction with the current job environment (TC198: Q05, Q07, Q09, and Q10)

Continued.

TABLE A1 | Continued.

Variable Name	Description
TCH_SATTEACH	Teacher's satisfaction with teaching profession (TC198: Q01, Q02, Q04, and Q6)
TCH_SEFFCM	Teacher's self-efficacy in classroom management (TC199)
TCH_SEFFREL	Teacher's self-efficacy in maintaining positive relations with students (TC199)
TCH_SEFFINS	Teacher's self-efficacy in instructional settings (TC 199)
TCH_TCOTLCOMP	Opportunity to learn (OTL) aspects of reading comprehension (TC155)
TCH_TCSTIMREAD	Teacher's stimulation of reading engagement (TC156)
TCH_TCSTRATREAD	Teacher's initiation of reading strategies (TC157)
TCH_TCICTUSE	Teacher's use of specific ICT applications (TC169)
TCH_TCDISCLIMA	Disciplinary climate in test language lessons (TC170)
TCH_TCDIRINS	Direct teacher's instruction (TC171)
TCH_FEEDBACK	Feedback provided by the teachers (TC192)
TCH_ADAPTINSTR	Student assessment/use (adaption of instruction) (TC202: Q01–Q04)
TCH_FEEDBINSTR	Feedback provided by the teachers (TC202: Q05–Q09)

The code in brackets indicates which items compose each of the variables used in this study. More information is available in chapter 16 of the PISA 2018 Technical Report (https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018_Technical-Report-Chapter-16-Background-Questionnaires.pdf).

frontiers in Psychology

OPINION
published: 09 December 2020
doi: 10.3389/fpsyg.2020.577410

Analyzing Large-Scale Studies: Benefits and Challenges

Bernhard Ertl[1]*, *Florian G. Hartmann*[2] and *Jörg-Henrik Heine*[3]

[1] Department of Human Sciences, Learning and Teaching With Media, Institute for Education, Universität der Bundeswehr München, Neubiberg, Germany, [2] Department of Human Sciences, Methodology in the Social Sciences, Institute for Education, Universität der Bundeswehr München, Neubiberg, Germany, [3] Center for International Student Assessment, TUM School of Education, Technical University of Munich, Munich, Germany

Keywords: large-scale assessments, ILSA, PISA, PIAAC, NEPS, educational psychology, learning and teaching

INTRODUCTION

The analysis of (inter)national large-scale assessments (LSAs) promises representativity of their results and statistical power and has the ability to reveal even minor effects. LSAs' international grounding verifies previous findings that might previously have been biased by their focus on Western and industrialized countries. This contribution will discuss these promises, contextualizing them via methodical challenges and interpretation caveats that are able to tap the potential of LSAs for educational psychology. Evidence of this contribution is grounded in previous analyses of Program for International Student Assessment (PISA; Schleicher, 2019) and Program for the International Assessment of Adult Competencies (PIAAC; OECD, 2013), two internationally repeated cross-sectional studies. Many aspects we bring up can also apply to several other international large-scale studies, such as TIMSS, PIRLS, and ICILS.[1] We also refer to the national longitudinal study German National Educational Panel Study (NEPS; Blossfeld et al., 2011) to include a perspective on longitudinal studies in this paper. Implications for large-scale studies within the context of learning and teaching round off our paper in its closing section.

OPEN ACCESS

Edited by:
Ching Sing Chai, The Chinese University of Hong Kong, China

Reviewed by:
Trude Nilsen, University of Oslo, Norway
Hui Luan, National Taiwan Normal University, Taiwan
Rebecca J. Collie, University of New South Wales, Australia

***Correspondence:**
Bernhard Ertl
bernhard.ertl@unibw.de

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *29 June 2020*
Accepted: *26 October 2020*
Published: *09 December 2020*

Citation:
Ertl B, Hartmann FG and Heine J-H (2020) Analyzing Large-Scale Studies: Benefits and Challenges. Front. Psychol. 11:577410. doi: 10.3389/fpsyg.2020.577410

PROMISES

Representativity and Impact

LSAs aim to survey representative (sub)samples of defined populations (e.g., OECD, 2013, section Caveats). This representativity can help them be more informative and provide stronger evidence for policymaking than traditional educational or psychological studies that often rely on convenience samples. Wagemaker (2014) discusses changes in educational policies as one of LSAs' impacts. Fischman et al. (2019) looked deeper inside the issue of LSAs' direct impact on educational policy, finding that several countries worldwide have established PISA-based educational goals (p. 12). They further report that LSA results are often used as triggers or levers for educational reforms, while also showing that several stakeholders mentioned that these kinds of studies actually hinder reforms when their focus is too much on simply reaching the stated indicators (see Rutkowski and Rutkowski, 2018).

Longitudinal Perspective

A second LSA benefit is their long-time perspective. They either have been repeated cross-sectionally in several cycles (e.g., the PISA study takes place every 3 years; Schleicher, 2019) or show a longitudinal panel design, such as with NEPS that recently surveyed six starting cohorts in

[1] See, e.g., Lenkeit and Schwippert (2018), Gustafsson (2018), von Maurice et al. (2017), and Rutkowski et al. (2010) for an overview of international large-scale studies.

Germany over the past 10 years (Blossfeld and Roßbach, 2019). While the trend-study approach of PISA allows a measurement of how changes in educational policy or society may impact a defined sample (e.g., 15-year-old students in PISA; Schleicher, 2019), the longitudinal approach of NEPS enables background variables to be revealed, shedding light on how an individual's characteristics affect educational trajectories (Blossfeld and Roßbach, 2019). These procedures can be especially informative if a study like NEPS follows several cohorts that overlap at a certain point in time.

Standardization

Besides representativity and the longitudinal perspective, LSAs provide standardized procedures, instruments, item pools, and test booklets (e.g., OECD, 2013). These standardizations ensure a survey setting and data that allow international comparisons (PIAAC and PISA) as well as comparisons between survey cycles (PIAAC and PISA) or waves (NEPS). An essential prerequisite for supporting these comparisons is the international cooperation for developing competency and performance measures as well as questionnaires (see, e.g., OECD, 2013). Furthermore, the standardized coding of survey data allows a certain level of matching to contextual and/or official data, e.g., labor market data, national examination statistics, or even geodata from microcom in NEPS (Schönberger and Koberg, 2018).[2]

Statistical Power

Finally, the large sample sizes with LSAs provide a statistical power for analyses that allows detection on the individual level of even small effects, even if subsamples of the original population are analyzed. This helps to reveal effects that would have been overlooked in traditional educational or psychological studies. However, statistical power here decreases when analyses go beyond the individual level and focus on class, school, or national realms.

CHALLENGES

Complexity of Analysis

These promises go along with analysis and interpretation challenges. The advantage of representativity in the context of economic sample sizes requires a complex weighting of each case. Consequently, all further analyses must include weights to be able to maintain representativity during analyses. Using stratification variables for sampling that differ across the participating countries to reflect different (educational) structures in their population requires complex variance estimation procedures. This is typically based on replicated estimation or bootstrap procedures (Rust, 1985; Lin et al., 2013) to prove significance statements. In addition, the principle of item sampling (e.g., Lord, 1965) typically used in competence assessment (see Rutkowski et al., 2013) results in design-related missing data points (see below), which are compensated by the plausible value (PV) techniques (e.g., von Davier et al., 2009; von Davier, 2013, and Marsman et al., 2016). Here, analysis procedures have to take not only one but also multiple (e.g., five, ten, or even more) variables (PV) as competence measures into account. However, these kinds of procedures are rare with traditional statistics programs,[3] meaning representative analyses need either add-ons such as the IDB Analyzer[4] or specifically developed packages for R (e.g., survey; BIFIEsurvey, or intsvy; see Heine and Reiss, 2019).

Test Time

Another aspect relates to the extent of the questionnaires. People being surveyed can offer only a limited amount of time. This is typically compensated for in LSAs via two alternative approaches. A pragmatic and easily implemented approach is to apply very short scales for measuring traits and competencies. The NEPS panel, for example, measures the Big Five[5] personality domains with only two items per dimension and vocational interests (the Big Six) with three items per dimension (see Wohlkinger et al., 2011). The issue of expectably low reliabilities and the respective validity is increasingly being discussed in psychological research (Rammstedt and Beierlein, 2014). A more demanding approach in terms of both implementation and later analysis is to use rotated booklet designs (e.g., Frey et al., 2009 and Heine et al., 2016). For computer-based assessments, adaptive test scenarios can usually further reduce the number of items (e.g., Kubinger, 2017). In both test designs, the items are appropriately distributed across different test booklets or even test scenarios. Test takers here often do not answer every item, which inevitably results in missing data points. With a suitable test design, this loss of data is typically completely random, although it still might require the use of data imputation methods which can be complicated to apply.[6]

Missing Data and Imputation

Correspondingly, for the construction of short scales or *within-scale*[7] booklet designs, LSAs often require general design decisions for the assessment of competencies. The NEPS data set for instance surveyed competencies for only about a third of the student cohort (FDZ-LIfBi, 2018), while PIAAC

[2]Matching to contextual data is typically required to preserve the anonymity of individuals and schools. Here, different levels of anonymization, starting from a segment of households up to the municipality level, may be observable (see Schönberger and Koberg, 2018). This kind of matching is usually implemented by the provider of the data set and may require further data access restrictions, e.g., that access is granted only in rooms with specific security precautions. Microcom enrichment may be restricted in some countries and for some studies.

[3]Analyses would be supported by multilevel structural equation modeling, e.g., in MPLUS, if the correct weights are appropriately used and the plausible values are correctly applied. However, the usability of this modeling is dependent on the complexity of the data set and decreases dramatically when nested plausible values are used, for example.

[4]https://www.iea.nl/data-tools/tools

[5]The Big Five is a set of personality variables including the dimensions of openness, conscientiousness, extraversion, agreeableness, and neuroticism (see Goldberg, 1990 and McCrae and John, 1992).

[6]The use of rotated booklet designs and/or adaptive testing usually leads to the imputation of data by the provision of plausible values for estimating test results (see next section). This increases the complexity of analyses (as mentioned in the previous section).

[7]The *within-scale booklet design* is used to describe the phenomenon that all constructs or scales are represented in all booklets, albeit with different and a reduced number of items.

assessed the competency of problem solving in technology-rich environments just for parts of the sample (OECD, 2013) with the booklet designs described above. This means that there is no discrete competency value for an individual; the estimate for competency is based on PVs (e.g., von Davier et al., 2009), which are based on the theory of data imputation (see Rubin, 1987). Modeling longitudinal effects, e.g., by structural equation modeling, furthermore requires the availability of the target variables at specific waves in order to construct valid models.

Invariance of Measurement

A recent OECD conference related to cross-country comparability of questionnaire scales (see Avvisati et al., 2019) identified measurement invariance as a core challenge for LSAs in general and for PISA studies as well (Van de Vijver et al., 2019). Among other methodological topics, participants from different countries discussed typical forms of analysis for verification of measurement invariance. A classical approach for the verification of the measurement invariance uses multigroup confirmatory factor analysis (MGCFA). Based on this, a widely accepted taxonomy includes configurational, metric, scalar, and residual measurement invariance (e.g., Putnick and Bornstein, 2016). The MGCFA approach however also has critical aspects ranging from insufficient subgroup sizes (even for LSA data), reduced test strength, and unknown distribution properties of the test statistics—especially when global model validation tests are used to assess the relative model fit of varyingly nested MGCFA models for levels of measurement invariance. Moreover, MGCFA rests on the assumption of a continuous scale for both the latent variable of interest and the response scales of the manifest indicators. When these strong assumptions of interval scales can be seriously questioned, different models from the IRT domain can be used for ordinal scales or methodology for classification like (multigroup) latent class analysis (MG-LCA—Eid et al., 2003 and Eid, 2019) for nominal scales. Some recent approaches in the LSA framework are founded upon Bayesian IRT models (e.g., Fox, 2010) or IRT residual fit statistics (see, e.g., Buchholz and Hartig, 2017). To establish an invariant scale on the item level, there are in fact some promising approaches to automated item selection to determine a scale, which fulfill predefined target criteria such as invariance across subsamples and cultures (e.g., Schultze and Eid, 2018).

Item Formats and Response Sets

Extreme and middle response endorsement, cheating, socially desirable responding, and flat-lined response behavior are phenomena closely related to the issue of invariant measurement (see Heine, 2020). A critical discussion is currently taking place regarding whether innovative item formats (Kyllonen, 2013) such as *forced choice* measures (e.g., Bürkner et al., 2019) or *anchoring vignettes* to adjust distorted responses (e.g., Stankov et al., 2018) might lead to improved measurement when compared to classical rating scales.

Classification Issues and Different Standards

Standardization and international comparability require the classification of responses, e.g., of vocational aspirations, by standardized classification schemes such as the ISCO-08. However, standardization is always subject to national practice and legislation, and although these schemes are in fact well-defined, they usually do not unambiguously map in alignment with national peculiarities; i.e., they often are only able to partially map national differences. Nursing is widely discussed as a prototypical challenge when it comes to international classification issues (see, e.g., Baumann, 2013 and Palmer and Miles, 2019) because it is distinguished with respect to the educational path (vocational vs. university background) as well as in terms of the scope of medical treatment a nurse is allowed to perform (see, e.g., Currie and Carr-Hill, 2013 and Gunn et al., 2019).

CAVEATS

Significance Does Not Mean Big Effects

Along with these challenges, LSAs also provide some interpretation caveats. The high sample sizes of large-scale studies support big statistical power (on the level of the individual) as a result frequent significance levels of $p < 0.001$ (or lower). Although this is strong when it comes to detecting even marginal differences, it also allows marginal effect sizes (zero effects) to become significant. So merely showing the significance of differences is not sufficient (e.g., Cohen, 1994 and Hunter, 1997) when analyzing large-scale studies; it is necessary to additionally discuss effect sizes (e.g., Snyder and Lawson, 1993).

Horse Race Communication

Countries and states participating in international large-scale studies differ in both their schooling systems and general societal aspects. Just one example of this involves socioeconomic background variables and basic political and social convictions. Different immigration policies in different countries (see, e.g., Entorf and Minoiu, 2005 and Hunger and Krannich, 2015) can lead to a different population composition in so-called "non-native speaker groups," or groups of people with low socioeconomic status might in turn influence (bias) the outcomes of these studies in cross-country comparisons much more than the factor of different school systems. Many international large-scale studies have very complex designs and analyses, and as a result, local or national aspects might be the most illustrative ones to communicate, even if they are not the most relevant ones when considering other educational factors. This often leads to a horse race discussion focusing on the position rather than on the peculiarities of the respective systems. While Rutkowski and Rutkowski (2018) describe how to deal with these peculiarities, the NEPS data use agreement prohibits comparisons between the German federal states[8] to avoid precisely these issues.

[8]https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Datenzugangswege/Vertraege/NEPS_DataUseAgreement_en.pdf

IMPLICATIONS FOR LEARNING AND TEACHING

We have discussed the promises, challenges, and caveats of LSAs. Benefits such as representativity and the long-time perspective go along with challenges such as the complexity of analysis and limited information (e.g., information loss due to classification issues, missing values, constructs not covered, and panel loss) as well as with further caveats for interpretation. This reflects a general issue of these studies, i.e., that their result might have the power to influence educational policies (see Fischman et al., 2019) while at the same time displaying difficulties in being appropriately communicated to teachers, principals, and policymakers due to their complexity. This makes it essential to communicate and transfer LSA evidence into practice in a manner that this is appropriate and understandable for a non-scientific audience, without trivializing its results.

The international perspective of many large-scale studies allows the stereotypes and preconditions that national studies cannot overcome to be reflected upon (see also Else-Quest et al., 2010). These include for example stereotyped gender differences in mathematics and science that in the Western world often favor boys—while PISA results on the other hand have disclosed that several countries show scores favoring girls in mathematics and an almost even distribution in science scores (OECD, 2015, p. 28f.). The study design thereby allows an analysis of the extent to which phenomena develop over time and between different countries, which is an essential aspect for evaluating changes in really any educational system. Incidentally, education always targets the development of individuals. So longitudinal follow-up surveys and analyses of cohorts may increase the benefits of these studies as they relate to learning and teaching.

To sum up, (inter)national large-scale studies can provide several benefits for research on learning and teaching in how they achieve a solid data set for investigating relevant effects. However, the formal comparability of study scores does not exactly reflect actual differences between states or educational systems without considering background variables and national social and educational specifics. Although these studies may mitigate the methodical shortcomings of traditional studies, especially the focus on Western white populations, they at the same time may reveal methodical challenges.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

Conceptual analyses resulting in this article were partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - project ER470/2-1. The publication of this article was funded by the Open Access Fund of the Bundeswehr Universität München.

REFERENCES

Avvisati, F., Le Donné, N., and Paccagnella, M. (2019). A meeting report: cross-cultural comparability of questionnaire measures in large-scale international surveys. *Meas. Instrum. Soc. Sci.* 1:8. doi: 10.1186/s42409-019-0010-z

Baumann, A. (2013). What's in a name? The importance of definition and comparable data. *Int. Nurs. Rev.* 60, 75–77. doi: 10.1111/j.1466-7657.2012.01046.x

Blossfeld, H. P., and Roßbach, H. G. (Eds.). (2019). *Education as a Lifelong Process: The German National Educational Panel Study (NEPS),* 2nd Edn. Wiesbaden: SpringerVS. doi: 10.1007/978-3-658-23162-0

Blossfeld, H. P., Roßbach, H. G., and von Maurice, J. (2011). Education as a lifelong process: The German National Educational Panel Study (NEPS). *Zeitschrift Erziehungswissenschaft Sonderheft.* 14, 19–34. doi: 10.1007/s11618-011-0179-2

Buchholz, J., and Hartig, J. (2017). Comparing attitudes across groups: An IRT-based item-fit statistic for the analysis of measurement invariance. *Appl. Psychol. Meas.* 5, 1–10. doi: 10.1177/0146621617748323

Bürkner, P. C., Schulte, N., and Holling, H. (2019). On the statistical and practical limitations of thurstonian IRT models. *Educ. Psychol. Meas.* 79, 827–854. doi: 10.1177/0013164419832063

Cohen, J. (1994). The earth is round ($p<.05$). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Currie, E. J., and Carr-Hill, R. A. (2013). What is a nurse? Is there an international consensus? *Int. Nurs. Rev.* 60, 67–74. doi: 10.1111/j.1466-7657.2012.00997.x

Eid, M. (2019). "Multigroup and multilevel latent class analysis," in *Invariance Analyses in Large-Scale Studies,* ed F. J. van de Vijver (Paris: OECD Publishing), 70–90.

Eid, M., Langeheine, R., and Diener, E. (2003). Comparing typological structures across cultures by multigroup latent class analysis. *J. Cross Cult. Psychol.* 34, 195–210. doi: 10.1177/0022022102250427

Else-Quest, N. M., Hyde, J. S., and Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: a meta-analysis. *Psychol. Bull.* 136, 103–127. doi: 10.1037/a0018053

Entorf, H., and Minoiu, N. (2005). What a difference immigration policy makes: a comparison of PISA scores in Europe and traditional countries of immigration. *German Econ. Rev.* 6, 355–376. doi: 10.1111/j.1468-0475.2005.00137.x

FDZ-LIfBi (2018). *Codebook. NEPS Starting Cohort 5—First-Year Students. From Higher Education to the Labor Market. Scientific Use File Version 11.0.0.* Retrieved from https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC5/11-0-0/SC5_11-0-0_Codebook_en.pdf

Fischman, G. E., Topper, A. M., Silova, I., Goebel, J., and Holloway, J. L. (2019). Examining the influence of international large-scale assessments on national education policies. *J. Educ. Policy* 34, 470–499. doi: 10.1080/02680939.2018.1460493

Fox, J. (2010). *Bayesian Item Response Modeling.* New York, NY: Springer New York. doi: 10.1007/978-1-4419-0742-4

Frey, A., Hartig, J., and Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: theory and practice. *Educ. Meas.* 28, 39–53. doi: 10.1111/j.1745-3992.2009.00154.x

Goldberg, L. R. (1990). An alternative description of personality: the big-five factor structure. *J. Pers. Soc. Psychol.* 59, 1216–1229. doi: 10.1037/0022-3514.59.6.1216

Gunn, V., Muntaner, C., Ng, E., Villeneuve, M., Gea-Sanchez, M., and Chung, H. (2019). Gender equality policies, nursing professionalization, and the nursing workforce: a cross-sectional, time-series analysis of 22 countries, 2000–2015. *Int. J. Nurs. Stud.* 99:103388. doi: 10.1016/j.ijnurstu.2019.103388

Gustafsson, J. E. (2018). International large-scale assessments: current status and ways forward. *Scand. J. Educ. Res.* 62, 328–332. doi: 10.1080/00313831.2018.1443573

Heine, J. H. (2020). *Untersuchungen zum Antwortverhalten und zu Modellen der Skalierung bei der Messung psychologischer Konstrukte.* München; Neubiberg: Universität der Bundeswehr.

Heine, J. H., Mang, J., Borchert, L., Gomolka, J., Kröhne, U., Goldhammer, F., and Sälzer, C. (2016). "Kompetenzmessung in PISA 2015," in *PISA 2015: Eine Studie zwischen Kontinuität und Innovation,* eds K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, and O. Köller (Münster: Waxmann), 383–430.

Heine, J. H., and Reiss, K. (2019). "Pisa 2018 – die Methodologie," in *PISA 2018 Grundbildung im internationalen Vergleich*, eds K. Reiss, M. Weis, E. Klieme, and O. Köller (Münster: Waxmann), 241–258.

Hunger, U., and Krannich, S. (2015). *Einwanderungsregelungen im Vergleich: was Deutschland von anderen Ländern lernen kann*. Bonn: Friedrich-Ebert-Stiftung.

Hunter, J. E. (1997). Needed: a ban on the significance test. *Psychol. Sci.* 8, 3–7. doi: 10.1111/j.1467-9280.1997.tb00534.x

Kubinger, K. D. (2017). "Adaptive testing," in *Principles and Methods of Test Construction: Standards and Recent Advances*. Vol. 3, *Psychological Assessment - Science and Practice,* eds K. Schweizer and C. DiStefano (Göttingen: Hogrefe), 104–119.

Kyllonen, P. and Bertling, J. (2013). "Innovative questionnaire assessment methods to increase cross-country comparability," in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis,* eds L. Rutkowski, L. M. von Davier and D. Rutkowski (Boca Raton: Chapman and Hall/CRC), 277–285.

Lenkeit, J., and Schwippert, K. (2018). Doing research with international assessment studies: methodological and conceptual challenges and ways forward. *Assess. Educ.* 25, 1–4. doi: 10.1080/0969594X.2017.1352137

Lin, C., Devon, W., Lu, W., Rust, K., and Sitter, R. R. (2013). Replication variance estimation in unequal probability sampling without replacement: One-stage and two-stage. *Can. J. Stat. Revue Canad. Stat.* 41, 696–716. doi: 10.1002/cjs.11200

Lord, F. M. (1965). Item sampling in test theory and in research design. *ETS Res. Bull. Series* 1965, i—39. doi: 10.1002/j.2333-8504.1965.tb00968.x

Marsman, M., Maris, G., Bechger, T., and Glas, C. (2016). What can we learn from plausible values? *Psychometrika* 81, 274–289. doi: 10.1007/s11336-016-9497-x

McCrae, R. R., and John, O. P. (1992). An introduction to the Five-Factor model and its applications. *J. Pers.* 60, 175–215. doi: 10.1111/j.1467-6494.1992.tb00970.x

OECD (2013). *Technical Report of the Survey of Adult Skills (PIAAC)*. Paris: OECD Publishing. doi: 10.1787/9789264204027-en

OECD (2015). *The ABC of Gender Equality in Education: Aptitude, Behavior, Confidence*. Paris: OECD Publishing. doi: 10.1787/9789264229945-en

Palmer, S. P., and Miles, L. W. (2019). Students' observations of the nursing role in seven nations. *Nurs. Educ. Perspect.* 40, 283–290. doi: 10.1097/01.NEP.0000000000000560

Putnick, D. L., and Bornstein, M. H. (2016). Measurement invariance conventions and reporting: the state of the art and future directions for psychological research. *Dev. Rev.* 41, 71–90. doi: 10.1016/j.dr.2016.06.004

Rammstedt, B., and Beierlein, C. (2014). Can't we make it any shorter? *J. Ind. Diff.* 35, 212–220. doi: 10.1027/1614-0001/a000141

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley. doi: 10.1002/9780470316696

Rust, K. F. (1985). Variance estimation for complex estimators in sample surveys. *J. Off. Stat.* 1, 381–397.

Rutkowski, L., Gonzalez, E., Joncas, M., and von Davier. M. (2010). International large-scale assessment data: issues in secondary analysis and reporting. *Educ. Res.* 39, 142–151. doi: 10.3102/0013189X10363170

Rutkowski, L., Gonzalez, E., Von Davier, M., and Zhou, Y. (2013). "Assessment design for international large-scale assessments," in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis,* eds L. Rutkowski, M. V. Davier, and D. Rutkowski (Boca Raton, FL: CRC Press), 75–95. doi: 10.1201/b16061

Rutkowski, L., and Rutkowski, D. (2018). Improving the comparability and local usefulness of international assessments: a look back and a way forward. *Scand. J. Educ. Res.* 62, 354–367. doi: 10.1080/00313831.2016.1261044

Schleicher, A. (2019). *PISA 2018 Insights and Interpretations*. Paris: OECD Publishing.

Schönberger, K., and Koberg, T. (2018). *Regional Data: Microcom*. Bamberg: Research Data Center LIfBi.

Schultze, M., and Eid, M. (2018). Identifying measurement invariant item sets in cross-cultural settings using an automated item selection procedure. *Methodology* 14, 177–188. doi: 10.1027/1614-2241/a000155

Snyder, P., and Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *J. Exp. Educ.* 61, 334–349. doi: 10.1080/00220973.1993.10806594

Stankov, L., Lee, J., and von Davier, M. (2018). A note on construct validity of the anchoring method in PISA 2012. *J. Psychoeduc. Assess.* 36, 709–724. doi: 10.1177/0734282917702270

Van de Vijver, F. J. R., Avvisati, F. Davidov, E., Eid, M., Fox, J. P. Le Donné, N. et al., (2019). "Invariance analyses in large-scale studies," in *OECD Education Working Papers* (Paris: OECD Publishing).

von Davier, M. (2013). "Imputing proficiency data under planned missingness in population models," in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis,* eds L. Rutkowski, M. V. Davier, and D. Rutkowski (Boca Raton, FL: CRC Press), 175–202.

von Davier, M., Gonzalez, E., and Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Monogr. Series* 2, 9–36.

von Maurice, J., Zinn, S., and Wolter, I. (2017). Large-scale assessments: potentials and challenges in longitudinal designs. *Psychol. Test Assess, Model.* 59, 35–54.

Wagemaker, H. (2014). "International Large-scale assessments: from research to policy," in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis,* eds L. Rutkowski, M. V. Davier, and D. Rutkowski (Boca Raton; London; New York, NY: CRC Press), 11–36.

Wohlkinger, F., Ditton, H., von Maurice, J., Haugwitz, M., and Blossfeld, H. P. (2011). 10 Motivational concepts and personality aspects across the life course. *Zeitschrift Erziehungswissenschaft* 14:155. doi: 10.1007/s11618-011-0184-5

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Psychology

BRIEF RESEARCH REPORT
published: 15 January 2021
doi: 10.3389/fpsyg.2020.564327

Stricter Teacher, More Motivated Students? Comparing the Associations Between Teacher Behaviors and Motivational Beliefs of Western and East Asian Learners

Yushan Jiang[1], Chi-Kin John Lee[1], Zhi Hong Wan[1]* and Junjun Chen[2]*

[1] Department of Curriculum and Instruction, The Education University of Hong Kong, Tai Po, Hong Kong, [2] Department of Education Policy and Leadership, The Education University of Hong Kong, Tai Po, Hong Kong

Teacher behaviors are one of the most significant factors influencing student learning. Students from different cultures may have different interpretations of their teachers' behaviors. This study compared the associations between teacher strictness, teacher feedback, and students' motivational beliefs using data from six Western countries (the United States, the United Kingdom, Finland, Norway, Australia, and New Zealand) and six East Asian regions (Japan, Korea, mainland China, Hong Kong, Macau, and Taiwan) in the Program for International Student Assessment (PISA) 2015. A total of 89,869 15-year-old students were included in data analysis. The findings indicate that (i) teacher strictness was negatively associated with Western students' motivation, but positively related to that of East Asian students; (ii) teacher feedback had significant positive associations with the motivational beliefs of both Western and East Asian students; and (iii) there was a positive relationship between teacher strictness and teacher feedback in East Asian context. These results highlight the need to consider cultural factors when interpreting students' reactions to teacher behaviors.

Keywords: teacher feedback, teacher strictness, motivational beliefs, Western and East Asian learners, interpersonal behavior and communication, teacher behavior and classroom practice

OPEN ACCESS

Edited by:
Ronnel B. King,
University of Macau, China

Reviewed by:
Pei-Yi Lin,
The Chinese University of Hong Kong, China
Ma. Jenina N. Nalipay,
The Education University of Hong Kong, Hong Kong

***Correspondence:**
Chi-Kin John Lee
jcklee@eduhk.hk
Zhi Hong Wan
wanzh@eduhk.hk

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *21 May 2020*
Accepted: *03 December 2020*
Published: *15 January 2021*

Citation:
Jiang Y, Lee C-KJ, Wan ZH and Chen J (2021) Stricter Teacher, More Motivated Students? Comparing the Associations Between Teacher Behaviors and Motivational Beliefs of Western and East Asian Learners. Front. Psychol. 11:564327. doi: 10.3389/fpsyg.2020.564327

INTRODUCTION

In the past few decades, teacher behaviors have attracted considerable attention in the fields of learning environment and educational effectiveness (e.g., den Brok et al., 2004; Kyriakides et al., 2020). A growing body of studies have revealed the significant influence of teacher behaviors on students' engagement, motivation, and achievement (Brekelmans et al., 2000, 2002; Roorda et al., 2011; Wubbels et al., 2016).

Teacher behaviors can be broadly classified as interpersonal and teaching behaviors. Interpersonal behaviors are usually conceptualized and investigated using the Model of Interpersonal Teacher Behavior (MITB; Wubbels and Brekelmans, 2005), which encompasses eight sectors of behaviors, namely, leadership, helpful/friendly, understanding, student freedom, uncertainty, dissatisfaction, and strictness. Previous empirical studies showed that teachers' favorable interpersonal behaviors are strongly correlated with student motivation (e.g.,

den Brok et al., 2004; Lapointe et al., 2005). Maulana et al. (2011) found high student motivation was moderately related to teachers' proximity and influence behaviors with Indonesian samples. Likewise, van Uden et al. (2014) reported positive correlations between student engagement and teachers' influence and proximity in the Dutch context. In addition, teacher's interpersonal behaviors have been identified to influence students' learning attitudes (Quek et al., 2007).

As for teaching behaviors, previous studies have investigated the impacts of feedback, clarity, modeling, questioning, reinforcement, and communication of teacher expectations on student learning (e.g., Creemers, 1994; Soh, 2017; Gentrup et al., 2020; Kyriakides et al., 2020). Among these constructs, teacher feedback has been considered as one of the most important practices for improving student learning (Gentrup et al., 2020). Previous research has revealed a direct positive impact of teacher feedback on students' self-efficacy (Rakoczy et al., 2019) and motivation (e.g., Hamidun et al., 2012; Pat-El et al., 2012). In particular, scaffolding behaviors in the form of giving extra information about how to improve performance on tasks have shown to have a positive influence on student motivation (Dresel and Haugwitz, 2008).

Although previous studies have revealed a close relationship between teacher behaviors and student learning, most of these studies were conducted in the West (Pennings et al., 2014; Pennings and Hollenstein, 2019; Sun et al., 2019). However, as revealed in some comparative studies, teacher behaviors might be interpreted differently by East Asian and Western learners. For example, Chinese students in Lewis et al.'s (2008) study believed their teachers' disciplinary actions are more justified when compared to Australian students. Zhou et al. (2012) have found that Chinese students perceived less controls when their teachers provide corrective feedbacks, while their American counterparts perceived more controls. Given the differences between East Asian and Western learners in their interpretations of same teacher behaviors, their motivational and behavioral reactions to same teacher behaviors might also differ (Tsai et al., 2016). In other words, there may exist cultural differences in the relationship between teacher behaviors and student learning. To date, there is a scarcity of research to explore such cultural differences.

This study aims to explore the associations between two kinds of teacher behaviors (i.e., teacher strictness and teacher feedback) and four motivational beliefs (i.e., intrinsic motivation, instrumental motivation, achievement motivation, and self-efficacy) of Western and East Asian learners. *Intrinsic motivation* refers to the enjoyment and interests that students may experience from the learning process, *instrumental motivation* is the perceived usefulness of learning in students' future studies and career, and *achievement motivation* encompasses students' needs for success and excellence (Cheng and Wan, 2016). In addition, *self-efficacy* refers to students' beliefs in the extent to which s/he will perform well in a task (Wang et al., 2014). Together, these motivational beliefs are strong predictors of students learning and achievement.

Current study utilized PISA 2015 data to explore the cultural difference between Western and East Asian learners for two reasons: (i) PISA utilizes standardized tools across the Organization for Economic Cooperation and Development (OECD) countries, providing an opportunity to make fair comparisons; and (ii) it adopts a very strict sampling procedure that enables accurate statistical analyses. Therefore, it might be meaningful to explore whether there are significant correlations between teacher strictness and teacher feedback in this study and whether their relationship is also culturally embedded. In sum, the following questions were investigated in this study:

1. How do teacher strictness and teacher feedback affect the motivational beliefs of Western and East Asian learners?
2. How do teacher strictness and teacher feedback correlate in Western and East Asian contexts?

METHODS

Data

The empirical analysis in this study relies on the PISA 2015 data downloaded in January 2020. Approximately 540,000 15-year-old students from 72 countries and economies were asked to fill out questionnaires and assessments to evaluate their attitudes, motivation, and academic performance (OECD, 2016). The database is publicly available at the OECD website[1].

To ensure cross-cultural investigation, 12 countries and economies were selected for this study. Six were Western countries (i.e., the United States, the United Kingdom, Finland, Norway, Australia, and New Zealand) and six were East Asian countries or economies (i.e., Japan, Korea, mainland China, Hong Kong, Macau, and Taiwan). The study data came from 89,869 students from the selected countries and regions. Of these students, 44,149 (49.1%) were girls and 45,720 (50.9%) were boys; 50,257 (55.9%) were from the West and 39,612 (44.1%) were from the East Asia.

Variables

The PISA 2015 included two scales related to teachers' behavior (teacher feedback and teacher strictness) and four scales related to students' motivational beliefs (intrinsic motivation, instrumental motivation, achievement motivation, and self-efficacy). The questionnaires can be accessed through the PISA website[2], and the items of measurements for current study are illustrated in **Table A1**.

Teacher Strictness

This index included four items that capture students' perceptions of the ways their teachers treat them. Sample items are "Teachers disciplined me more harshly than other students." and "Teachers graded me harder than they graded other students." The four-point Likert scale was adopted with 1 indicating "never or almost never," 2 indicating "a few times a year," 3 indicating "a few times a month," and 4 indicating "once a week or more."

[1] http://www.oecd.org/pisa/data/

[2] http://www.oecd.org/pisa/data/2015database/

Teacher Feedback

This index included five items that capture students' perception of the frequency of receiving teacher's formative feedback. Sample items include "The teacher tells me how I am performing in this course."; "The teacher tells me in which areas I can still improve." The four-point Likert scale was used with 1 indicating "never or almost never," 2 indicating "some lessons," 3 indicating "many lessons," and 4 indicating "every lesson or almost every lesson."

Motivational Beliefs

As the focus of PISA 2015 was science subjects, the structure and design of the questionnaire was specifically targeted at students' science learning. Students' motivational beliefs were measured through: intrinsic motivation (5 items), instrumental motivation (4 items), achievement motivation (5 items), and science self-efficacy (8 items). Measures of intrinsic and instrumental motivation and self-efficacy assessed students' motivational beliefs within a science learning context, whereas achievement motivation assessed students' overall motivation. Sample items are "I am interested in learning about science," "Many things I learn in my science subject(s) will help me to get a job," and "I want to be the best, whatever I do," respectively, for intrinsic, instrumental, and achievement motivation. In order to ensure consistency in construct scaling, the responses for instrumental motivation have been reverse coded, so that 1 indicates strongly disagree and 4 indicates strongly agree for all three motivation measurements. For science self-efficacy, the students were asked to rate their confidence in completing particular science-related tasks, such as "Explain why earthquakes occur more frequently in some areas than in others." Responses were reverse coded on the four-point scale with 1 being "I could not do this," 2 being "I would struggle to do this on my own," 3 being "I could do this with a bit of effort," and 4 being "I could do this easily."

Data Analysis

Cronbach's alpha was calculated for the six scales in the survey as an indicator of their reliability (**Table 1**). The overall alpha coefficient for teacher strictness was 0.718, and the overall alpha coefficient for teacher feedback was 0.932. For constructs under motivational beliefs, the alpha coefficients ranged from 0.848 to 0.951. As suggested by Fink (2015), the criterion for Cronbach's alpha coefficient is 0.70, therefore all these scales can be considered to have a good reliability.

TABLE 1 | Alpha coefficients for the six constructs.

Constructs	Alpha coefficients
Teacher behavior	
Teacher strictness	0.718
Teacher feedback	0.932
Motivational beliefs	
Intrinsic motivation	0.951
Instrumental motivation	0.935
Achievement motivation	0.848
Science self-efficacy	0.906

Item-to-scale correlation was calculated to estimate the scale validity (**Table 2**). The average item-to-scale correlation coefficients for teacher strictness and teacher feedback were 0.743 and 0.886, respectively. The average item-to-scale correlation coefficients for motivational beliefs were 0.914, 0.914, 0.789, and 0.774, respectively, for intrinsic motivation, instrumental motivation, achievement motivation, and self-efficacy. A score above.30 indicates internal consistency (Gable, 1986), thus, the scales were all valid.

Next, the Pearson correlation coefficients of all constructs were estimated to check if they were significantly correlated before performing a structural equation modeling (SEM) analysis. The relationships among teacher strictness, teacher feedback, and students' motivation beliefs were then further explored using SEM to have an estimation of whether these relationships varied across Western and East Asian learners. Finally, a multi-group analysis was conducted to determine whether such variations in their relationships were statistically significant.

RESULTS

Correlation Analyses

To explore the relationship among teacher feedback, teacher strictness, and students' motivational beliefs, Pearson correlations were performed. As shown in **Table 3**, the correlation between teacher strictness and teacher feedback was negative and significant for students from Western cultures ($r = -0.051$, $p < 0.01$). The correlations between teacher strictness and motivational beliefs were significantly negative. Teacher strictness was most strongly correlated with intrinsic motivation ($r = -0.154$, $p < 0.01$), followed by instrumental motivation ($r = -0.094$, $p < 0.01$), self-efficacy ($r = -0.074$, $p < 0.01$), and achievement motivation ($r = -0.026$, $p < 0.01$). Teacher feedback was most strongly correlated with intrinsic motivation ($r = 0.235$, $p < 0.01$), followed by instrumental motivation ($r = 0.167$, $p < 0.01$), self-efficacy ($r = 0.154$, $p < 0.01$), and achievement motivation ($r = 0.149$, $p < 0.01$). Using Bonferroni adjusted significance level of.003, the motivational beliefs constructs were significantly and positively correlated with one another (r ranged from 0.211 to 0.437).

As shown in **Table 4**, for East Asian learners the correlation between teacher strictness and teacher feedback was significantly positive ($r = 0.069$, $p < 0.01$). There were significantly positive correlations between teacher strictness and motivational beliefs. Specifically, achievement motivation ($r = 0.058$, $p < 0.01$) had the strongest correlation with perceived teacher strictness, followed by intrinsic motivation ($r = 0.046$, $p < 0.01$), instrumental motivation ($r = 0.021$, $p < 0.01$), and self-efficacy ($r = 0.010$, $p = 0.031$). The strongest correlation between teacher feedback and students' motivational beliefs was for intrinsic motivation ($r = 0.265$, $p < 0.01$), followed by instrumental motivation ($r = 0.247$, $p < 0.01$), self-efficacy ($r = 0.175$, $p < 0.01$), and achievement motivation ($r = 0.136$, $p < 0.01$). In addition, the motivational beliefs were significantly and positively correlated with one another at the Bonferroni adjusted 0.003 significance

TABLE 2 | Item-to-scale correlations of the six constructs.

Teacher strictness		Teacher feedback		Intrinsic motivation		Instrumental motivation		Achievement motivation		Self-efficacy	
Item	**Corr.**	**Item**	**Corr.**	**Item**	**Corr.**	**Item**	**Corr.**	**Item**	**Corr.**	**Item**	**Corr.**
TST1	0.691	TFB1	0.841	INT1	0.899	INS1	0.907	ACH1	0.802	SEF1	0.761
TST2	0.769	TFB2	0.888	INT2	0.897	INS2	0.928	ACH2	0.752	SEF2	0.739
TST3	0.760	TFB3	0.912	INT3	0.921	INS3	0.925	ACH3	0.827	SEF3	0.785
TST4	0.753	TFB4	0.906	INT4	0.924	INS4	0.897	ACH4	0.736	SEF4	0.789
		TFB5	0.884	INT5	0.927			ACH5	0.830	SEF5	0.799
										SEF6	0.776
										SEF7	0.776
										SEF8	0.766
Mean	0.743	Mean	0.886	Mean	0.914	Mean	0.914	Mean	0.789	Mean	0.774

TABLE 3 | Correlations among the 6 constructs for Western learners.

	1	2	3	4	5	6
1. Teacher strictness	1					
2. Teacher feedback	−0.051*	1				
3. Intrinsic motivation	−0.154*	0.235*	1			
4. Instrumental motivation	−0.094*	0.167*	0.430*	1		
5. Achievement motivation	−0.026*	0.149*	0.231*	0.211*	1	
6. Self-efficacy	−0.074*	0.154*	0.437*	0.324*	0.216*	1

**Correlation is significant at the adjusted 0.03 level.*

TABLE 4 | Correlations among the 6 constructs for East Asian learners.

	1	2	3	4	5	6
1. Teacher strictness	1					
2. Teacher feedback	0.069*	1				
3. Intrinsic motivation	0.046*	0.265*	1			
4. Instrumental motivation	0.021*	0.247*	0.460*	1		
5. Achievement motivation	0.058*	0.136*	0.250*	0.220*	1	
6. Self-efficacy	0.010	0.175*	0.386*	0.289*	0.233*	1

**Correlation is significant at the adjusted 0.03 level.*

level (r ranged from 0.220 to 0.460), except for teacher strictness and self-efficacy.

SEM Analyses

SEM analyses were performed to further explore the relationships among the variables. We separated Western and East Asian students and used the same model to illustrate the relationships between perceived teacher strictness, teacher feedback, and motivational beliefs.

Multiple fit indices are used for SEM, one of the most common is the ratio of chi-square (χ^2) statistic (Lehman et al., 2013). However, the value of chi-square statistic is sensitive to sample size (Anderson and Gerbing, 1988) as a large sample size may generate a significant chi-square result with minor discrepancies. Given a rather large sample size of the current study (n = 89,869), it will be more robust to adopt multiple goodness-of-fit indices, including CFI, PNFI, and RMSEA.

As shown in **Figure 1**, for students from Western cultures, teacher strictness had negative and significant effects on all constructs of motivational beliefs, including intrinsic motivation ($\beta = -0.205$, $p < 0.01$), instrumental motivation ($\beta = -0.141$, $p < 0.01$), achievement motivation ($\beta = -0.062$, $p < 0.01$), and self-efficacy ($\beta = -0.117$, $p < 0.01$). Teacher feedback had positive and significant effects on intrinsic motivation ($\beta = 0.269$, $p < 0.01$), instrumental motivation ($\beta = 0.205$, $p < 0.01$), achievement motivation ($\beta = 0.168$, $p < 0.01$), and self-efficacy ($\beta = 0.205$, $p < 0.01$). As suggested by Hu and Bentler (1999), the indices of this model (CFI = 0.950; PNFI = 0.814; RMSEA = 0.047) indicate an excellent fit to the data.

Figure 2 shows the effects of teacher strictness and teacher feedback on East Asian students' motivational beliefs. Teacher strictness had a weak but positive and significant effect on intrinsic motivation ($\beta = 0.053$, $p < 0.01$), instrumental motivation ($\beta = 0.026$, $p < 0.01$), achievement motivation ($\beta = 0.106$, $p < 0.01$), and self-efficacy ($\beta = 0.026$, $p < 0.01$).

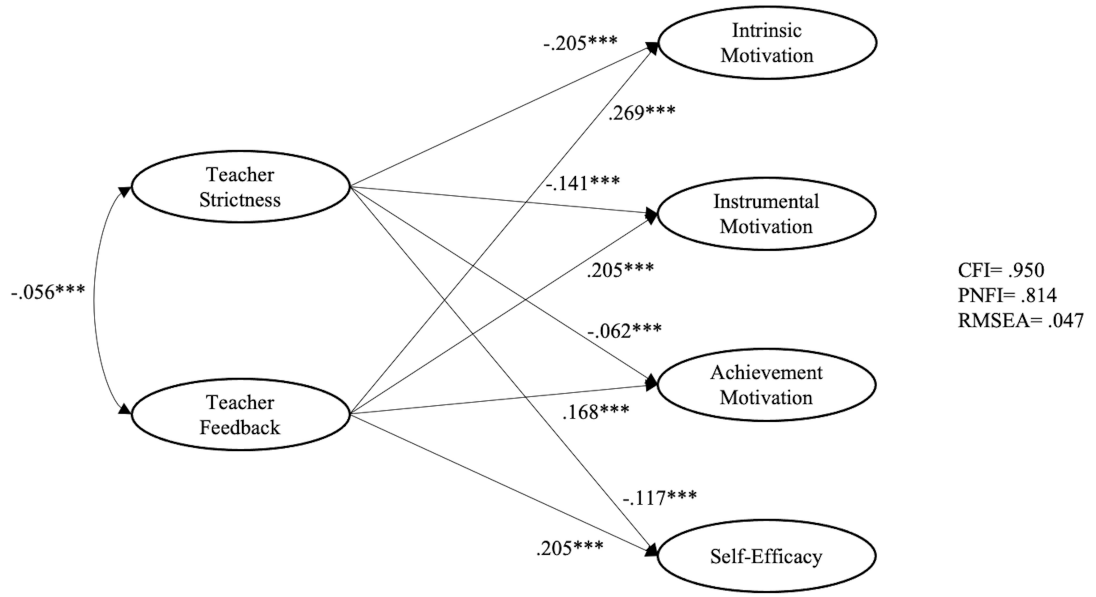

FIGURE 1 | Relationships among teacher feedback, teacher strictness, and motivational beliefs for Western students.

Teacher feedback also had a positive and significant effect on intrinsic motivation (β = 0.310, $p < 0.01$), instrumental motivation (β = 0.298, $p < 0.01$), achievement motivation (β = 0.164, $p < 0.01$), and self-efficacy (β = 0.226, $p < 0.01$). The model fit indices (CFI = 0.945; PNFI = 0.810; RMSEA = 0.050) indicate an excellent fit to the data.

Since considerable differences were found between **Figures 1**, **2** in the relationships between (i) teacher strictness and motivational beliefs and (ii) teacher feedback and strictness, multi-group analysis was performed to examine if such differences were statistically significant across Western and East Asian learners. Four models were imposed, with good model fits (**Table 5**). First, no equality constraints were imposed in the baseline model (M1). Following M1, M2 was imposed by forcing equal constraints on measurement weights. The change in CFI between the models was.002, which is below the.01 threshold for invariance as suggested by Cheung and Rensvold (2002). Equality constraints were further imposed on measurement weights and structural weights (M3). The change in CFI was 0.015, which is above the 0.01 threshold. For the fourth model, equality constraints were imposed on the measurement weights, structural weights, and structural covariance. The change in CFI was 0.001. All these findings indicated that there exists significant difference in the structural weights between Western and East Asian learners.

DISCUSSION

Teacher Feedback, Teacher Strictness, and Students' Motivational Beliefs

As indicated in **Figures 1**, **2**, teacher feedback had significant positive association with both Western and East Asian students' motivational beliefs. For Western students, teacher feedback had the greatest association with intrinsic motivation, followed by self-efficacy, instrumental motivation, and achievement motivation. For East Asian students, teacher feedback had the greatest association with intrinsic motivation, followed by instrumental motivation, self-efficacy, and achievement motivation. This finding is in line with previous findings that teacher feedback is positively related to students' motivational beliefs (e.g., Hamidun et al., 2012; Pat-El et al., 2012).

It is interesting to note that while teacher strictness is negatively related to Western students' motivational beliefs, its relations were significantly positive for Eastern students. This finding echoes the results of Maulana et al. (2011) study in Indonesia that the more teachers exhibit dominance and cooperation, the more students are motivated to engage in learning. However, these results are opposite to findings from the West (e.g., Brekelmans et al., 1993; Wubbels and Brekelmans, 2005).

The disparities in the associations between teacher strictness and students' motivations to learn might be caused by different social and cultural factors in interpreting the roles and expectations of teachers and students. Eastern Asia is characterized by Confucian heritage cultures, such as China, Korea, and Japan (e.g., Kim, 2009; Min, 2016; Xiao and Hu, 2019), which often practice a large power distance (i.e., high acceptance of an unequal power distribution) (Hofstede et al., 2010). Under the influence of Confucianism, Eastern societies often emphasize obedience to authority figures and compliance with group interests (Chang et al., 2011), whereas in Western cultures, the power distance between social members is relatively small (i.e., superiority over others is often considered unacceptable) (Hofstede et al., 2010). Hence, in Western societies, individual thinking and

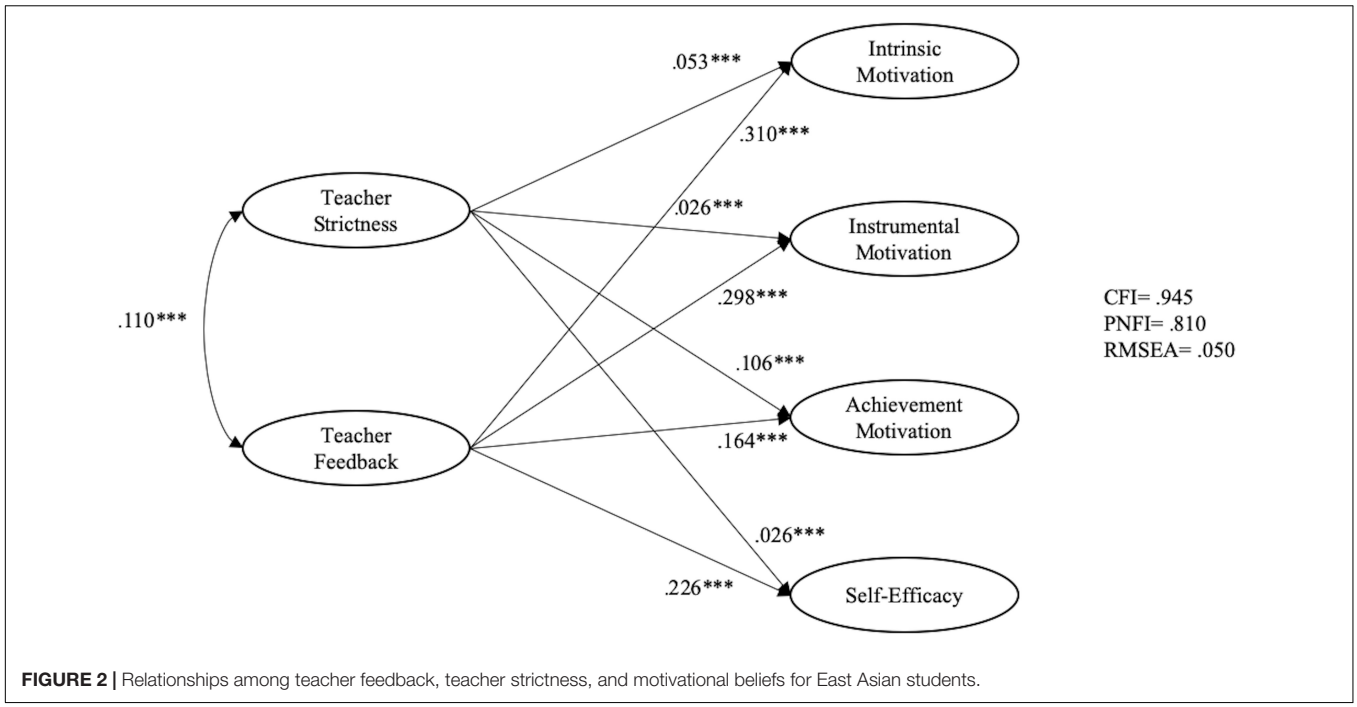

FIGURE 2 | Relationships among teacher feedback, teacher strictness, and motivational beliefs for East Asian students.

TABLE 5 | Summary of goodness-fit statistics for tests of multigroup invariance across the Western and East Asian students.

	RMSEA	PNFI	CFI	Change in CFI
M1 Baseline model (no constraints imposed)	0.034	0.812	0.948	–
M2 Invariant measurement weights	0.034	0.834	0.946	0.002
M3 Invariant measurement weights and structural weights	0.038	0.857	0.931	0.015
M4 Invariant measurement weights, structural weights, and structural covariances	0.038	0.860	0.930	0.001

interest are valued, and individual differences are appreciated (Hofstede et al., 2010).

Cultural differences also exist in the expectations and roles of teachers and students. In the Confucian context, teachers are expected to become the models to help students to realize their good natures, or to introduce models for students to emulate (Shim, 2008). This suggests teachers have significant influences and controls in students' learning as means to cultivate their excellence. Most classrooms in contemporary East Asia are featured as teacher-centered practices. A good teacher is considered as being able to strictly control classroom processes (Zhu et al., 2010; Sun et al., 2018), while a good student is someone who respects and is obedient to their teachers (Zhou et al., 2012). Hence, East Asian students tend to have high expectations and acceptance of teachers' strict or dominant behaviors in class (Hofstede et al., 2010; Wei et al., 2015). In contrast, some of the philosophical underpinnings in Western cultures (such as the philosophy of Socrates) emphasize that teachers should not simply pass on knowledge but also investigate and explore with students together, and that students should be able to think and express their own views, and teachers could correct their views through conversations with them (Shim, 2008). With such philosophical roots, classrooms in Western cultures are often featured as student-centered processes. Teachers are valued for supporting students' autonomy, freedom, and choices, treating each student as a unique individual, as well as maintaining companionate communications with students, while students are valued for demonstrating autonomy and independence in their learning (Hofstede et al., 2010; Chang et al., 2011).

Given the differences in social norms and cultural roots as discussed above, teacher strictness may be more acceptable to East Asian students because it meets their expectations and aligns with the cultural and social values of Eastern societies, which in turn has no negative connections with these students' motivational beliefs. In contrast, Western students may consider teacher strictness to interfere with their freedom, independence, and autonomy in learning (Chan and Rao, 2010). Therefore, when they feel that their teachers are stricter, their learning motivations and efficacy will be reduced.

The positive influence of teacher strictness on East Asian students' motivation can be further explained by the deeply rooted beliefs that teacher strictness is an indication of high expectations from East Asian cultures (Watkins and Biggs, 2001). There is an old Chinese saying, "a strict teacher produces outstanding students" (严师出高徒). This implies that if a teacher is stricter with a student, he or she has high expectations of that student. Therefore, when a Chinese student perceives teacher

strictness, the student may consider it to be recognition of the importance of his or her learning, and so be more confident and motivated to learn. This inference seems to be supported by the positive relationship between teacher strictness and teacher feedback in the East Asian culture revealed in this study. When an East Asian teacher is stricter with some students, more feedback will be provided to these students, which may be commonly regarded by teachers and students as a sign of high expectations. In contrast, but not surprisingly, the relationship between teacher strictness and teacher feedback in the Western culture was found to be weak and negative.

Teacher Strictness and Teacher Feedback

In classroom learning environment, close correlations between teacher interpersonal and teaching behaviors were reported in two previous studies (Cheng and Wan, 2017; Wan and Cheng, 2019). In these two studies, teaching behaviors included assigning challenging tasks to students, stimulating multiple perspectives and encouraging students to communicate with each other, while interpersonal behaviors included sharing control with students and allowing skeptical voice. However, in this study, although the significant correlation was found between teacher strictness and teacher feedback in the East Asian context, their correlation coefficient was rather weak in the Western context. Compared with the two previous studies of classroom learning environment, the correlation coefficients generated in this study were rather small, which indicates the strengths of the correlations between different teacher interpersonal and teaching behaviors may be various.

This study further revealed that with the significant correlation between teacher interpersonal and teaching behaviors (i.e., teacher strictness and teacher feedback), there might still exist cultural differences. The positive significant correlation between teacher strictness and teacher feedback in the East Asian context may be interpreted by the examination-oriented culture. It is well-known that examination culture is prevailing in China and other East Asian regions (Cheng, 2004; Zhan and Wan, 2010). Within such culture, if a teacher is stricter with their students, it implies that he/she has a higher expectation for their students' examination performance, which may cause them to give more feedback to their students so as to enhance their performance. Therefore, it is logical to have a significant and positive correlation between East Asian teacher strictness and feedback. In contrast, in the regions where the examination-oriented culture is not dominant, such a correlation between teacher strictness and teacher feedback may not exist because the connection cannot be established between teacher strictness and a high expectation for students' examination performance.

Implications, Limitations, and Future Studies

Echoing the argument made by Sun et al. (2019) that findings from Western countries may not be directly generalizable to Eastern countries given the multi-layered differences that exist in Western and Eastern cultures, the current study has the following implications. First, although both Western and East Asian teachers should consider giving more constructive feedback to their students, Western teachers should be careful when doing some behaviors that may be perceived as strict by Western students. At the same time, East Asian teachers can be a bit strict to their students, but they should pay attention to the reaction of their students (especially low performing students) since over-strictness may harm their learning motivations and efficacy. Second, there has been a growing trend of pedagogical reform in East Asian countries in recent years, and the reforms usually include adopting sophisticated research findings from mostly Western countries into teaching practice. However, teachers should be cautious and selective when adopting teaching strategies in accordance with local cultural environments. Third, taking a culturally responsive classroom management perspective (Weinstein et al., 2004), it is important for teachers, especially those from Western countries who work in schools with populations of students from multicultural backgrounds, to be conscious of the potential for different interpretations of interpersonal behaviors in different cultures.

A number of limitations should be acknowledged in the current study. The first limitation lays on insufficient information that could be provided by pre-collected PISA data. The current study revealed a significant difference in the effects of teacher behaviors on students' motivational beliefs between East Asian and Western learners. However, given the pre-collected nature of PISA data, there lacks data to further reveal why such difference exists. Second, teachers' behaviors and students' motivational beliefs were retrieved from students' self-reported perceptions, and the results might be limited by self-reported data.

Stemming from current findings, further research should be conducted to explore the complex connections between teacher strictness and teacher feedback utilizing multiple methods. In addition, the current research is a brief report that only make a concise investigation of the issues by comparing the differences between East Asian and Western students. Further investigations may examine not only cross-cultural differences but also intracultural differences using comparative data, such as TIMSS and PIRLS. The comparisons in terms of gender or students' achievement are interesting directions for future research as well.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://www.oecd.org/pisa/data/.

AUTHOR CONTRIBUTIONS

YJ: analysis and interpretation of data for the work and drafting of the work. C-KL: design of the work and revising the draft. ZW: analysis and interpretation of data for the work and revising the draft. JC: analysis and interpretation of data for the work. All authors contributed to the article and approved the submitted version.

REFERENCES

Anderson, J. C., and Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychol. Bull.* 103:411. doi: 10.1037/0033-2909.103.3.411

Brekelmans, M., Sleegers, P., and Fraser, B. J. (2000). "Teaching for active learning," in *New Learning*, eds P. R. J. Simons, J. L. van der Linden, and T. Duffy (Dordrecht: Kluwer), 227–242.

Brekelmans, M., Wubbels, T., and den Brok, P. (2002). "Teacher experience and the teacher–student relationship in the classroom environment," in *Studies in educational learning environments: An international perspective*, eds S. C. Goh and M. S. Khine (Singapore: World Scientific), 73–100. doi: 10.1142/9789812777133_0004

Brekelmans, M., Wubbels, T., and Levy, J. (1993). "Student performance, attitudes, instructional strategies and teacher-communication style," in *Do you know what you look like?*, eds T. Wubbels and J. Levy (London: Falmer Press), 56–63.

Chan, C. K., and Rao, N. (2010). *Revisiting the Chinese Learner: Changing Contexts, Changing Education*. Dordrecht: Springer. doi: 10.1007/978-90-481-3840-1

Chang, L., Mak, M. C., Li, T., Wu, B. P., Chen, B. B., Lu, H. J. et al. (2011). Cultural adaptations to environmental variability: An evolutionary account of East–West differences. *Edu. Psychol. Rev.* 23, 99–129. doi: 10.1007/s10648-010-9149-0

Cheng, K. M. (2004). Examination Culture. *Education* 11:10.

Cheng, M. H. M., and Wan, Z. H. (2016). Unpacking the paradox of chinese science learners: insights from research into asian chinese school students' attitudes towards learning science, science learning strategies, and scientific epistemological views. *Stud. Sci. Edu.* 52, 29–62. doi: 10.1080/03057267.2015.1112471

Cheng, M. M. H., and Wan, Z. H. (2017). Exploring the effects of classroom learning environment on critical thinking skills and disposition: a study of hong kong 12th graders in liberal studies. *Thinking Skills Creativity* 24, 152–163. doi: 10.1016/j.tsc.2017.03.001

Cheung, G. W., and Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Struc. Equ. Model.* 9, 233–255. doi: 10.1207/S15328007SEM0902_5

Creemers, B. P. M. (1994). *The effective classroom*. London: Cassell.

den Brok, P., Brekelmans, M., and Wubbels, T. (2004). Interpersonal teacher behaviour and student outcomes. *School Effect. School Improv.* 15, 407–442. doi: 10.1080/09243450512331383262

Dresel, M., and Haugwitz, M. (2008). A computer-based approach to fostering motivation and self-regulated learning. *J. Exp. Edu.* 7, 3–18. doi: 10.3200/JEXE.77.1.3-20

Fink, A. (2015). *How to conduct surveys: A step-by-step guide*. California: Sage Publications.

Gable, R. K. (1986). *Instrument development in the affective domain*. Boston, MA: Kluwer-Nijhoff.

Gentrup, S., Lorenz, G., Kristen, C., and Kogan, I. (2020). Self-fulfilling prophecies in the classroom: Teacher expectations, teacher feedback and student achievement. *Learn. Instruc.* 66:101296. doi: 10.1016/j.learninstruc.2019.101296

Hamidun, N., Hashim, S. H. M., and Othman, N. F. (2012). Enhancing students' motivation by providing feedback on writing: the case of international students from Thailand. *Int. J. Soc. Sci. Hum.* 2:591. doi: 10.7763/IJSSH.2012.V2.179

Hofstede, G., Hofstede, G. J., and Minkov, M. (2010). *Cultures and Organizations: Software of the Mind. Revised and expanded*, 3rd Edn. New York: McGraw-Hill.

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struc. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Kim, T. (2009). *Confucianism, modernities and knowledge: China, South Korea and Japan. In International handbook of comparative education*. Dordrecht: Springer, 857–872.

Kyriakides, L., Anthimou, M., and Panayiotou, A. (2020). Searching for the Impact of teacher behavior on promoting students' cognitive and metacognitive skills. *Stud. Edu. Evalu.* 64:100810. doi: 10.1016/j.stueduc.2019.100810

Lapointe, J. M., Legault, F., and Batiste, S. J. (2005). Teacher interpersonal behavior and adolescents' motivation in mathematics: a comparison of learning disabled, average, and talented students. *Int. J. Edu. Res.* 43, 39–54. doi: 10.1016/j.ijer.2006.03.005

Lehman, A., O'Rourke, N., Hatcher, L., and Stepanski, E. (2013). *JMP for basic univariate and multivariate statistics: methods for researchers and social scientists*. North Carolina: Sas Institute.

Lewis, R., Romi, S., Katz, Y. J., and Qui, X. (2008). Students' reaction to classroom discipline in Australia, Isreel, and China. *Teach. Teach. Educ.* 24, 715–724. doi: 10.1016/j.tate.2007.05.003

Maulana, R., Opdenakker, M. C., den Brok, P., and Bosker, R. (2011). Teacher–student interpersonal relationships in Indonesia: profiles and importance to student motivation. *Asia Pacific J. Educ.* 31, 33–49. doi: 10.1080/02188791.2011.544061

Min, A. K. (ed.) (2016). *Korean Religions in Relation: Buddhism, Confucianism*. Christianity: SUNY Press.

OECD (2016). *PISA 2015 results (Volume I): Excellence and equity in education*. Paris: OECD Publishing.

Pat-El, R., Tillema, H., and van Koppen, S. W. (2012). Effects of formative feedback on intrinsic motivation: Examining ethnic differences. *Learn. Indiv. Differ.* 22, 449–454. doi: 10.1016/j.lindif.2012.04.001

Pennings, H. J., and Hollenstein, T. (2019). Teacher-student interactions and teacher interpersonal styles: a state space grid analysis. *J. Exp. Educ.* 2019, 1–25. doi: 10.1080/00220973.2019.1578724

Pennings, H. J. M., Van Tartwijk, J., Wubbels, T., Claessens, L. C. A., Van der Want, A. C., Brekelmans, M. et al. (2014). Real-time teacher-student interactions: A dynamic systems approach. *Teachi. Teach. Educ.* 37, 183–193. doi: 10.1016/j.tate.2013.07.016

Quek, C. L., Wong, A. F. L., Divaharan, S., Liu, W. C., Peer, J., Williams, M. D. et al. (2007). Secondary school students' perceptions of teacher-student interaction and students' attitude toward project work. *Learn. Environ. Res.* 10, 177–187. doi: 10.1007/s10984-007-9030-3

Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., Besser, M. et al. (2019). Formative assessment in mathematics: mediated by feedback's perceived usefulness and students' self-efficacy. *Learn. Instruc.* 60, 154–165. doi: 10.1016/j.learninstruc.2018.01.004

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., and Oort, E. J. (2011). The influence of affective teacher students relationships on students' school engagement and achievement. *Rev. Edu. Res.* 81, 493–529. doi: 10.3102/0034654311421793

Shim, S. H. (2008). A philosophical investigation of the role of teachers: a synthesis of plato, confucius, buber, and freire. *Teach. Teach. Edu.* 24, 515–535. doi: 10.1016/j.tate.2007.09.014

Soh, K. (2017). Fostering student creativity through teacher behaviors. *Thinking Skills Creat.* 23, 58–66. doi: 10.1016/j.tsc.2016.11.002

Sun, X., Mainhard, T., and Wubbels, T. (2018). Development and evaluation of a Chinese version of the Questionnaire on Teacher Interaction (QTI). *Learn. Environ. Res.* 21, 1–17. doi: 10.1007/s10984-017-9243-z

Sun, X., Pennings, H. J., Mainhard, T., and Wubbels, T. (2019). Teacher interpersonal behavior in the context of positive teacher-student interpersonal relationships in east asian classrooms: examining the applicability of western findings. *Teach. Teach. Edu.* 86:102898. doi: 10.1016/j.tate.2019.102898

Tsai, W., Sun, M., Wang, S. W., and Lau, A. S. (2016). Implications of emotion expressivity for daily and trait interpersonal and intrapersonal functioning across ethnic groups. *Asian Am. J. Psychol.* 7:52. doi: 10.1037/aap0000043

van Uden, J. M., Ritzen, H., and Pieters, J. M. (2014). Engaging students: the role of teachers beliefs and interpersonal teacher behaviour in fostering student engagement in vocational education. *Teach. Teach. Edu.* 37, 21–32. doi: 10.1016/j.tate.2013.08.005

Wan, Z. H., and Cheng, M. H. M. (2019). Classroom learning environment, critical thinking, and achievement in an interdisciplinary subject: a study of hong kong secondary school graduates. *Edu. Stud.* 45, 285–304. doi: 10.1080/03055698.2018.1446331

Wang, C., Kim, D. H., Bai, R., and Hu, J. (2014). Psychometric properties of a self-efficacy scale for english language learners in China. *System* 44, 24–33. doi: 10.1016/j.system.2014.01.015

Watkins, D. A., and Biggs, J. B. (eds) (2001). *Teaching the Chinese Learner: Psychological and pedagogical perspectives. Hong Kong/Melbourne: Comparative Education Research Centre*. Hong Kong: Australian Council for Educational Research.

Wei, M., Zhou, Y., Barber, C., and Den Brok, P. (2015). Chinese students' perceptions of teacher–student interpersonal behavior and implications. *System* 55, 134–144. doi: 10.1016/j.system.2015.09.007

Weinstein, C. S., Tomlinson-Clarke, S., and Curran, M. (2004). Toward a conception of culturally responsive classroom management. *J. Teach. Edu.* 55, 25–38. doi: 10.1177/0022487103259812

Wubbels, T., Brekelmans, J. M. G., Mainhard, T., den Brok, P. J., and van, J. W. F. (2016). "Teacher-student relationships and student achievement," in *Handbook of social influences in school contexts: social-emotional, motivation, and cognitive outcomes*, eds K. Wentzel and G. Ramani (New York: Taylor and Francis Ltd).

Wubbels, T., and Brekelmans, M. (2005). Two decades of research on teacher-student relationships in class. *Int. J. Edu. Res.* 43, 6–24. doi: 10.1016/j.ijer.2006.03.003

Xiao, Y., and Hu, J. (2019). *The inheritance and spreading of Confucianism in modern China and South Korea. In 2019 5th International Conference on Social Science and Higher Education (ICSSHE 2019)*. Amsterdam, Netherlands: Atlantis Press.

Zhan, Y., and Wan, Z. H. (2010). Perspectives on the cultural appropriacy of assessment for learning in Chinese context. *Educate* 10, 9–16.

Zhou, N., Lam, S. F., and Chan, K. C. (2012). The Chinese classroom paradox: A cross-cultural comparison of teacher controlling behaviors. *J. Edu. Psychol.* 104:1162. doi: 10.1037/a0027609

Zhu, C., Valcke, M., and Schellens, T. (2010). A cross-cultural study of teacher perspectives on teacher roles and adoption of online collaborative learning in higher education. *Eur. J. Teach. Edu.* 33, 147–165. doi: 10.1080/02619761003631849

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer MN declared a shared affiliation with the authors, to the handling editor at the time of review.

APPENDIX

TABLE A1 | Items of measurements.

Construct	Name	Content	Scaling
Teacher strictness	TST1	Teachers called on me less often than they called on other students.	1 = never or almost never; 2 = a few times a year; 3 = a few times a month; 4 = once a week or more
	TST2	Teachers graded me harder than they graded other students.	
	TST3	Teachers disciplined me more harshly than other students.	
	TST4	Teachers gave me the impression that they think I am less smart than I really am.	
Teacher feedback	TFB1	The teacher tells me how I am performing in this course.	1 = never or almost never; 2 = some lessons; 3 = many lessons; 4 = every lesson or almost every lesson.
	TFB2	The teacher gives me feedback on my strengths in the science subject.	
	TFB3	The teacher tells me in which areas I can still improve.	
	TFB4	The teacher tells me how I can improve my performance.	
	TFB5	The teacher advises me on how to reach my learning goals.	
Intrinsic motivation	INT1	I generally have fun when I am learning science topics.	1 = strongly disagree; 2 = disagree; 3 = agree; 4 = strongly agree
	INT2	I like reading about science.	
	INT3	I am happy working on science topics.	
	INT4	I enjoy acquiring new knowledge in science.	
	INT5	I am interested in learning about science.	
Instrumental motivation	INS1	Making an effort in my science subject(s) is worth it because this will help me in the work I want to do later on.	
	INS2	What I learn in my science subject(s) is important for me because I need this for what I want to do later on.	
	INS3	Studying my science subject(s) is worthwhile for me because what I learn will improve my career prospects.	
	INS4	Many things I learn in my science subject(s) will help me to get a job.	
Achievement motivation	ACH1	I want top grades in most or all of my courses.	
	ACH2	I want to be able to select from among the best opportunities available when I graduate.	
	ACH3	I want to be the best, whatever I do.	
	ACH4	I see myself as an ambitious person.	
	ACH5	I want to be one of the best students in my class.	
Self-efficacy	SEF1	Recognize the science question that underlies a newspaper report on a health issue.	1 = I couldn't so this; 2 = I could struggle to do this on my own; 3 = I could do this with a bit of effort; 4 = I could do this easily.
	SEF2	Explain why earthquakes occur more frequently in some areas than in others.	
	SEF3	Describe the role of antibiotics in the treatment of disease.	
	SEF4	Identify the science question associated with the disposal of garbage.	
	SEF5	Predict how changes to an environment will affect the survival of certain species.	
	SEF6	Interpret the scientific information provided on the labeling of food items.	
	SEF7	Discuss how new evidence can lead you to change your understanding about the possibility of life on Mars	
	SEF8	Identify the better of two explanations for the formation of acid rain.	

frontiers in Psychology

ORIGINAL RESEARCH
published: 19 February 2021
doi: 10.3389/fpsyg.2021.581193

Intrinsic Motivation and Sophisticated Epistemic Beliefs Are Promising Pathways to Science Achievement: Evidence From High Achieving Regions in the East and the West

Ching Sing Chai[1], Pei-Yi Lin[2], Ronnel B. King[3] and Morris Siu-Yung Jong[1]*

[1] Department of Curriculum and Instruction, Centre for Learning Sciences & Technologies, The Chinese University of Hong Kong, Shatin, Hong Kong, [2] Department of Education, National Kaohsiung Normal University, Kaohsiung, Taiwan, [3] Faculty of Education, University of Macau, Macau, China

OPEN ACCESS

Edited by:
Bernhard Ertl, Munich University of the Federal Armed Forces, Germany

Reviewed by:
Ricardo Scott, University of Alicante, Spain
Cheng Yong Tan, The University of Hong Kong, Hong Kong
Nikki Shure, University College London, United Kingdom

***Correspondence:**
Pei-Yi Lin
pylin@nknu.edu.tw; hanapeiyi@gmail.com

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *08 July 2020*
Accepted: *01 February 2021*
Published: *19 February 2021*

Citation:
Chai CS, Lin P-Y, King RB and Jong MS-Y (2021) Intrinsic Motivation and Sophisticated Epistemic Beliefs Are Promising Pathways to Science Achievement: Evidence From High Achieving Regions in the East and the West. Front. Psychol. 12:581193. doi: 10.3389/fpsyg.2021.581193

Research on self-determination theory emphasizes the importance of the internalization of motivation as a crucial factor for determining the quality of motivation. Hence, intrinsic motivation is deemed as an important predictor of learning. Research on epistemic beliefs, on the other hand, focuses on the nature of knowledge, and learning with more sophisticated epistemic beliefs associated with more adaptive outcomes. While learning and achievement are multiply determined, a more comprehensive theoretical model that takes into account both motivational quality and epistemic beliefs is needed. Hence, this study aims to examine the role of intrinsic and instrumental motivation alongside epistemic beliefs in predicting students' achievement in science. Data were drawn from the PISA 2015 survey. We focused on four of the top-performing societies. Two were Eastern societies – Singapore and Hong Kong, and the other two were Western societies: Canada and Finland. We found both common and specific patterns among the four societies. Regarding the common patterns, we found that intrinsic motivation and epistemic beliefs had direct positive effects on science achievement. As for the regionally-specific findings, instrumental motivation positively predicted achievement only in Western societies (i.e., Finland and Canada), but not in Eastern societies (i.e., Singapore and Hong Kong). The interaction effect between motivation and epistemic beliefs also demonstrated different patterns across the four societies. Implications for the role of motivation and epistemic beliefs in optimizing student learning and achievement are discussed.

Keywords: intrinsic motivation, instrumental motivation, epistemic beliefs, science achievement, PISA 2015, Eastern and Western learners

INTRODUCTION

Scientific and technological advances have greatly improved human life. In addition, emerging global issues such as the Covid-19 pandemic, global warming, and food shortage could only be resolved with more people having strong scientific knowledge and scientific ways of knowing. Despite the critical importance of science, not many students aspire to become scientists

(Nugent et al., 2015). Moreover, there is a worrying trend that students' intrinsic motivation to learn science in school and their aspiration to engage in a science-related career declines from elementary to high school (Alexander et al., 2012; Potvin and Hasni, 2014). Students may also possess unsophisticated assumptions about what science is and how it works (Li et al., 2018). Hence, there is a clear need to look into students' motivation and science-related epistemic beliefs.

Identifying the factors that would optimize science learning and achievement is an urgent educational issue. In this study, we focus particularly on the role of motivation and epistemic beliefs in predicting science achievement. This study is novel because it integrates research on motivation which usually focuses on *why* students learn science with research on epistemic beliefs which pertains to students' perceptions of *what* science is. While these two bodies of research have been quite active (e.g., Chen et al., 2014; Lin and Tsai, 2017), there is little research attempt to study them together. There is theoretical value in exploring their synergies as science learning is likely to be multiply determined. Researchers have increasingly warned against devoting exclusive attention to one key variable and neglecting a broader view of the critical factors underpinning key outcomes (Pettigrew and Hewstone, 2017; Yarkoni and Westfall, 2017).

Students who have high levels of motivation have a "why" for engaging in science-related learning activities. These students might be either intrinsically motivated as they just love learning science for its own sake or instrumentally motivated when they engage in learning science to advance their careers or to graduate from school. However, being motivated might not be enough to yield high levels of achievement. Numerous studies have shown that the relationship between motivational factors and achievement though statistically significant is smaller than other psycho-educational factors (e.g., Hulleman et al., 2010; Howard et al., 2017; Kriegbaum et al., 2018). This suggests the need to examine other potentially important factors that underpin science achievement and learning.

This brings us to the importance of recognizing that optimal science learning happens when students have a strong *why* (i.e., motivation) but also have a sophisticated understanding *what* science and scientific knowledge is all about. The investigation of epistemic beliefs about scientific knowledge is increasingly important in a post-truth society where scientific truth is contested and when an increasing number of people hold unscientific beliefs (e.g., Hornsey and Fielding, 2017; Hornsey et al., 2018). For example, researchers have found a large number of individuals holding anti-vaccination beliefs and harboring skepticism about climate change (Ecklund et al., 2017; Rizeq et al., 2020). These trends are associated with a strong resistance to evidence-based reasoning posing serious threats to societal progress (Hornsey and Fielding, 2017). Research on epistemic beliefs may hold potential implications for these critical problems (Hartman et al., 2017; Wilson, 2018).

Hence, the main research objective of this study is to explore the role of both motivation and epistemic beliefs in predicting science learning. To achieve this objective, we analyze data from the Program for International Student Assessment (PISA) from four different regions (Singapore, Hong Kong, Finland, and Canada) representing high-achieving societies across both East and West thereby allowing us to identify the possible cross-cultural factors that are common in predicting science achievement.

This study also addresses methodological shortcomings of past research. Past studies on science learning and achievement have been hampered by their exclusive focus on one cultural context (Chen et al., 2014; Lin and Tsai, 2017; Wong et al., 2019; Kaderavek et al., 2020). Hence, the possible cross-cultural applicability of the results might be questioned. This is a particularly important issue as researchers have shown the importance of culture in influencing students' learning and motivational processes and their epistemic beliefs (Zusho and Clayton, 2011; Lee et al., 2012; King and McInerney, 2014; King et al., 2018).

In an attempt to address how epistemic beliefs may influence academic achievement, Greene et al.'s (2018) meta-analysis revealed that sophisticated epistemic beliefs (i.e., adaptive view on the development and justification of knowledge as constructed and evidence-based) are more influential of academic achievement than unsophisticated epistemic beliefs (i.e., view knowledge as absolute and certain). In addition, most epistemic beliefs studies have primarily relied on self-reports (see Debacker et al., 2008), and its effect on achievement needs to be further explored. It thus becomes important to identify the generalizations and contextually adaptive views on knowledge and knowing when assessing what constitutes a set of sophisticated beliefs in a certain discipline.

Hence, the purpose of this study was to empirically examine an integrated theoretical framework to assess whether students' motivations, epistemic beliefs, and the interaction between their motivation and epistemic beliefs are predictive of science achievement across different societies representing different cultures.

LITERATURE REVIEW

Motivation to Learn Science

Student motivation to refers to why students undertake a learning task (Deci and Ryan, 2000; Pintrich, 2003). Though motivation is a complex phenomenon, self-determination theory suggests a common model that explains the process of how learners' innate behavior and inherent propensity drive them to accomplish the desired educational outcomes (Elliott and Dweck, 1988; Deci et al., 1991; Deci and Ryan, 2000). Students who are intrinsically motivated view learning science as interesting and working on scientific issues enjoyable (Ryan and Deci, 2009). Studies have shown that students who are intrinsically motivated in science participate more in science-related activities (Lin and Schunn, 2016), and these factors would consequently influence students' science achievement (Burns et al., 2019).

On the other hand, instrumental motivation (also called utility value) to learn science reflects students' desire to learn science as a means to achieve a certain goal (i.e., to pursue further studies or for career progression) (Nagengast and Marsh, 2014). Instrumental motivation is a predictor of achievement

and career choice (Canning et al., 2018). Previous research supports that students were more likely to learn science when they perceived the instrumental value of studying science in order to attain STEM-related career expectations or have successful work outcomes later on (Rozek et al., 2015). Nonetheless, instrumental motivation seems to have weaker association with science achievement compared to intrinsic motivation (Liang and Tsai, 2010).

More importantly, the two types of motivations could co-exist; an individual can be both instrumentally and intrinsically motivated (Hidi and Harackiewicz, 2000). In this study, we investigated motivational variables (i.e., intrinsic motivation and instrumental motivation) in predicting science achievement across the selected societies.

Epistemic Beliefs About Science

Epistemology is a sub-discipline of philosophy that is concerned with the nature and grounds of knowledge, and ways of knowing (Hofer, 2002). Within the fields of psychology and education, epistemic beliefs focus on students' beliefs about the nature of knowledge and knowing process (Schommer, 1990; Hofer and Pintrich, 1997; Hofer, 2002). The evolution of the thinking process about knowledge and knowing has become prominent in science education (Scott et al., 2006; Lin and Tsai, 2017). In general, science epistemic beliefs are associated with students' scientific reasoning, interpreting and justifying scientific ideas based upon empirical evidence and through critical thinking (Hofer and Pintrich, 1997).

In this study, epistemic beliefs are posited as students' beliefs about science and scientific knowledge. This involves how students scientifically explain phenomenon, interpret data and evidence, and approach science issues (OECD, 2016a). Students with sophisticated epistemic beliefs are more likely to hold intrinsic goal orientation to make inferences and comparisons from one or multiple texts, construct perspectives from integrated information, and apply scientific ideas and concepts to make evaluations and justifications (Paulsen and Feldman, 2005; Chen and Pajares, 2010; Tsai et al., 2011). Most importantly, sophisticated epistemic belief entails an understanding about the evolving and constructed nature of scientific knowledge (Muis, 2007; Krist, 2020).

Sophisticated epistemic beliefs about science generally are associated with higher levels of achievement (Greene et al., 2018). In addition, middle-school students with sophisticated epistemic beliefs undertake scientific inquiry in a qualitatively different manner. They could use scientific standards to provide insights into their understanding of the explanatory and descriptive goals, conceptual coherence and clarity, and empirically evidence evaluation for scientific models (Pluta et al., 2011; Belland et al., 2016).

In the domain of science, PISA measures students' sophisticated epistemic beliefs about science as tentative and evolving. Epistemic beliefs encompass students' views about the need for scientific experiments to justify scientific knowledge, and a recognition of the limitations of scientific experiments (OECD, 2016a). The investigation of epistemic beliefs about science is extremely important in the context of a post-truth society where it is imperative that students develop the skills to evaluate scientific evidence and explanations (Sinatra and Lombardi, 2020).

Relationship Between Motivations and Epistemic Beliefs

Past research has explored the associations among learning motivation, epistemic beliefs and achievement, and indicated that students' motivation and how they view science impact the learning process (e.g., Chen, 2012; Mason et al., 2013; Ho and Liang, 2015). Research has indicated that students with a strong intrinsic motivation tend to invest their time and effort in seeking in-depth understanding (Chen and Pajares, 2010; Burns et al., 2019). For example, students' intrinsic motivation is associated with adopting constructive learning strategies to construct scientific knowledge (Lin et al., 2013; Ho and Liang, 2015; Shen et al., 2018). Nonetheless, the decline of students' motivation to learn science (Vedder-Weiss and Fortus, 2011) and promotion of students' sophisticated epistemic beliefs (Lee et al., 2016) are critical issues. Therefore, this study aimed to explore the generalizability of motivation, epistemic beliefs, and achievement across societies.

Commonality and Specificity

A critical issue in examining the pattern of relationships among the variables is whether they are common across cultures or whether they are culturally-specific. Much of the existing research in motivation and epistemic beliefs have been conducted in WEIRD (Western, educated, industrialized, rich democratic societies) (Henrich et al., 2010). Though many motivational phenomena are commonly observed across different cultures (e.g., Pintrich, 2003), students may also have different motivational orientations (e.g., Brown et al., 2018; Liu et al., 2020). The critical factors that underpin learning and achievement are also strongly influenced by sociocultural factors (Chiu and Chow, 2010; Chiu et al., 2016; King and McInerney, 2019; Li and Yamamoto, 2020). Hence, it is important to test the cross-cultural applicability of the models (King and McInerney, 2014; King et al., 2018).

Science epistemic beliefs, which refer to individuals' beliefs about the nature of knowledge and knowing has been found to be associated with cultural factors (Hofer, 2008). For example, Schommer-Aikins and Easter (2008) argued that Euro-American students had significantly higher epistemic belief scores (i.e., student beliefs about the speed of knowledge acquisition and knowledge construction and modification) compared to Asian American students. More recently, Yang (2016) reviewed 106 studies and concluded that there are cultural differences with epistemic beliefs in the context of science learning. More specifically, it seems that American and Taiwanese students may have more sophisticated epistemic beliefs, while Turkish and Chinese students may rely more on authority.

Cultural differences are also reflected in teaching practices. In Asian educational contexts, science learning is dominated by

traditional didactic approaches wherein students are asked to provide certain and correct answers (e.g., Ho and Liang, 2015). In contrast, science learning in Western societies is more dominated by inquiry-based approaches which could foster more sophisticated epistemic beliefs (Yang, 2016). Hence, further comparative work is needed to explore the contextually and culturally situated nature of epistemic beliefs.

SCIENCE LEARNING CONTEXT IN SINGAPORE, HONG KONG, FINLAND AND CANADA

Given the excellent performance by Singapore, Finland, Canada, and Hong Kong in the science literacy test in PISA 2015, an introduction to the four societies' science learning context will allow for better interpretations of students' learning motivation, science epistemic beliefs, and its relation to science literacy. We focus on these four societies given that they represent high-performing regions in the West and the East. Moreover, all four societies are considered highly economically developed thereby minimizing potential confounds.

We are aware that these four societies do not completely represent the West and the East as there are numerous countries that could be classified into West–East. Hence, we invite readers to be cautious in making over-generalizations. Adding too many societies, however, would work against model parsimony as there might be too many country-level confounds that might potentially bias the results (e.g., differences in geography, cultural values, climate, political system, demographic profile) (e.g., Oishi, 2014; Krems et al., 2017). For example, though Estonia is also a top-performing Western country, the country's governance and cultural values differ from Canada and Finland. Similarly, one could classify Vietnam as a top-performing Eastern country but it is demographically very different from Hong Kong and Singapore which both share a British colonial history and have relatively similar economic profiles. Bearing this caveat in mind, we discuss each of the four societies we included in our study.

Singaporean Context

In the Singapore education system, science classes start in the 3rd grade and in secondary schools, students will learn general science until the eighth grade. The center of science education is focused on promoting "science as an inquiry" for students to relate science to society, daily life, and the environment (Ministry of Education [MOE], 2013, 2014). The curriculum emphasizes students' acquisition of science knowledge, understanding, and application; scientific skills and knowing processes; and scientific attitudes with ethical handling of scientific issues (Ministry of Education [MOE], 2013, 2014). Recent studies indicate that the inquiry-oriented science pedagogy enhances Singaporean students' interest in school science and science learning (Jocz et al., 2014; Sun et al., 2016). Nonetheless, only a small group of students in Singapore reported that they like learning science in TIMSS 2011 assessment, and thus examination of students' science motivation is considered in the current study (Lay and Chandrasegaran, 2016).

Hong Kong Context

The science education system in Hong Kong is implemented in the general studies curriculum at the primary school level that integrates the disciplines of social science, science, and technology, and at the secondary school level, science education is positioned to strengthen students' science knowledge and ability to integrate and apply science knowledge across disciplines (Curriculum Development Council, 2017). Science inquiry is positioned as a pedagogical means to engage students in acquiring science knowledge and advanced scientific skills (Wan and Lee, 2017; Cheung, 2018), and to prepare students' readiness for the workplace and solving daily life problems (Jong, 2017; So et al., 2018). Moreover, the national science curriculum guidelines also highlights the importance of enhancing students' motivation through connecting science-related issues to their daily life, and encourages teachers to adopt inquiry-based, or hands-on activities to develop students' interest in science (Curriculum Development Council, 2017).

Finnish Context

In Finland, primary science education (Grades 1–6) is taught as an integrated course that aims to transmit the nature of science (Finnish National Agency for Education, 2017). At the secondary school level, science could be taught as an integrated subject or as more specialized into the separate subjects of physics, chemistry, geography, and biology (Lavonen and Juuti, 2016). The Finnish science curriculum may be characterized as an inquiry- or context-based approach to raise students' interest and motivation toward science subjects (Kang et al., 2019; Lehtinen et al., 2019). It highlights the importance of personal relevance by linking science content to their lives, which apparently leads to a positive correlation with interest and achievement (Kang and Keinonen, 2018). Past research showed that, compared to students in the United States, Finnish students felt confident, successful, and happy during their science classes (Schneider et al., 2016).

Canadian Context

In Canada, science education varies across the 13 jurisdictions (Milford and Tippett, 2019). The Council of Ministers of Education, Canada (2016) aims that students develop (i) an understanding of the nature of science, technology, society, and the environment (STSE), (ii) scientific and technological inquiry, (iii) knowledge in life sciences, physical sciences, and earth and space sciences, and (iv) attitudes that support the scientific and technological acquisition and application. Studies have shown that Canadian students are able to extend and deepen their understanding of fundamental science concepts and learn to use science knowledge and processes as a scientist does (Hasni et al., 2016; Asghar et al., 2019). On the other hand, in a local study conducted by Potvin and Hasni (2014), there is a slight decrease of students' interests in science learning from 5-grade through 11-grade.

The present study includes data from the four top-performing countries and regions, and aims to investigate whether there

is a general relationship with the four factors – science epistemic beliefs, intrinsic motivation, instrumental motivation, and science achievement – assessed in PISA 2015. The research questions are:

1. Do students' science motivations (i.e., intrinsic and instrumental motivation) predict their science achievement?
2. Do students' epistemic beliefs predict their science achievement?
3. Do students' motivations, epistemic beliefs, and the interaction between motivations and epistemic beliefs predict their science achievement?

METHODS

Sample

The sample for this study adopted data released from the PISA 2015 database. PISA 2015 measured how 15-year-old students in 72 participating countries and regions meet the challenges of today's knowledge societies (OECD, 2016a). In 2015, science was the major assessment domain. The present study includes two Eastern societies – Singapore, Hong Kong – and two Western societies – Finland and Canada – from the top-10 performing countries and regions in PISA 2015 to validate a cross-contexts comparison. The total number of participants from all participating countries and regions was 418,458 students (50.1% female). In this study, we only focused on four societies: the Singapore 5,748 students (48.6% female); Hong Kong 5,011 students (49.9% female), Canada 17,220 students (50% female), and Finland 5,060 students (48.7% female).

Variables

The Program for International Student Assessment is an international assessment administered by the OECD. PISA data were examined in different analyses to ensure the quality of data meet designed criteria. Research also has used PISA 2015 to provide insight into students' science learning and literacy (Aditomo and Klieme, 2020; Tang and Zhang, 2020). In the current study, variables were chosen from the student questionnaire in PISA 2015. This study includes the following variables taken from the student questionnaire in PISA 2015.

Intrinsic Motivation to Learn Science

Intrinsic motivation pertains to students' enjoyment of engaging in science learning activities based on their responses to questions such as whether they have fun when learning science topics, like reading about science, enjoy learning new science topics and acquiring new knowledge in science. PISA 2015 measures students' enjoyment of learning science through a four-point Likert scale from "1 = *strongly disagree*" to "4 = *strongly agree*." A sample item is, "I have fun when I am learning <broad science>." Higher levels agreement indicates that students enjoy learning science for its own sake. Reliabilities (Cronbach's α) measured in this study ranged from 0.93 to 0.96, which was in line with OECD's technical report (2016b).

Instrumental Motivation to Learn Science

Instrumental motivation measured students' agreement to whether that making an effort to learn science is worthwhile because school science is helpful for later-on work and career plans. Students' responses on a four-point Likert scale with categories from "1 = *strongly agree*" to "4 = *strongly disagree*." The responses were reverse-coded so that higher values refer to higher levels of instrumental motivation. A sample item is, "Studying my <school science> subject(s) is worthwhile for me because what I learn will improve my career prospects." Reliabilities (Cronbach's α) measured in this study ranged from 0.91 to 0.95, which was in line with OECD's technical report (2016b).

Epistemic Beliefs About Science

Epistemic beliefs about science investigated students' views on scientific approaches, understanding of scientific knowledge as derived from experimentation, and that scientific knowledge is revisable based on the experimental evidence. A four-point Likert scale with the answering categories from "1 = *strongly disagree*" to "4 = *strongly agree*" was measured. A sample item is, "Good answers are based on evidence from many different experiments." Higher levels of agreement indicate that students possess more sophisticated epistemic beliefs about science. Reliabilities (Cronbach's α) measured in this study ranged from 0.88 to 0.91, which was in line with OECD's technical report (2016b).

Science Achievement

The PISA 2015 science achievement score was viewed as the cognitive learning outcome in this study. The PISA 2015 described a clear framework in measuring students' scientific competencies (i.e., explain phenomena scientifically, evaluate and design scientific inquiry, and interpret data and evidence scientifically). The test content is not confined by school science content, but rather by contexts and problems for which science knowledge, scientific methods can be applied.

Data Analyses

Data were analyzed in accordance with the research questions of the study. Firstly, the univariate normality was examined in accordance with Kline's (2005) criteria. The values of skewness (ranged from −1.02 to −0.79) and kurtosis (ranged from −0.70 to 0.10) (see **Table 1**) indicated the dataset was normally distributed following the recommended value that skewness and kurtosis should be under | 3| and | 10|, respectively. In the preliminary analyses, exploratory factor analyses (EFA) with SPSS version 21 (IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY, United States: IBM Corp.) were employed to examine the construct validity of the responses to the Singapore, Hong Kong, Finland, and Canada datasets. A-three factor (i.e., intrinsic motivation, instrumental motivation, and epistemic beliefs) model was established.

Pearson's correlation analysis was conducted. In PISA 2015, there were 10 plausible values that presented students' achievement, we conducted plausible values analysis using each plausible value separately, then, computed and averaged them (OECD, 2009). Multilevel modeling is used to analyze data

TABLE 1 | Means and SD of measured items.

	All PISA participants (*N* = 418458)		Singapore (*n* = 5748)		Hong Kong (*n* = 5011)		Canada (*n* = 17220)		Finland (*n* = 5060)	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
1. Intrinsic motivation	2.73	0.78	3.01	0.68	2.80	0.75	2.85	0.80	2.57	0.73
2. Instrumental motivation	2.90	0.79	3.08	0.65	2.85	0.77	3.04	0.78	2.79	0.75
3. Epistemic beliefs	3.02	0.58	3.15	0.50	3.06	0.54	3.18	0.59	2.99	0.56
Skewness	−0.90	−0.25	−0.94	−0.51	−0.79	−0.28	−1.02	−0.26	−0.84	−0.03
Kurtosis	−0.19	1.04	0.10	2.08	−0.47	2.25	−0.70	1.65	−0.55	1.91

because students were nested in schools. This study employed a two level model (level 1 = student level, level 2 = school level) to examine the influence of schools on students' science achievement. We ran four multilevel models for each region. The first model was a null model to partition the between- and within-groups variance in science achievement. The intra-class correlation coefficient (ICC) is the ratio of between-group variance to the total variance. In the second model, we specified a random intercept model. The following level 1 predictors were included: gender, students' economic, social, and cultural status (which is based on students' scores in PISA 2015 ESCS measure) intrinsic motivation, and instrumental motivation. The third model is also a random intercept model that included the following predictors: gender, ESCS and epistemic beliefs. The fourth model is a full model including interaction effects. The predictors were: gender, ESCS, intrinsic motivation, instrumental motivation, epistemic beliefs, and interaction between motivation and epistemic beliefs. Gender and ESCS measure are controlled as covariates to predict science achievement in the model 2 to 4. The data file downloaded from OECD website[1] is weighted at the student level with normalized student final weights (OECD's technical report, 2016b) and listwise deletion is used to treat missing data.

RESULTS

The results are presented in the following sections. First, preliminary analyses included the EFA and the bivariate correlations, established a structural model and explored relationships between students' intrinsic motivation, instrumental motivation, and epistemic beliefs among the four countries and regions. The main analyses were about examining how intrinsic motivation and instrumental motivation, and epistemic beliefs and their interactions predict science achievement.

Establishing the Factor Structure

We first tested the factor structure using exploratory factor analysis to examine the factors of the measurement. Principal axis factor analyses with direct oblimin rotation were run on the data. A three-component structure among the four selected societies was established. Follow Hair et al.'s (2010) recommendation, three latent factors were specified by factor loadings greater than 0.5, and eigenvalues greater than one. The intrinsic motivation includes five items, the instrumental motivation includes four items, and the epistemic beliefs includes six items, of the measurement are listed in **Appendix 1**.

The Kaiser–Meyer–Olkin (KMO) Value and Bartlett's test of sphericity were calculated before the EFA to determine the applicability of the factor analyses. In the present study, all KMO values greater than 0.50 (KMO = 0.91 in the Singapore, Canada, and Finland dataset; KMO = 0.93 in the Hong Kong dataset; see **Table 2**) indicated that factor analysis sampling was appropriate. Bartlett's test of sphericity indicated significance for EFA (X^2 = 67809.839, df = 105, $p < 0.001$ in the Singapore dataset; X^2 = 72282.874, df = 105, $p < 0.001$ in the Hong Kong dataset; X^2 = 219084.577, df = 105, $p < 0.001$ in the Canada dataset; X^2 = 61184.599, df = 105, $p < 0.001$ in the Finland dataset; see **Table 2**). Factor loadings of measured items ranged from 0.68 to 0.94 in the Singapore dataset; ranged from 0.70 to 0.94 in the Hong Kong dataset; ranged from 0.72 to 0.93 in the Canada dataset; and ranged from 0.72 to 0.91 in the Finland dataset (see **Table 2**). Total explained variance was found to be 68.82% in the Singapore dataset; 75.80% in the Hong Kong dataset; 73.05% in the Canada dataset; and 71.23% in the Finland dataset.

Next, we addressed the relationships among the latent factors in **Table 3**. Correlations were computed for Singaporean, Hong Kong's, Canadian, and Finnish students. Intrinsic motivation, instrumental motivation and epistemic beliefs were all positively and significantly correlated (ranging from 0.09 to 0.50). The lowest correlation was found for the association between instrumental motivation and science achievement in the Singapore dataset. High correlations between intrinsic motivation and instrumental motivation were found in the four societies. Regarding science achievement, epistemic beliefs were strongly and positively correlated to science achievement in the Hong Kong and Finland datasets, whereas intrinsic motivation was found to be strongly and positively related to science achievement in the Singapore and Canada datasets.

Predicting Students' Science Achievement

We hypothesized that (i) motivations (i.e., intrinsic and instrumental motivation), (ii) epistemic beliefs, and (iii) their interactions would predict science achievement when entered separately into the regression equation. We analyzed the

[1]https://www.oecd.org/pisa/data/2015database/

TABLE 2 | EFA of measured items.

	Singapore		Hong Kong		Canada		Finland	
	Factor loadings	**% of variance**	**Factor loadings**	**% of variance**	**Factor loadings**	**% of variance**	**Factor loadings**	**% of variance**
1. Intrinsic motivation	0.87–0.94	40.20	0.80–0.94	46.67	0.85–0.92	40.55	0.81–0.91	38.98
2. Instrumental motivation	0.77–0.90	12.56	0.88–0.93	10.94	0.84–0.93	12.48	0.86–0.91	12.48
3. Epistemic beliefs	0.68–0.77	16.07	0.70–0.86	18.20	0.72–0.83	20.03	0.72–0.82	19.77
Kaiser-Meyer-Olkin value	0.91		0.93		0.91		0.91	
Bartlett's test of sphericity	X^2 = 67809.839 df = 105 $p < 0.001$		X^2 = 72282.874 df = 105 $p < 0.001$		X^2 = 219084.577 df = 105 $p < 0.001$		X^2 = 61184.599 df = 105 $p < 0.001$	
Total % of variance	68.82		75.80		73.05		71.23	

TABLE 3 | Correlations of motivations, epistemic beliefs, and science achievement.

	1	2	3	4	1	2	3	4
1. Intrinsic motivation	–	0.50	0.41	0.26	–	0.41	0.29	0.32
2. Instrumental motivation	0.39	–	0.26	0.12	0.41	–	0.15	0.18
3. Epistemic beliefs	0.37	0.21	–	0.28	0.32	0.16	–	0.37
4. Achievement	0.33	0.09	0.29	–	0.33	0.16	0.31	–

All correlations were statistically significant at $p < 0.001$. The lower triangle in the left column is the Singapore data; the upper triangle in the left column is the Hong Kong data. The lower triangle in the right column is the Canada data; the upper triangle in the right column is the Finland data.

predictive effect of science achievement using four models to respectively, answer our three research questions in **Table 4**. The ICC for model 1 was 35% in Singapore, 32% in Hong Kong, 16% in Canada, and 9% in Finland. The intercepts varied significantly across schools (Wald $Z = 8.53$, $p < 0.001$, in Singapore; Wald $Z = 7.83$, $p < 0.001$, in Hong Kong; Wald $Z = 15.45$, $p < 0.001$, in Canada; Wald $Z = 6.44$, $p < 0.001$, in Finland). The results support the use of multilevel modeling.

In model 2, gender and ESCS were control covariates. Intrinsic motivation and instrumental motivation were entered as independent variables. The results indicated that intrinsic motivation significantly predicted science achievement across the four societies. Instrumental motivation was a negative predictor of science achievement in Singapore, yet, a positive predictor of science achievement in Canada and Finland.

In model 3, we found that epistemic beliefs were a positive predictor of science achievement across the four societies.

In model 4, intrinsic motivation and instrumental motivation, epistemic beliefs, and interaction of motivations and epistemic beliefs (i.e., intrinsic motivation × epistemic beliefs and instrumental motivation × epistemic beliefs) were entered as predictors of science achievement. In this model, intrinsic motivation and epistemic beliefs were both positively associated with science achievement across the four societies.

Gender differences across cultures were also observed. Males had higher science scores in Hong Kong and Canada, females scored higher than males in Finland. In Singapore, there was no gender difference. ESCS was a positive predictor in the four regions.

However, cross-cultural differences were observed as regards instrumental motivation. Instrumental motivation positively predicted science achievement only in the Western countries such as Canada and Finland. In the East (Singapore), instrumental motivation was a negative predictor. Instrumental motivation was not significantly related to achievement in Hong Kong.

To help with the interpretation of the finding, **Figure 1** illustrates the interaction between intrinsic motivation and epistemic beliefs, while **Figure 2** depicts the interaction between instrumental motivation and epistemic beliefs. The *X*-axis represents motivation (intrinsic or instrumental), while the *Y*-axis represents achievement. Science achievement was particularly high when both intrinsic motivation and epistemic beliefs were high in the four societies. This demonstrates that the factors are important across the four regions.

We also found culturally specific findings. In Singapore, students' epistemic beliefs had a stronger association with achievement when instrumental motivation was low. In Singapore and Canada, students' instrumental motivation had a stronger association with achievement when they had less sophisticated epistemic beliefs (e.g., −1SD and −2SD below the mean). In Finland, students' instrumental motivation had a stronger association with achievement when they had sophisticated epistemic beliefs (e.g., +1SD and +2SD below the mean).

DISCUSSION

We examined the associations among intrinsic motivation, instrumental motivation, epistemic beliefs, and their interactions to predict science achievement in a large sample of 15-years

TABLE 4 | Multilevel analyses for predicting science achievement.

	Singapore				Hong Kong			
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Intercept	545.07***	456.69***	427.90***	363.77***	524.81***	461.57***	433.44***	422.87***
Level 1								
Gender (female)		−4.25	−7.82**	−4.02		−7.18***	−11.82**	−8.00***
ESCS (SES)		24.17***	25.47***	23.23***		3.22**	3.63**	2.58*
INTR		35.61***		20.05**		25.48***		12.73*
INST		−4.48**		12.19*		−0.51		−0.35
EB			39.19***	37.65***			32.89***	19.38***
INTR × EB				3.20				2.17
INST × EB				−6.03**				−0.74
Residual variance	7151.25	6133.35	6282.84	5953.52	4398.33	3982.80	4015.12	3849.49
Intercept variance (School level)	3838.69	2317.38	2373.90	2184.47	2027.22	1881.81	1779.16	1800.18
Intra-class correlation	0.35	0.27	0.27	0.27	0.32	0.32	0.31	0.31
Model fit: −2 Log likelihood	72142.68	68465.16	68308.00	66607.25	60563.80	56819.86	57234.82	55655.61
	Canada				**Finland**			
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 1**	**Model 2**	**Model 3**	**Model 4**
Intercept	514.37***	413.86***	384.11***	329.27***	530.58***	426.52***	374.60***	373.44***
Level 1								
Gender (female)		−2.75*	−6.01***	−4.29**		14.79***	10.52***	11.41***
ESCS (SES)		20.63***	22.17***	19.20***		30.24***	29.78***	26.10***
INTR		31.43***		18.60***		33.59***		12.72*
INST		2.63**		12.04**		3.58*		−8.96
EB			40.39***	34.04***			50.39***	19.98**
INTR × EB				2.03*				4.31
INST × EB				−3.23**				3.91*
Residual variance	7086.60	5970.55	6113.90	5703.93	8384.53	6827.76	6663.75	6251.62
Intercept variance (School level)	1361.85	826.82	736.05	722.14	797.44	355.39	275.04	266.65
Intra-class correlation	0.16	0.12	0.11	0.11	0.09	0.05	0.04	0.04
Model fit: −2 Log likelihood	236063.32	207675.21	208073.18	198240.75	70065.23	62252.85	61088.81	58630.71

*INTR, intrinsic motivation; INST, instrumental motivation; EB, epistemic belief; Gender: Control variable (1 = female, 2 = male). All the residual variance and intercept variance are significant at ***p < 0.001 level. *p < 0.05; **p < 0.01; ***p < 0.001.*

old students across four societies. Given that studies about the interrelationships these factors are usually culturally specific (Chen et al., 2014; Lin and Tsai, 2017; Wong et al., 2019; Kaderavek et al., 2020), this study first established the construct validity of the factors measured for the four different societies. This effort allows us to discuss the findings with some confidence about cross-cultural applicability.

In relation to the first research question, we found empirical support that intrinsic motivation is predictive of students' science achievement for the four regions. This finding extends the current understanding that intrinsic motivation could be a common factor that predicts science learning achievement (Ryan and Deci, 2009; Lin and Schunn, 2016; Burns et al., 2019). A practical implication of this finding is that teachers are encouraged to foster students' intrinsic motivation to learn science regardless of cultural or contextual differences.

As for instrumental motivation, our findings indicate that it was positively associated with achievement in Canada and Finland, yet negatively associated with achievement in Singapore. The case of Canada and Finland may reflect a stronger emphasis in Western societies about the use of instrumental motivation to encourage students to learn science (Rozek et al., 2015; Canning et al., 2018). In the Asian context, Liang and Tsai (2010) reported a weaker association between instrumental motivation and achievement. Our finding also indicates that instrumental motivation is not a significant predictor for Hong Kong students' achievement when both forms of motivation are considered. However, in model 4, instrumental motivation is a negative predictor for the Hong Kong sample. There could be a higher emphasis on the instrumental value of science in Hong Kong (So et al., 2018). In general, it seems that leveraging on instrumental motivation may not enhance students' achievement in the two Eastern regions. In addition, given that the correlation between the achievement and the instrumental motivation is significant and positive (r = 0.09 for Singapore), the negative regression weight for the Singapore sample could be due to suppression effects. The situation warrants more specific cross-cultural research in this area.

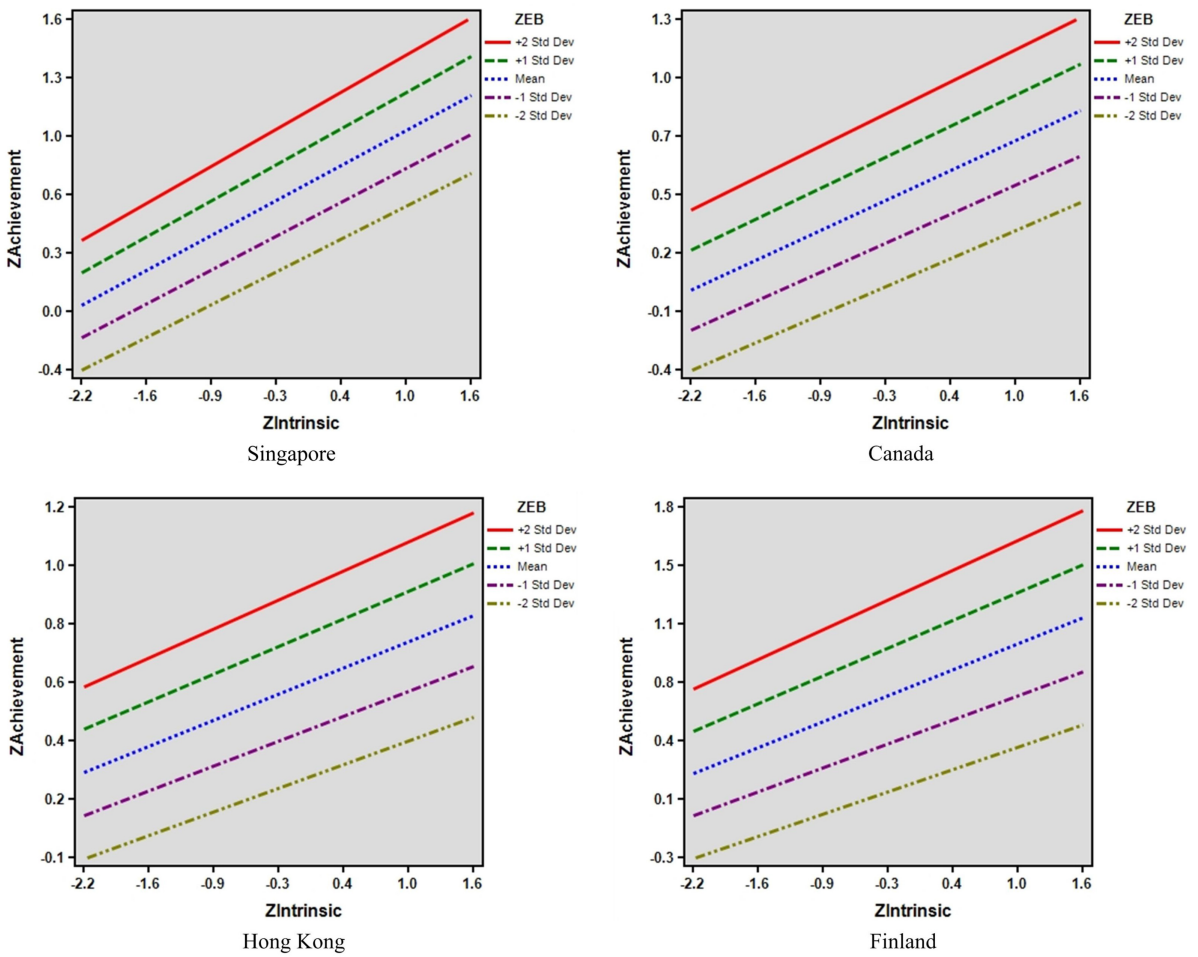

FIGURE 1 | Predicting science achievement: a graphical illustration of interaction of intrinsic motivation × epistemic beliefs.

Second, epistemic beliefs significantly predicted science achievement across all societies for both model 3 and model 4. The importance of facilitating development of sophisticated epistemic beliefs for science has received constant attention (Scott et al., 2006; Lin and Tsai, 2017). This study affirms the importance of epistemic beliefs for science achievement and science learning (Pluta et al., 2011; Belland et al., 2016; Greene et al., 2018) through the regression analyses among the four top-performing countries or regions. The implication to science education would be that epistemic beliefs about science need to be emphasized and explicitly discussed in class. The four samples we analyzed have a common emphasis on teaching science through inquiry with the aim of providing students with opportunities to be scientists rather than just science learners (Jocz et al., 2014; Cheung, 2018; Asghar et al., 2019; Inkinen et al., 2020). In particular, to develop more sophisticated epistemic beliefs, students need to be able to question knowledge claims and make justification from multiple references and sources (Belland et al., 2016).

Third, in model 4, intrinsic motivation and epistemic beliefs are both positive predictors of science achievement when both are entered into the regression equation (i.e., model 4), and this finding is commonly reflected across the four societies. This finding affirms previous research that has investigated the structural relationships among epistemic beliefs and motivation (Chen, 2012; Ho and Liang, 2015). In particular, Ho and Liang (2015) illustrate that sophisticated epistemic beliefs are predictive of deep intrinsic motivation to learn science mediated by constructive conceptions of learning science. This study extends the previous study with the science achievement as the predicted outcome to provide more support for science educators to structure intrinsically motivating science learning activities that concurrently challenges students to draw on sophisticated epistemic beliefs (Mason et al., 2013). Our finding also reveals that the interaction between intrinsic motivation and epistemic beliefs positively predicted science achievement in the Singapore dataset.

Instrumental motivation, on the other hand, showed a culturally specific pattern. In the Singapore context, instrumental motivation was a negative predictor of science achievement after taking into account the variance predicted by intrinsic motivation, whereas in the Western context, it is a positive predictor.

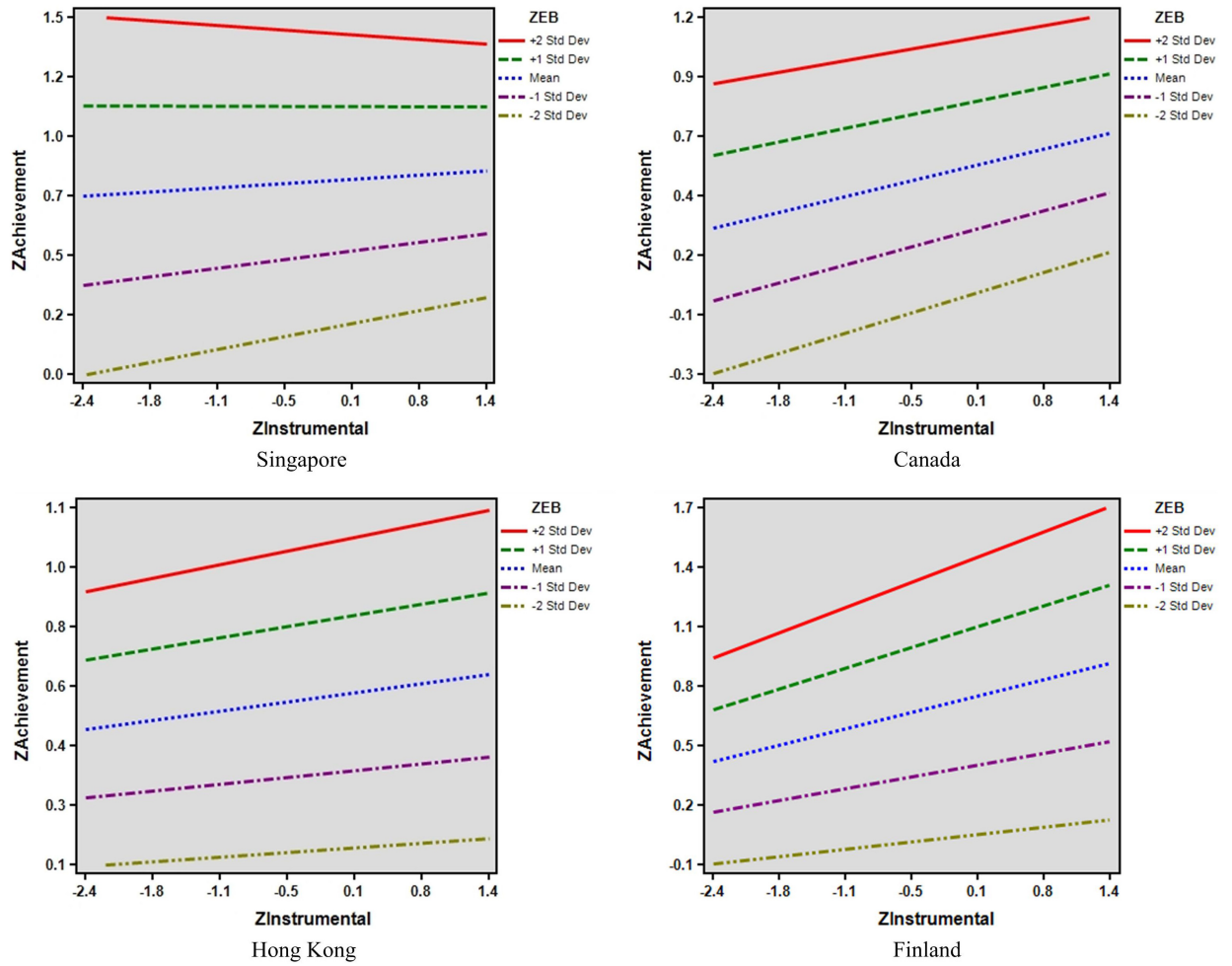

FIGURE 2 | Predicting science achievement: a graphical illustration of interaction of instrumental motivation × epistemic beliefs.

The interaction between instrumental motivation and epistemic beliefs showed culturally specific patterns. In Finland, the relationship between instrumental motivation and achievement was strongest for those with the most sophisticated epistemic beliefs. However, in Singapore and Canada, the relationship between instrumental motivation and achievement was strongest for students with less sophisticated epistemic beliefs. For those with more sophisticated epistemic beliefs, the relationship between instrumental motivation and achievement was weaker. These differential patterns might reflect differences in the educational system across countries though further research is needed to understand these patterns.

In this study, we also detected that gender influences achievement differently based on societies. Students' SES was an influential predictive of science achievement. Science educators may therefore need to pay specific attention to the gender issue while they design interesting and enjoyable science learning activities, depending on where they are located. Overall, there was no gender difference in Singaporean students' motivation and epistemic beliefs. In Hong Kong, male students had higher intrinsic motivation than female students in learning science. For Canadian students, male students had higher scores in motivation to learn science and epistemic beliefs than female students. On the other hand, Finnish female students had higher scores in motivation to learn science and epistemic beliefs than male students.

The results of the present study provide support for the complexity of factors that predict science achievement. We found that intrinsic motivation and epistemic beliefs are closely associated with science achievement, which may provide insights on the importance of intrinsic motivation or sophisticated epistemic beliefs. In line with our findings, instrumental motivation was found to be positively or negatively associated with science achievement, which needs to be appropriately researched. Overall, the findings support the importance of recognizing both cultural universals and about cultural/contextual differences (Yang, 2016).

LIMITATIONS

Some limitations should be noted. First, our findings showing the importance of intrinsic motivation and sophisticated epistemic

beliefs in facilitating science learning need to be replicated across different ages as PISA focuses on 15-year old students. Second, it might also be useful to test validity using confirmatory factor analysis and explore these relationships across different regions. There are 72 regions included in PISA and we decided to focus on only four regions especially because adding more regions would make our discussion unwieldy. Societies can differ on so many dimensions (e.g., government system, colonial background, GDP per capita, income inequality, ethnicity, demographic factors). However, future studies can examine the commonality of the results to other cultural contexts. Third, our study uses a cross-sectional correlational design and we cannot make causal conclusions. Future studies can utilize longitudinal or experimental designs to establish stronger causal conclusions. Fourth, because we relied on secondary data from PISA, the current study is also limited by PISA's sampling design and analytic framework.

CONCLUSION

Our study demonstrates that, to enhance science achievement, students need to be both intrinsically motivated and possess sophisticated epistemic beliefs. This pattern is common across the selected regions with notable differences in cultural contexts. However, instrumental motivation in the present study shows a regionally specific pattern. It seems that instrumental motivation was more adaptive in Western than Eastern societies. Our study suggests both commonality and specificity and indicates that increasing students' intrinsic motivation in science learning and helping them develop more sophisticated epistemic beliefs might be promising pathways to optimizing science achievement. This may also provide implications for science educators to motivate students' intrinsically to learn science and incorporate pedagogical strategies that will enhance more sophisticated and deeper epistemic processes and judgment.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.oecd.org/pisa/data/2015database/.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

REFERENCES

Aditomo, A., and Klieme, E. (2020). Forms of inquiry-based science instruction and their relations with learning outcomes: evidence from high and low-performing education systems. *Int. J. Sci. Educ.* 42, 504–525. doi: 10.1080/09500693.2020.1716093

Alexander, J. M., Johnson, K. E., and Kelley, K. (2012). Longitudinal analysis of the relations between opportunities to learn about science and the development of interests related to science. *Sci. Educ.* 96, 763–786. doi: 10.1002/sce.21018

Asghar, A., Huang, Y.-S., Elliott, K., and Skelling, Y. (2019). Exploring secondary students' alternative conceptions about engineering design technology. *Educ. Sci.* 9:45. doi: 10.3390/educsci9010045

Belland, B. R., Gu, J., Kim, N. J., and Turner, D. J. (2016). An ethnomethodological perspective on how middle school students addressed a water quality problem. *Educ. Technol. Res. Dev.* 64, 1135–1161. doi: 10.1007/s11423-016-9451-8

Brown, E. R., Steinberg, M., Lu, Y., and Diekman, A. B. (2018). Is the lone scientist an American dream? Perceived communal opportunities in stem offer a pathway to closing US–Asia gaps in interest and positivity. *Soc. Psychol. Pers. Sci.* 9, 11–23. doi: 10.1177/1948550617703173

Burns, E. C., Martin, A. J., and Collie, R. J. (2019). Examining the yields of growth feedback from science teachers and students' intrinsic valuing of science: implications for student- and school-level science achievement. *J. Res. Sci. Teach.* 56, 1060–1082. doi: 10.1002/tea.21546

Canning, E. A., Harackiewicz, J. M., Priniski, S. J., Hecht, C. A., Tibbetts, Y., and Hyde, J. S. (2018). Improving performance and retention in introductory biology with a utility-value intervention. *J. Educ. Psychol.* 110, 834–849. doi: 10.1037/edu0000244

Chen, J. A. (2012). Implicit theories, epistemic beliefs, and science motivation: a person-centered approach. *Learn. Indiv. Differ.* 22, 724–735. doi: 10.1016/j.lindif.2012.07.013

Chen, J. A., Metcalf, S. J., and Tutwiler, M. S. (2014). Motivation and beliefs about the nature of scientific knowledge within an immersive virtual ecosystems environment. *Contemp. Educ. Psychol.* 39, 112–123. doi: 10.1016/j.cedpsych.2014.02.004

Chen, J. A., and Pajares, F. (2010). Implicit theories of ability of Grade 6 science students: relation to epistemological beliefs and academic motivation and achievement in science. *Contemp. Educ. Psychol.* 35, 75–87.

Cheung, D. (2018). The key factors affecting students' individual interest in school science lessons. *Int. J. Sci. Educ.* 40, 1–23. doi: 10.1080/09500693.2017.1362711

Chiu, M. M., and Chow, B. W. Y. (2010). Culture, motivation, and reading achievement: high school students in 41 countries. *Learn. Indiv. Differ.* 20, 579–592. doi: 10.1016/j.lindif.2010.03.007

Chiu, M. M., Wing-Yin Chow, B., Mcbride, C., and Mol, S. T. (2016). Students' sense of belonging at school in 41 countries: cross-cultural variability. *J. Cross Cult. Psychol.* 47, 175–196. doi: 10.1177/0022022115617031

Curriculum Development Council (2017). *Science Education Key Learning Area Curriculum Guide (Primary 1-Secondary 6)*. Hong Kong.

Debacker, T. K., Crowson, H. M., Beesley, A. D., Thoma, S. J., and Hestevold, N. L. (2008). The challenge of measuring epistemic beliefs: an analysis of three self-report instruments. *J. Exp. Educ.* 76, 281–312. doi: 10.3200/jexe.76.3.281-314

Deci, E. L., and Ryan, R. M. (2000). The" what" and" why" of goal pursuits: human needs and the self-determination of behavior. *Psychol. Inq.* 11, 227–268.

Deci, E. L., Vallerand, R. J., Pelletier, L. G., and Ryan, R. M. (1991). Motivation and education: the self-determination perspective. *Educ. Psychol.* 26, 325–346. doi: 10.1080/00461520.1991.9653137

Ecklund, E. H., Scheitle, C. P., Peifer, J., and Bolger, D. (2017). Examining links between religion, evolution views, and climate change skepticism. *Environ. Behav.* 49, 985–1006. doi: 10.1177/0013916516674246

Elliott, E. S., and Dweck, C. S. (1988). Goals: an approach to motivation and achievement. *J. Pers. Soc. Psychol.* 54, 5–12. doi: 10.1037/0022-3514.54.1.5

Finnish National Agency for Education (2017). *National Core Curriculum for Basic Education for Adults 2017*. Helsinki: Finnish National Agency for Education.

Greene, J. A., Cartiff, B. M., and Duke, R. F. (2018). A meta-analytic review of the relationship between epistemic cognition and academic achievement. *J. Educ. Psychol.* 110, 1084–1111. doi: 10.1037/edu0000263

Hair, J. F. Jr., Black, W. C., Babin, B. J., Anderson, R. E., and Tatham, R. L. (2010). *SEM: An Introduction. Multivariate Data Analysis: A Global Perspective*, 7th Edn. Upper Saddle River, NJ: Pearson Education.

Hartman, R. O., Dieckmann, N. F., Sprenger, A. M., Stastny, B. J., and Demarree, K. G. (2017). Modeling attitudes toward science: development and validation of the credibility of science scale. *Basic Appl. Soc. Psychol.* 39, 358–371. doi: 10.1080/01973533.2017.1372284

Hasni, A., Roy, P., and Dumais, N. (2016). The teaching and learning of diffusion and osmosis: what can we learn from analysis of classroom practices? A case study. *EURASIA J. Math. Sci. Technol. Educ.* 12, 1507–1531. doi: 10.12973/eurasia.2016.1242a

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not WEIRD. *Nature* 466, 29–29. doi: 10.1038/466029a

Hidi, S., and Harackiewicz, J. M. (2000). Motivating the academically unmotivated: a critical issue for the 21st century. *Rev. Educ. Res.* 70, 151–179. doi: 10.3102/00346543070002151

Ho, H.-N. J., and Liang, J.-C. (2015). The relationships among scientific epistemic beliefs, conceptions of learning science, and motivation of learning science: a study of Taiwan high school students. *Int. J. Sci. Educ.* 37, 2688–2707. doi: 10.1080/09500693.2015.1100346

Hofer, B. K. (2002). "Personal epistemology as a psychological and educational construct: an introduction," in *Personal Epistemology: The Psychology of Beliefs About Knowledge and Knowing*, eds B. K. Hofer and P. R. Pintrich (Mahwah, NJ: Lawrence Erlbaum), 3–14.

Hofer, B. K. (2008). "Personal epistemology and culture," in *Knowing, Knowledge and Beliefs*, ed. M. S. Khine (Dordrecht: Springer), 3–22.

Hofer, B. K., and Pintrich, P. R. (1997). The development of epistemological theories: beliefs about knowledge and knowing and their relation to learning. *Rev. Educ. Res.* 67, 88–140. doi: 10.3102/00346543067001088

Hornsey, M., Harris, E. A., and Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: a 24-nation investigation. *Health Psychol.* 37, 307–315.

Hornsey, M. J., and Fielding, K. S. (2017). Attitude roots and jiu jitsu persuasion: understanding and overcoming the motivated rejection of science. *Am. Psychol.* 72, 459–473. doi: 10.1037/a0040437

Howard, J. L., Gagné, M., and Bureau, J. S. (2017). Testing a continuum structure of self-determined motivation: a meta-analysis. *Psychol. Bull.* 143, 1346–1377. doi: 10.1037/bul0000125

Hulleman, C. S., Schrager, S. M., Bodmann, S. M., and Harackiewicz, J. M. (2010). A meta-analytic review of achievement goal measures: different labels for the same constructs or different constructs with similar labels? *Psychol. Bull.* 136, 422–449. doi: 10.1037/a0018947

Inkinen, J., Klager, C., Juuti, K., Schneider, B., Salmela-Aro, K., Krajcik, J., et al. (2020). High school students' situational engagement associated with scientific practices in designed science learning situations. *Sci. Educ.* 104, 667–692. doi: 10.1002/sce.21570

Jocz, J. A., Zhai, J., and Tan, A. L. (2014). Inquiry learning in the Singaporean context: factors affecting student interest in school science. *Int. J. Sci. Educ.* 36, 2596–2618. doi: 10.1080/09500693.2014.908327

Jong, M. S. Y. (2017). Empowering students in the process of social inquiry learning through flipping the classroom. *Educ. Technol. Soc.* 20, 306–322.

Kaderavek, J. N., Paprzycki, P., Czerniak, C. M., Hapgood, S., Mentzer, G., Molitor, S., et al. (2020). Longitudinal impact of early childhood science instruction on 5th grade science achievement. *Int. J. Sci. Educ.* 42, 1124–1143. doi: 10.1080/09500693.2020.1749908

Kang, J., and Keinonen, T. (2018). The effect of student-centered approaches on students' interest and achievement in science: relevant topic-based, open and guided inquiry-based, and discussion-based approaches. *Res. Sci. Educ.* 48, 865–885. doi: 10.1007/s11165-016-9590-2

Kang, J., Keinonen, T., and Salonen, A. (2019). Role of interest and self-concept in predicting science aspirations: gender study. *Res. Sci. Educ.* 1–23. doi: 10.1007/s11165-019-09905-w

King, R. B., and McInerney, D. M. (2014). Culture's consequences on student motivation: capturing cross-cultural universality and variability through personal investment theory. *Educ. Psychol.* 49, 175–198. doi: 10.1080/00461520.2014.926813

King, R. B., and McInerney, D. M. (2019). Family-support goals drive engagement and achievement in a collectivist context: integrating etic and emic approaches in goal research. *Contemp. Educ. Psychol.* 58, 338–353. doi: 10.1016/j.cedpsych.2019.04.003

King, R. B., Mcinerney, D. M., and Pitliya, R. J. (2018). Envisioning a culturally imaginative educational psychology. *Educ. Psychol. Rev.* 30, 1031–1065. doi: 10.1007/s10648-018-9440-z

Kline, R. B. (2005). *Principles and Practices of Structural Equation Modeling*. New York, NY: The Guildford Press.

Krems, J. A., Varnum, M. E. W., and Van Lange, P. A. M. (2017). More than just climate: income inequality and sex ratio are better predictors of cross-cultural variations in aggression. *Behav. Brain Sci.* 40, 26–27.

Kriegbaum, K., Becker, N., and Spinath, B. (2018). The relative importance of intelligence and motivation as predictors of school achievement: a meta-analysis. *Educ. Res. Rev.* 25, 120–148. doi: 10.1016/j.edurev.2018.10.001

Krist, C. (2020). Examining how classroom communities developed practice-based epistemologies for science through analysis of longitudinal video data. *J. Educ. Psychol.* 112, 420–443. doi: 10.1037/edu0000417

Lavonen, J., and Juuti, K. (2016). "Science at finnish compulsory school," in *The Miracle of Education: The Principles and Practices of Teaching and Learning in Finnish Schools*, eds H. Niemi, A. Toom, and A. Kallioniemi (Rotterdam: Sense Publishers), 131–147.

Lay, Y. F., and Chandrasegaran, A. (2016). The predictive effects of motivation toward learning science on TIMSS grade 8 students' science achievement: a comparative study between Malaysia and Singapore. *EURASIA J. Math. Sci. Technol. Educ.* 12, 2949–2959. doi: 10.12973/eurasia.2016.02315a

Lee, M. H., Tsai, C. C., and Chai, C. S. (2012). A comparative study of Taiwan, Singapore, and China preservice teachers' epistemic beliefs. *Asia Pac. Educ. Res.* 21, 599–609.

Lee, S. W.-Y., Liang, J.-C., and Tsai, C.-C. (2016). Do sophisticated epistemic beliefs predict meaningful learning? Findings from a structural equation model of undergraduate biology learning. *Int. J. Sci. Educ.* 38, 2327–2345. doi: 10.1080/09500693.2016.1240384

Lehtinen, A., Lehesvuori, S., and Viiri, J. (2019). The connection between forms of guidance for inquiry-based learning and the communicative approaches applied—a case study in the context of pre-service teachers. *Res. Sci. Educ.* 49, 1547–1567. doi: 10.1007/s11165-017-9666-7

Li, J., and Yamamoto, Y. (2020). Western and east Asian sociocultural learning models: Evidence from cross-cultural and immigrant research.ăAsian J. Soc. Psychol. 23, 174–186. doi: 10.1111/ajsp.12384

Li, M., Zheng, C., Liang, J.-C., Zhang, Y., and Tsai, C.-C. (2018). Conceptions, self-regulation, and strategies of learning science among Chinese high school students. *Int. J. Sci. Math. Educ.* 16, 69–87. doi: 10.1007/s10763-016-9766-2

Liang, J. C., and Tsai, C. C. (2010). Relational analysis of college science-major students' epistemological beliefs toward science and conceptions of learning science. *Int. J. Sci. Educ.* 32, 2273–2289. doi: 10.1080/09500690903397796

Lin, P.-Y., and Schunn, C. D. (2016). The dimensions and impact of informal science learning experiences on middle schoolers' attitudes and abilities in science. *Int. J. Sci. Educ.* 38, 2551–2572. doi: 10.1080/09500693.2016.1251631

Lin, T.-J., Deng, F., Chai, C. S., and Tsai, C.-C. (2013). High school students' scientific epistemological beliefs, motivation in learning science, and their relationships: a comparative study within the Chinese culture. *Int. J. Educ. Dev.* 33, 37–47.

Lin, T.-J., and Tsai, C.-C. (2017). Developing instruments concerning scientific epistemic beliefs and goal orientations in learning science: a validation study. *Int. J. Sci. Educ.* 39, 2382–2401. doi: 10.1080/09500693.2017.1384593

Liu, Y., Hau, K. T., and Zheng, X. (2020). Does instrumental motivation help students with low intrinsic motivation? Comparison between Western and Confucian students. *Int. J. Psychol.* 55, 182–191. doi: 10.1002/ijop.12563

Mason, L., Boscolo, P., Tornatora, M. C., and Ronconi, L. (2013). Besides knowledge: a cross-sectional study on the relations between epistemic beliefs, achievement goals, self-beliefs, and achievement in science. *Instruct. Sci.* 41, 49–79. doi: 10.1007/s11251-012-9210-0

Milford, T. M., and Tippett, C. D. (2019). "Introduction: setting the scene for a meso-level analysis of Canadian science education," in *Science Education in Canada*, eds C. D. Tippett and T. M. Milford (Cham: Springer), 1–12.

Ministry of Education [MOE] (2013). *Science Syllabus Lower Secondary Express Course*. Singapore: Ministry of Education.

Ministry of Education [MOE] (2014). *Primary Science Syllabus*. Singapore: Ministry of Education.

Muis, K. R. (2007). The role of epistemic beliefs in self-regulated learning. *Educ. Psychol.* 42, 173–190. doi: 10.1080/00461520701416306

Nagengast, B., and Marsh, H. W. (2014). "Motivation and engagement in science around the globe: testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006," in *Handbook of International Large-Scale Assessment: Background, Technical Issues, and Methods of Data Analysis*, eds M. von Davier Rutkowski and D. Rutkowski (Boca Raton, FL: CRC Press), 317–344.

Nugent, G., Barker, B., Welch, G., Grandgenett, N., Wu, C., and Nelson, C. (2015). A model of factors contributing to STEM learning and career orientation. *Int. J. Sci. Educ.* 37, 1067–1088. doi: 10.1080/09500693.2015.1017863

OECD (2009). *PISA Data Analysis Manual: SPSS*, 2nd Edn. Paris: OECD Publishing.

OECD (2016a). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing.

OECD (2016b). *PISA 2015 Technical Report*. Paris: OECD Publishing.

Oishi, S. (2014). Socioecological psychology. *Annu. Rev. Psychol.* 65, 581–609. doi: 10.1146/annurev-psych-030413-152156

Paulsen, M. B., and Feldman, K. A. (2005). The conditional and interaction effects of epistemological beliefs on the self-regulated learning of college students: motivational strategies. *Res. High. Educ.* 46, 731–768.

Pettigrew, T. F., and Hewstone, M. (2017). The single factor fallacy: implications of missing critical variables from an analysis of intergroup contact theory 1. *Soc. Issues Policy Rev.* 11, 8–37. doi: 10.1111/sipr.12026

Pintrich, P. R. (2003). A motivational science perspective on the role of student motivation in learning and teaching contexts. *J. Educ. Psychol.* 95, 667–686. doi: 10.1037/0022-0663.95.4.667

Pluta, W. J., Chinn, C. A., and Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *J. Res. Sci. Teach.* 48, 486–511. doi: 10.1002/tea.20415

Potvin, P., and Hasni, A. (2014). Analysis of the decline in interest towards school science and technology from grades 5 through 11. *J. Sci. Educ. Technol.* 23, 784–802. doi: 10.1007/s10956-014-9512-x

Rizeq, J., Flora, D. B., and Toplak, M. E. (2020). An examination of the underlying dimensional structure of three domains of contaminated mindware: paranormal beliefs, conspiracy beliefs, and anti-science attitudes. *Think. Reason.* 1–25. doi: 10.1080/13546783.2020.1759688

Rozek, C. S., Hyde, J. S., Svoboda, R. C., Hulleman, C. S., and Harackiewicz, J. M. (2015). Gender differences in the effects of a utility-value intervention to help parents motivate adolescents in mathematics and science. *J. Educ. Psychol.* 107, 195–206. doi: 10.1037/a0036981

Ryan, R. M., and Deci, E. L. (2009). "Promoting self-determined school engagement," in *Handbook of Motivation at School*, eds K. Wentzel, A. Wigfield, and D. Miele (New York, NY: Routledge), 171–195.

Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Broda, M., Spicer, J., et al. (2016). Investigating optimal learning moments in U.S. and finnish science classes. *J. Res. Sci. Teach.* 53, 400–421. doi: 10.1002/tea.21306

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *J. Educ. Psychol.* 82, 498–504. doi: 10.1037/0022-0663.82.3.498

Schommer-Aikins, M., and Easter, M. (2008). Epistemological beliefs' contributions to study strategies of Asian Americans and European Americans. *J. Educ. Psychol.* 100, 920–929. doi: 10.1037/0022-0663.100.4.920

Scott, P. H., Mortimer, E. F., and Aguiar, O. G. (2006). The tension between authoritative and dialogic discourse: a fundamental characteristic of meaning making interactions in high school science lessons. *Sci. Educ.* 90, 605–631. doi: 10.1002/sce.20131

Shen, K. M., Li, T. L., and Lee, M. H. (2018). Learning biology as 'Increase ones' knowledge and understanding': studying Taiwanese high school students' learning strategies in relation to their epistemic views and conceptions of learning in biology. *Int. J. Sci. Educ.* 40, 2137–2157.

Sinatra, G. M., and Lombardi, D. (2020). Evaluating sources of scientific evidence and claims in the post-truth era may require reappraising plausibility judgments. *Educ. Psychol.* 55, 120–131.

So, W. M. W., Wan, Z. H., and Chen, Y. (2018). "Primary science education in Hong Kong," in *Primary Science Education in East Asia*, eds Y. J. Lee and J. Tan (Cham: Springer), 19–48.

Sun, D., Looi, C.-K., Wu, L., and Xie, W. (2016). The innovative immersion of mobile learning into a science curriculum in singapore: an exploratory study. *Res. Sci. Educ.* 46, 547–573. doi: 10.1007/s11165-015-9471-0

Tang, X., and Zhang, D. (2020). How informal science learning experience influences students' science performance: a cross-cultural study based on PISA 2015. *Int. J. Sci. Educ.* 42, 598–616. doi: 10.1080/09500693.2020.1719290

The Council of Ministers of Education, Canada (2016). *Common Framework of Science Learning Outcomes*. Available online at: http://science.cmec.ca/framework/Pages/english/1.html (accessed on 07 July 2020)

Tsai, C. C., Ho, H. N. J., Liang, J. C., and Lin, H. M. (2011). Scientific epistemic beliefs, conceptions of learning science and self-efficacy of learning science among high school students. *Learn. Instruct.* 21, 757–769.

Vedder-Weiss, D., and Fortus, D. (2011). Adolescents' declining motivation to learn science: inevitable or not? *J. Res. Sci. Teach.* 48, 199–216. doi: 10.1002/tea.20398

Wan, Z. H., and Lee, J. C. K. (2017). Hong Kong secondary school students' attitudes towards science: a study of structural models and gender differences. *Int. J. Sci. Educ.* 39, 507–527. doi: 10.1080/09500693.2017.1292015

Wilson, J. A. (2018). Reducing pseudoscientific and paranormal beliefs in University Students through a course in science and critical thinking. *Sci. Educ.* 27, 183–210. doi: 10.1007/s11191-018-9956-0

Wong, S. Y., Liang, J.-C., and Tsai, C.-C. (2019). Uncovering Malaysian secondary school students' academic hardiness in science, conceptions of learning science, and science learning self-efficacy: a structural equation modelling analysis. *Res. Sci. Educ.* 1–28. doi: 10.1007/s11165-019-09908-7

Yang, F. Y. (2016). "Learners' epistemic beliefs and their relations with science learning—exploring the cultural differences," in *Science Education Research and Practices in Taiwan*, ed. M. H. Chiu (Singapore: Springer), 133–146.

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci.* 12, 1100–1122. doi: 10.1177/1745691617693393

Zusho, A., and Clayton, K. (2011). Culturalizing achievement goal theory and research. *Educ. Psychol.* 46, 239–260. doi: 10.1080/00461520.2011.614526

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

APPENDIX

TABLE A1 | The three latent variables and their assessment items in PISA 2015.

1. Intrinsic motivation	
ST094Q01NA	I have fun when I am learning <broad science>
ST094Q02NA	I like reading about <broad science> topics.
ST094Q03NA	I am happy working on <broad science> topics.
ST094Q04NA	I enjoy acquiring new knowledge in <broad science>.
ST094Q05NA	I am interested in learning about <broad science>.
2. Instrumental motivation	
ST113Q01TA	Making an effort in my <school science> subject(s) is worth it because this will help me in the work I want to do later on
ST113Q02TA	What I learn in my <school science> subject(s) is important for me because I need this for what I want to do later on
ST113Q03TA	Studying my <school science> subject(s) is worthwhile for me because what I learn will improve my career prospects.
ST113Q04TA	Many things I learn in my <school science> subject(s) will help me to get a job.
3. Epistemic beliefs	
ST131Q01NA	A good way to know if something is true is to do an experiment.
ST131Q03NA	How much do you disagree or agree with the statements below? Ideas in <broad science> sometimes change.
ST131Q04NA	Good answers are based on evidence from many different experiments.
ST131Q06NA	It is good to try experiments more than once to make sure of your findings.
ST131Q08NA	Sometimes <broad science> scientists change their minds about what is true
ST131Q11NA	The ideas in <broad science> science books sometimes change.

frontiers in Psychology

ORIGINAL RESEARCH
published: 18 October 2021
doi: 10.3389/fpsyg.2021.580763

The Relation Between Television Viewing Time and Reading Achievement in Elementary School Children: A Test of Substitution and Inhibition Hypotheses

Wilfried Supper[1], Frédéric Guay[1] and Denis Talbot[2]*

[1] Département des fondements et pratiques en éducation, Faculté des sciences de l'éducation, Université Laval, Quebec, QC, Canada, [2] Département de médecine sociale et préventive, Faculté de Médecine, Université Laval, Quebec, QC, Canada

OPEN ACCESS

Edited by:
Ronnel B. King,
University of Macau, China

Reviewed by:
Kimmo Eriksson,
Mälardalen University College, Sweden
Elmer Dela Rosa,
Central Luzon State University, Philippines

****Correspondence:***
Frédéric Guay
Frederic.Guay@fse.ulaval.ca

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *07 July 2020*
Accepted: *16 September 2021*
Published: *18 October 2021*

Citation:
Supper W, Guay F and Talbot D (2021) The Relation Between Television Viewing Time and Reading Achievement in Elementary School Children: A Test of Substitution and Inhibition Hypotheses. Front. Psychol. 12:580763. doi: 10.3389/fpsyg.2021.580763

Research has focused on the relations between television (TV) viewing time and children's reading achievement. Two hypotheses have been proposed to explain this relation. The substitution hypothesis proposes that TV viewing distracts students from activities that are important for their learning. The inhibition hypothesis proposes that watching television inhibits important affective/cognitive skills. In this study, we test both hypotheses by estimating the relation between TV viewing time and reading achievement. We use the frequency of students' leisure reading and the frequency of interactions between students and their parents as potential mediators to test the substitution hypothesis, whereas for the inhibition one, we use students' intrinsic motivation to read and their level of inattention. Data come from the Québec Longitudinal Study of Child Development (QLSCD). Designed by the *Institut de la statistique du Québec*, QLSCD covers a wide range of themes. The QLSCD is representative of children in Québec and contains 2223 participants who were followed from 0 to 21 years old. The four structural models tested are built as follows: the TV viewing time at 6 years old predicts the four mediating variables at 8 years old, which in turn predicts reading achievement at 10 years old. In addition, we have tested models' gender invariance. Results indicate that TV viewing time is not directly or indirectly associated with reading achievement. Specifically, it is not associated with the mediating variables of child-parent interactions, intrinsic motivation, and inattention. However, the frequency of leisure reading is negatively associated with the time spent watching TV. This association is very small (−0.07) and has no indirect effect on reading achievement. Finally, results do not vary according to the gender of the participants. Our results are in line with those of previous studies in the field and cast some doubts on the potential negative effects of TV viewing time on reading achievement.

Keywords: television viewing time, reading achievement, reading for leisure, intrinsic motivation, inattention

INTRODUCTION

Television (TV) viewing time has been widely criticized for its negative influence on children's learning to read (Winn, 1977; Postman, 1986; Ennemoser and Schneider, 2007). Major concerns are that time spent on watching TV replaces reading activities, reduces children's interest for reading, lowers language skills, makes children intellectually lazy, inattentive, and inhibits their imagination (Himmelweit et al., 1958; Winn, 1977; Hornik, 1981; Postman, 1986; Popper et al., 1995; Ennemoser and Schneider, 2007; Desmurget, 2011). In addition, the important portion of time children devote to this leisure activity might be a cause of concern. Indeed, several studies reveal that, on average, children spend as much time watching TV, as they do performing classroom tasks (Paik, 2000; Desmurget, 2011). This makes TV watching one of the most frequent hobbies for a majority of children (Rideout, 2016). It is therefore important to identify the extent to which the time spent watching TV affects children's reading achievement (RA). For more than 60 years, research has focused on this relation. However, very few studies have investigated the processes that explain why TV viewing has been associated with lower RA. Among studies testing mediators, results are divergent (Kostyrka-Allchorne et al., 2017; Supper et al., in-press article). Researchers have proposed several hypotheses to explain this divergence, including content type (Wright et al., 2001; Ennemoser and Schneider, 2007), cultural differences (Ennemoser and Schneider, 2007), differences in methodological quality between studies (Zavodny, 2006; Gentzkow and Shapiro, 2008; Munasib and Bhattacharya, 2010), the existence of moderators (Paik, 2000; Razel, 2001), and a non-linear relationship (Paik, 2000; Razel, 2001; Huang and Lee, 2010). However, studies that have focused on these potential confounding effects also show conflicting results (for research that specifically addresses this topic, see: Kostyrka-Allchorne et al., 2017; Supper et al., 2021). For example, cultural differences or content differences are aspects that we feel are secondary to consider in order to better understand this divergence in results. Indeed, most hypotheses on TV viewing time and RA are based on the video format of the TV and not on the type of content or messages likely to be broadcasted. Simply watching more TV is expected to predict a decrease in the amount of time children spend on educational activities such as reading, regardless of the type of program shown or the cultural practices between children in different countries. Based on past research, it is therefore difficult to draw clear conclusions about how TV viewing time is associated with RA.

Hypotheses Explaining the Negative Effects of Television Watching Time on Reading Achievement

Researchers propose several processes to explain why TV watching time could be negatively associated with RA (Himmelweit et al., 1958; Schramm, 1961; Hornik, 1981; Beentjes and Van der Voort, 1988). These processes can be classified into the substitution and the inhibition hypotheses.

The substitution hypothesis posits that the time children spend watching TV replaces the time they spend on other activities more susceptible to enhance their RA, such as reading, writing, and doing schoolwork (Hornik, 1981). This hypothesis therefore assumes that the different activities that make up children's schedules reflect a zero-sum game (Peaucelle, 1969). That is, an increase in the time spent on one type of activity, for example watching TV, inevitably decreases the time devoted to other types of activities favorable to RA. Among these educative activities, the frequencies of leisure reading and of non-educational interactions between children and their parents seem to be relevant mediating variables for testing the substitution hypothesis.

Leisure reading helps children to provide cognitive efforts when needed, to mobilize reading strategies learned in school, to discipline themselves, and to solve problems (Bergin, 1992). Thus, a decrease in the time that children spend on leisure reading will result in less time spent on cognitive, behavioral, and affective processes that enhance their RA. For this reason, we have chosen leisure reading as a mediating variable to test the substitution hypothesis. Some studies conclude that the time spent watching TV is associated with a decrease in leisure reading time (Koolstra et al., 1996; Shin, 2004; Ennemoser and Schneider, 2007), while other studies indicate that there is no association between these two variables (Himmelweit et al., 1958; Schramm, 1961; Ritchie et al., 1987; Anderson et al., 2001; Wright et al., 2001).

Furthermore, children who watch more TV could deprive themselves from the positive effects of spending time with their parents, such as better language development (Zimmerman et al., 2009) or support for their academic success (Guay et al., 2007). Indeed, TV watching time is associated with fewer parent-child interactions (Zimmerman et al., 2007). Television viewing time is also associated with a decrease in time spent on homeworks (Johnson et al., 2007), which are usually supervised by parents. However, some studies do not support these negative relationships (Schramm, 1961; Nakamuro et al., 2013).

The inhibition hypothesis posits that watching TV negatively affects RA by altering certain cognitive (mainly attention and concentration) or affective components (mainly valuing effort and interest in reading and school learning; Himmelweit et al., 1958; Beentjes and Van der Voort, 1988). The rationale underlying this hypothesis is that TV broadcasts content that does not require a sustained effort of understanding or concentration (Hornik, 1981). In addition, the images' speed, the quality of the visual and sound effects and the abundant supply of channels make this content entertaining and very stimulating, which gives children pleasure from the very first minutes (Winn, 1977; Postman, 1986). Thus, immediate and easy access to an entertaining activity could lead children to develop a certain "mental laziness" (Himmelweit et al., 1958; Beentjes and Van der Voort, 1988) which would discourage them from mobilizing efforts or concentration to be interested in reading activities. Intrinsic motivation (IM) to read, that is reading for the sole interest or pleasure it provides (Ryan and Deci, 2000), could therefore be negatively affected by the number of hours children spend watching TV. It is well known that a decrease

in IM to read is associated with a decrease in cognitive and emotional commitment to reading (Guthrie et al., 2012), in reading comprehension (Malanchini et al., 2017), in perceived reading skills and in RA (Morgan and Fuchs, 2007). For this reason, we have chosen IM to read as a mediating variable to test the inhibition hypothesis. In line with this, one study suggests that TV viewing is negatively associated with favorable attitudes toward reading (a concept similar to IM; Koolstra et al., 1996) or with reading motivation (Anderson et al., 2001), while another concludes that these variables are not significantly linked (Ennemoser and Schneider, 2007).

In addition, when asked to perform less interesting educational activities, children who watch a lot of TV may find it difficult to sustain their attention and, as a result, become more easily distracted (Beentjes and Van der Voort, 1988). An increase in inattention is associated with a decrease in RA (Rabiner et al., 2000). For this reason, we selected "attention" as an additional mediating variable to test the inhibition hypothesis. In this regard, the majority of studies indicate that watching TV is associated with an increase in inattention (Christakis et al., 2004; Mehmet-Radji, 2004; Acevedo-Polakovich et al., 2006; Johnson et al., 2007; Miller et al., 2007; Maass et al., 2010; Swing et al., 2010; Schmiedeler et al., 2014). However, an important number of studies have also found that there is no association between these variables (see **Table 1**).

In conclusion, the four mediating variables we chose to test the substitution and inhibition hypothesis are leisure reading, parent-child interaction, IM to read, and inattention. These mediators were chosen because they seem to be the ones that best fit the propositions of these two hypotheses (Hornik, 1981; Beentjes and Van der Voort, 1988; Koolstra et al., 1996; Ennemoser and Schneider, 2007).

Children's Age

The vast majority of children start watching TV long before they reach the age when they learn to read. Therefore, by the time these children begin to learn to read at school, they have developed the habit of watching TV frequently and then have potentially developed cognitive and affective components that could hamper their willingness to engage in more demanding activities, such as reading (Hornik, 1981; Beentjes and Van der Voort, 1988; Ennemoser and Schneider, 2007). Thus, it is relevant to test the inhibition and substitution hypotheses early in children's development.

Children's Gender as a Potential Moderator

Testing gender as a potential moderator is important for several reasons. First, boys watch TV for longer periods than girls (Sisson et al., 2012). They are exposed to more violent content and shows that portray stereotypical representations of masculinity (Coyne et al., 2014). These TV shows are thus more likely to provide the viewer with greater visual stimulation, and may thereby make boys more susceptible to inhibition effects. Second, the majority of children have female reader models (Morin, 2014) and boys more frequently report that reading is a female activity (Clark, 2012). This stereotypical view of reading could lead boys to value this activity less and thus to be less interested in reading (Morin, 2014). Third, boys read less frequently for leisure (Morin, 2014) and their book choices are more circumscribed around comic books (Morin, 2014). Boys therefore read more pictorial books, and choose more frequently reading formats that are close to TV content. For this reason, the type of reading that boys prefer may be more easily replaced by TV content. These gender differences may thus moderate the relationship between TV viewing time and RA. In light of the above, we expected the following: for boys, the relationships connecting TV viewing to the mediators and RA should be stronger than those observed for girls. In other words, boys would be more prone to the inhibition and substitution effects.

Contributions of This Study

This study contributes to the existing literature in several ways. First, no study has tested the moderating role of gender in the "TV→mediators→RA" sequence. Second, studies that have tested the relationship between TV viewing time and some mediating variables show contrasting results. It is therefore difficult to determine whether watching TV is associated with less leisure reading or increased inattention. Yet, according to studies testing the relationship between TV viewing and RA (Hornik, 1981; Beentjes and Van der Voort, 1988), these two variables are among those that allow the inhibition and substitution hypotheses to be tested in the most stringent way. Third, few studies have tested substitution and inhibition hypotheses simultaneously, aside from the one conducted by Ennemoser and Schneider (2007). However, this study has important limitations. On the one hand, the results presented are unclear and they are divided between those indicating that TV viewing time is negatively associated with RA and those revealing that there is no association between these variables. On the other hand, this study contains a very limited number of participants. It is therefore important to test the substitution and inhibition hypothesis with a larger representative sample. Indeed, it was important to conduct such research, both socially and scientifically. It allows for a more precise estimation of how TV viewing time is related to RA and, therefore, for more appropriate recommendations and interventions. For example, if TV viewing time decreased leisure time spent reading, but did not inhibit cognitive abilities, parents could ensure that viewing time does not interfere with their children's reading time. Conversely, if TV viewing time were not related to reading time, but was related to language development, then it would be important to recommend a more systematic reduction in TV viewing time, especially for younger children. Testing these two hypotheses together, therefore, provides a better understanding of the relationship between TV viewing time and RA. Fourth, longitudinal studies testing IM to read and the frequency of parent-to-child interactions as mediating mechanisms in the "TV→RA" relationship are scarce.

Goals and Hypotheses

The goal of this study is to better understand the processes likely to mediate the relationship between TV viewing time and children's RA. In order to test the inhibition and substitution

TABLE 1 | Studies on the association between TV viewing time and mediating variables.

Authors	Year	Mediator	Results
Tomopoulos et al.	2010	Cognitive skills	Negative association
Lonner et al.	1985	Cognitive skills	NS
Acevedo-Polakovich et al.	2006	Attention	Negative association
Cheng et al.	2010	Attention	Negative association
Christakis et al.	2004	Attention	Negative association
Maass et al.	2010	Attention	Negative association
Mehmet-Radji	2004	Attention	Negative association
Miller et al.	2007	Attention	Negative association
Schmiedeler et al.	2014	Attention	Negative association
Swing et al.	2010	Attention	Negative association
Johnson. Cohen et al.	2007	Attention	Negative association
Ansari and Crosnoe	2016	Attention	NS
Foster and Watkins	2010	Attention	NS
Landhuis et al.	2007	Attention	NS
Parkes et al.	2013	Attention	NS
Stevens et al.	2009	Attention	NS
Stevens et Mulsow	2006	Attention	NS
Zimmerman and Christakis	2007	Attention	NS
Koolstra et al.	1996	Attitude toward reading/Leisure reading	Negative association
Ennemoser and Schneider	2007	Attitude toward reading	NS
Johnson et al.	2007	Negative attitude toward school	Negative association
Huang & Lee	2010	Negative behavior at school	Negative association
Nakamuro et al.	2013	Negative behavior at school	NS
Parkes et al.	2013	Negative behavior at school	NS
Zimmerman et al.	2007	Language development	Negative association
Tomopoulos et al.	2010	Language development	Negative association
Barr et al.	2010	Language development	Negative association
Blankson et al.	2015	Language development	NS
Duch et al.	2013	Language development	Negative association
Schmidt et al.	2009	Language development	NS
Ruangdaraganon et al.	2009	Language development	Negative association
Pagani et al.	2013	Language development	Negative association
Linebarger et al.,	2005	Language development	Negative association
Bittman et al.	2011	Language development	NS
Johnson et al.	2007	Homework	Negative association
Nakamuro et al.	2013	Homework	NS
Koolstra et al.	1997	Mental effort	Negative association
Barr et al.	2010	Executive functions	Negative association
Hamer et al.	2010	Cognitive functions	Negative association
Ennemoser and Schneider	2007	Reading	Negative association
Anderson et al.	2001	Motivation	NS
Sharif et al.	2010	Search for sensation	Negative association
Blankson et al.	2015	Numeracy skills test	NS
Schmidt et al.	2009	Visual motor skills	NS

NS = Not statistically significant at the 5% level.

hypotheses, we will explore if the time that 6-year-old children spend watching TV predicts the frequency of their leisure reading, the frequency of their interaction with their parents, their IM to read and the level of inattention at the age of 8. Additionally, we will test if, in turn, these four variables predict their RA at the age of 10. If the substitution hypothesis is supported, the TV viewing time at 6 years old will be negatively associated with the frequency of their leisure reading and/or the frequency of their interaction with the parent at 8 years old. If the inhibition hypothesis is corroborated, then TV viewing time at 6 years of age will be negatively associated with IM to read at 8 years of age and/or will be positively associated with inattention at 8 years of age. Finally, we posit that boys are more likely than girls to be affected by substitution and

inhibition effects: at equal TV viewing time, we expect the substitution and inhibition effects will be more marked for boys than for girls.

MATERIALS AND METHODS

Participants

The data came from the Québec Longitudinal Study of Child Development (QLSCD). The QLSCD targeted the population of children who were born in Québec between 1997 and 1998. However, children living on Indigenous reserves, regions of Nord-du-Québec, Cree territory and Inuit territory as well as children born prematurely (gestation less than 24 weeks) were excluded (Jetté and Des Groseilliers, 2000).

Québec longitudinal study of child development contained data concerning 2223 children aged between 5 and 6 months at the time of recruitment. This sample was made up of 48.8% of females. Since young children could not have answered the various measures themselves, the QLSCD has asked parents or legal guardians to complete the measures. Mother is the «Person Most Knowledgeable» (PMK) of children's behaviors in 98.4% of cases. To obtain a better consistency, the QLSCD has encouraged the PMK to be the same respondent over time. In terms of education level, 44% of PMKs and 28% of fathers held a high school diploma or did not have a diploma, 29% of the PMK and 41% of fathers had a non-university post-secondary diploma and 26% of PMK and 30% of fathers had a university degree. Furthermore, 86.5% of mothers and 84.5% of fathers were Québec natives. Finally, French was the mother tongue for most of the participants' mothers (81%), followed by English (9%) and other languages (10%).

The variables used to test our hypotheses were measured at the ages of 5 months, and 6, 7, 8, and 10 years old. In **Table 2**, we have indicated the measurement times during at which each variable was measured.

Measures

The Average Daily Television Viewing Time

This variable was assessed by the PMK, when the child was 6 years old. The following questions were asked: "How much time does your child spend watching TV during the week?" and "How much time does your child spend watching TV during the weekend?" These two questions came from the Canadian Community Health Survey. This measure was similar to measures of TV viewing time used in other surveys (e.g., the National Longitudinal Survey of Children). The average TV viewing time per day for 6 years-old in the QLSCD was 1 h and 50 min. This amount is comparable to the average in other surveys (Zimmerman and Christakis, 2005; Rideout, 2016).

Academic Reading Achievement

This variable was assessed by teachers. In this study, we have selected scores when the children were 7 years old and when they were 10 years old. Scores on RA were strongly correlated with other measures of academic performance (see Forget-Dubois et al., 2007). The teacher has reported children's RA by answering the following question: "How would you assess the child's current academic success in reading?" The answers were given on a scale from 1 to 5: (1 = among the first in the class; 5 = among the last in the class). Students' scores were assessed at the end of the school year.

The Frequency of Leisure Reading

This measure came from the PMK's response to the question: "How often does your child enjoy reading?" This variable has been assessed when the child was 6 years old and when the child was 8 years old. Answers were rated on a scale from 1 to 7 (1 = rarely or never; 4 = a few times a month; 7 = every day). Previous studies showed that this measure of reading time for leisure was associated with children's RA at 8 years old (Tétreault and Desrosiers, 2014; Nanhou et al., 2016; Torppa et al., 2020; Manu et al., 2021).

Frequency of Non-educational Interactions Between the Child and the Parent

This scale was filled by the PMK. For our study, we used the responses that were provided when the child was 6 years old and when the child was 8 years old. This scale was adapted from a subscale of the Parenting Practices Scale by Strayhorn and Weidman (1988; Boivin et al., 2000) which aimed to measure the frequency of supportive and encouraging behaviors from parent to child. This scale presented an acceptable level of validity and fidelity (Strayhorn and Weidman, 1988; Boivin et al., 2000; Verhoeven et al., 2010). It includes 10 items when the child was 6 years old and 5 items when the child was 8 years old. To measure the frequency of non-educational interactions between the child and the parent, we only kept three items per measurement time: "How often do you talk or play with the child?"; "How often do you do a special activity that she/he enjoys?"; "How often do you get involved in sports, hobbies or games?". Answers were given on a scale from 1 to 5 (1 = never; 5 = several times a day). Other items on this scale were excluded for two reasons. First, the removed items mainly measured the quality of parenting practices rather than the frequency of parent-child interactions (i.e.,: "In the past 12 months, when you talked to the child about behavior, in what proportion of the time did you congratulate?"). However, our substitution hypothesis targeted the frequency of interactions and not their quality. The exclusion of items that did not measure the frequency of interactions therefore allowed us to be more consistent with our hypothesis. Second, the three items that were selected are identical between the 2 measurement times. The internal consistency of our three items was 0.61 when the child was 6 years old and 0.63 when the child was 8 years old (Cronbach's alpha).

Intrinsic Motivation for School Reading

This measure came from a scale that was filled by the child at 7 and 8 years old. Items were from a subscale of the "Elementary School Motivation Scale" (Guay et al., 2010) whose aim was to measure different forms of school motivation in reading, writing and mathematics for a population of elementary school children. The content validity, construct validity, and internal consistency of the scale has been supported (Guay et al., 2010). The IM for

TABLE 2 | Descriptive statistics.

	Girls		Boys		Total population			
	Mean	SD	Mean	SD	Mean	SD	Attrition	Age (years)
Parental practices	4.6	0.8	4.6	0.8	4.6	0.8	33.0%	6
Parental practices	4.3	0.8	4.4	0.9	4.4	0.9	34.8%	8
Inattention	1.3	0.4	1.5	0.5	1.4	0.5	47.6%	7
Inattention	1.3	0.4	1.5	0.5	1.4	0.5	49.7%	8
Talk about school activities a	6.9	0.4	6.9	0.5	6.9	0.5	31.5%	8
IQ	80.5	17.4	80.2	16.9	80.4	17.2	47.6%	6
Parental valorization of grades	3.5	0.6	3.5	0.6	3.5	0.6	31.7%	8
Leisure reading	5.3	2.0	4.5	2.2	4.9	2.2	42.7%	7
Leisure reading	4.3	1.0	3.8	1.3	4.1	1.1	43.2%	9
Motivation to read	4.3	0.8	4.0	1.0	4.1	0.9	34.4%	8
Motivation to read	4.3	0.8	4.0	1.0	4.1	0.9	34.4%	9
Reading score	3.7	1.3	3.4	1.4	3.5	1.3	42.1%	7
Reading score	3.6	1.2	3.2	1.3	3.4	1.3	57.2%	11
PMK diploma	2.7	1.1	2.7	1.0	2.7	1.1	0.1%	0.4
TV viewing time	1.8	0.9	1.9	0.8	1.8	0.8	33.0%	6
Gender	1	0	0	0	0.5	0.0	0.0%	0.4

Motivation, parental practices and inattention variables are items means corresponding to these constructs.

school reading was made up of 3 items ("I like reading"; "Reading interests me a lot"; "I read even when I don't have to") for which the answers were given on a scale from 1 to 4 (1 = always no, sometimes no, sometimes yes, 4 = always yes). In QLSCD, Cronbach's alpha was 0.68 when the child was 7 years old and 0.68 at 8 years old.

Children's Symptoms of Inattention

This measure was filled out by the teacher when the child was 7 years old and when the child was 8 years old. The items on this scale came from the Ontario Child Health Study (OCHS; Cardin et al., 2011). This scale had a good level of validity (Boyle et al., 1993; Romano et al., 2006; Cardin et al., 2011) and has been used in various studies for its ability to predict school achievement (Pingault et al., 2013). This scale was composed of 5 items ("was unable to concentrate"; "could not maintain her/his attention for a long time"; "was easily distracted"; "had difficulty pursuing any activity"; "was inattentive"). Answers were given on a scale from 1 to 3 (1 = never or not true; sometimes/a little; 3 = often or very). For two measurement times, when the child was 7 and 8 years old, the Cronbach's alpha was 0.88.

Control Variables

Several authors have highlighted the important influence that confounding variables have on the relationship between TV viewing time and RA (Ennemoser and Schneider, 2007; Munasib and Bhattacharya, 2010). More specifically, the intelligence quotient (IQ), the level of education of the parents and the parental involvement determine both the time spent watching TV and the RA (Ennemoser and Schneider, 2007). Consequently, we controlled in our models the following four variables: parents' education level, child's IQ, parents' interest in their child's education and parents' valorization of good grades (Ennemoser and Schneider, 2007; Munasib and Bhattacharya, 2010).

Some authors also suggest that the time spent watching TV and the RA potentially have reciprocal relationships (Munasib and Bhattacharya, 2010). In order to minimize the influence of these biases, we controlled our mediating and dependent variables by taking into account initial scores on these variables.

In this study, we used four covariates to test our hypotheses. First, the highest level of education that the PMK has achieved was measured by a Likert scale from 1 to 4 (1 = no diploma; 4 = university degree). Second, the Peabody Picture Vocabulary Test was administered to the children when they were 6 years old (Dunn et al., 1993). This IQ test was strongly correlated with other measures of intelligence (Childers et al., 1994) and it was used in several studies that focus on RA (Salla et al., 2016). Scores could range between 1 and 120. Third, the frequency of PMK talking to the child about school activities was assessed by the following question: "How often do you talk to your child about school activities or work?" Responses were given on a scale from 1 to 4 (1 = daily; 4 = rarely). This variable was measured when the child was 7 years old. Fourth, the value of academic performance was the PMK's response to the question: "How important is it to you that your child has good grades in school?" This variable was associated with the child's RA (Tétreault and Desrosiers, 2014). The answer was given on a scale from 1 to 4 (1 = very important; 4 = not important) and it was measured when the child was 7 years old.

Statistical Analysis

Missing Data

The QLSCD contained a significant number of missing data as shown in **Table 2**. We treated these missing data with the "full information maximum likelihood" (FIML) procedure of Mplus (Muthén and Muthén, 2012).

TABLE 3 | Fit indices for the 4 models with and without invariance test.

	Npar	χ^2	*dl*	CFI	NNFI	RMSEA	[CI]
Model with intrinsic motivation as a mediator							
Model for the whole population	82	84.16*	38	0.98	0.96	0.02	[0.01, 0.03]
1-Configural model (sex)	142	78.73*	66	0.99	0.99	0.01	[0.00, 0.02]
2-Saturation (S) between sexes	138	81.30*	70	0.99	0.99	0.01	[0.00, 0.02]
3-(S) + saturation of identical items over time (ST)	136	82.84*	72	0.99	0.99	0.01	[0.00, 0.02]
4-(S) + (ST) + intercepts (I)	130	94.03*	78	0.99	0.99	0.01	[0.00, 0.02]
5-(S) + (ST) + (I) + residual errors (U)	124	192.03*	84	0.96	0.93	0.03	[0.03, 0.04]
5a-(S) + (ST) + (I) + residual errors (U) of item 2 and 3 are relaxed	128	103.87*	80	0.99	0.98	0.016	[0.01, 0.03]
6-(S) + (ST) + (I) + U + correlation of u of identical items over time (CU)	125	105.16*	83	0.99	0.99	0.016	[0.00, 0.02]
7-(S) + (ST) + (I) + U + (CU) + Var-cov (CV)	98	158.81*	110	0.98	0.97	0.02	[0.01, 0.03]
8-(S) + (ST) + (I) + U + (CU) + (CV) + Path	89	202.03*	119	0.97	0.96	0.03	[0.02, 0.03]
Model with inattention as mediator							
Model with the whole population	82	50.10*	37	0.99	0.99	0.01	[0.00, 0.02]
1-Configural model (sex)	142	77.05*	66	0.99	0.99	0.01	[0.00, 0.02]
2-Saturation (S) between sexes	138	84.81*	70	0.99	0.99	0.01	[0.00, 0.02]
3-(S) + saturation of identical items over time (ST)	136	85.82*	72	0.99	0.99	0.01	[0.00, 0.02]
4-(S) + (ST) + intercepts (I)	130	90.09*	78	0.99	0.99	0.01	[0.00, 0.02]
5- (S) + (ST) + (I) + residual errors (U)	124	118.47*	84	0.99	0.99	0.02	[0.01, 0.03]
6-(S) + (ST) + (I) + U + correlation of u of identical items over time (CU)	121	118.94*	87	0.99	0.99	0.03	[0.01, 0.03]
7-(S) + (ST) + (I) + U + (CU) + Var-cov (CV)	94	176.76*	114	0.99	0.99	0.03	[0.02, 0.03]
8-(S) + (ST) + (I) + U + (CU) + (CV) + Path	85	234.15*	123	0.98	0.98	0.03	[0.02, 0.03]
Model with frequency of reading as mediator							
Model with the whole population	82	67.49*	37	0.98	0.96	0.02	[0.01, 0.03]
1-Configural model (sex)	142	104.27*	66	0.98	0.96	0.02	[0.01, 0.03]
2-Saturation (S) between sexes	138	112.20*	70	0.98	0.95	0.02	[0.02, 0.03]
3-(S) + saturation of identical items over time (ST)	136	117.47*	72	0.98	0.95	0.02	[0.02, 0.03]
4-(S) + (ST) + intercepts (I)	130	121.77*	78	0.98	0.96	0.02	[0.02, 0.03]
5- (S) + (ST) + (I) + residual errors (U)	124	126.39*	84	0.98	0.96	0.02	[0.01, 0.03]
6-(S) + (ST) + (I) + U + correlation of u of identical items over time (CU)	121	128.31*	87	0.98	0.96	0.02	[0.01, 0.03]
7-(S) + (ST) + (I) + U + (CU) + Var-cov (CV)	94	167.44*	114	0.98	0.96	0.02	[0.01, 0.03]
8-(S) + (ST) + (I) + U + (CU) + (CV) + Path	85	172.79*	123	0.97	0.96	0.02	[0.01, 0.03]
Model with leisure reading as mediator							
Model with the whole population	65	0.00*	0	1.00	1.00	0.00	[0.00, 0.00]
1-Configural model (sex)	108	0.00*	0	1.00	1.00	0.00	[0.00, 0.00]
2-Var-cov (CV)	81	44.52*	27	0.98	0.95	0.02	[0.01, 0.04]
3-(CV) + Path	72	58.82*	36	0.97	0.95	0.02	[0.01, 0.04]

*Npar is the number of parameters estimated; dl is the degree of freedom; CFI is the "Comparative Fit Index"; TLI is the "Tucker-Lewis Index" and RMSEA is the "Root Mean Square Error of Approximation." *Means that statistically significant at the 5% level.*

Structural Equation Models

Our statistical analyses were performed with Mplus software (Version 7.4; Muthén and Muthén, 2012) and the results presented are standardized. The four models have been tested with the Maximum Likelihood Robust (MLR) estimator. Only the model that included leisure reading as a mediating variable was fully saturated. For the other three models, which contained latent constructs (interaction with the parent, motivation and inattention), we have correlated the error terms of identical items appearing at several measurement times (Marsh and Hau, 1996). In addition, we have assessed whether these three models fitted the data adequately. To do this, we have selected three indices: the "Comparative Fit Index" (CFI), the "Tucker-Lewis Index" (TLI) and the "Root-Mean-Square Error of Approximation" (RMSEA). Our models were considered well adjusted if the CFI and TLI indices were greater than 0.90 and if the RMSEA was less than 0.08 (Hu and Bentler, 1999). We used the "indirect model" procedure to calculate the size of the total and indirect effects (Frazier et al., 2004) of the TV viewing time on the 10-year old RA.

Gender Invariance

In order to test our hypothesis regarding gender, we performed invariance analyses, which consisted in constraining certain parameters of our models to be equal between girls and boys. These analyses were composed of eight models (Caron, 2019) ranging from the unrestricted model (**Table 3**, line 1) to the most restrictive model (**Table 4**, line 8). Model 1 did not constrain any parameters across genders. In model 2, factor loadings were constrained to equality across genders. In

TABLE 4 | Correlations between not answering one of these variables with gender and the PMK diploma (0 = answer provided and 1 = no answer).

	Gender	PMK diploma
Parent-child interaction item 1 to 6 years old	−0.06	−0.10
Parent-child interaction item 2 to 6 years old	−0.07	−0.12
Parent-child interaction item 3 to 6 years old	−0.06	−0.10
Parent-child interaction item 1 to 8 years old	−0.07	−0.12
Parent-child interaction item 2 to 8 years old	−0.06	−0.10
Parent-child interaction item 3 to 8 years old	−0.07	−0.12
Inattention item 1 to 7 years old	−0.09	−0.08
Inattention item 2 to 7 years old	−0.10	−0.06
Inattention item 3 to 7 years old	−0.09	−0.08
Inattention item 1 to 8 years old	−0.10	−0.06
Inattention item 2 to 8 years old	−0.09	−0.08
Inattention item 3 to 8 years old	−0.10	−0.05
Leisure reading at 6 years old	−0.09	−0.15
Leisure reading at 8 years old	−0.07	−0.10
Intrinsic motivation item 1 to 7 years old	−0.10	−0.09
Intrinsic motivation item 2 to 7 years old	−0.10	−0.07
Intrinsic motivation item 3 to 7 years old	−0.10	−0.09
Intrinsic motivation item 1 to 8 years old	−0.10	−0.07
Intrinsic motivation item 2 to 8 years old	−0.10	−0.09
Intrinsic motivation item 3 to 8 years old	−0.10	−0.07
Reading achievement at 7 years old	−0.09	−0.07
Reading achievement at 10 years old	−0.06	−0.09
TV viewing time at 6 years old	−0.06	−0.09

All correlation coefficients are statistically significant at $p < 0.05$.

model 3, factor loadings were constrained to equality across measurement times. Models 2 and 3 offered the possibility to verify if the participants understood items in the same way over time and if gender differences existed in items comprehension. Thereafter, we constrained various parameters including the intercepts (model 4), the residual errors (model 5), the correlated uniquenesses (model 6), the variances and covariances (model 7) as well as the paths (model 8). The comparisons among these models were made as follows: when the more restrictive model indicated a decrease of 0.01 in CFI and TLI, but an increase of 0.015 in the RMSEA compared to the less restrictive one (e.g., model 2 vs. model 1), the least restrictive model would be considered as better fitting the data.

RESULTS

Preliminary Analyses

Missing Data

According to Koolstra et al. (1996, 1997), missing data are associated with lesser RA and higher amount of TV viewing. Thus, it was important to adjust for missing data to avoid bias. Among all of our variables, the gender and diploma of the PMK showed little missing data (see **Table 2**). These two variables allowed us to compare the participants in our sample who had missing data with those who did not. To compare these two

TABLE 5 | Correlations between variables of the 4 models.

	Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	TV viewing time at 6 years old	–														
2	Inattention at 7 years old	0.08*	–													
3	Inattention at 8 years old	0.10*	0.66*	–												
4	Motivation to read at 7 years old	−0.05	−0.25*	−0.19*	–											
5	Motivation to read at 8 years old	−0.05	−0.2*	−0.18*	0.37	–										
6	Parental practices at 6 years old	−0.14*	−0.03	0.01	−0.02	0.046	–									
7	Parental practices at 7 years old	−0.04	−0.10*	−0.04	0.10*	0.08*	0.61*	–								
8	Gender	−0.03	−0.22*	−0.22*	0.16*	0.18*	−0.03	−0.02	–							
9	PMK diploma	−0.18*	−0.22*	−0.21*	0.08*	0.04	0.10*	0.07	−0.00	–						
10	IQ at 6 years old	−0.08*	−0.30*	−0.29*	0.03	0.12*	0.08*	0.13*	0.019	0.24*	–					
11	Talk about school activities 7 years old	−0.03	0.00	−0.01	0.01	0.10*	0.12*	0.18*	0.03	0.08*	0.03	–				
12	parental valorization of grades	0.02	−0.05	−0.03	0.05	0.12*	−0.05	−0.03	0.02	−0.07*	−0.08*	0.027	–			
13	Reading score at 7 years old	−0.13*	−0.57*	−0.49*	0.27*	0.22*	0.07	0.04	0.12*	0.28*	0.40*	0.03	0.07*	–		
14	Reading score at 10 years old	−0.09*	−0.49*	−0.53*	0.16*	0.23*	0.03	0.09*	0.18*	0.24*	0.42*	0.02	0.05	0.62*	–	
15	Leisure reading at 6 years old	−0.08*	−0.16*	−0.07*	0.02*	0.12*	0.14*	0.11*	0.18*	0.07*	0.08*	0.07*	0.05	0.14*	0.14*	–
16	Leisure reading at 8 years old	−0.11*	−0.18*	−0.15*	0.22*	0.35*	0.13*	0.12*	0.23*	0.07*	0.10*	0.11*	0.05	0.23*	0.22*	0.26*

*The * indicates a p-value less than 5% Motivation, parental practices and inattention variables are latent constructs.*

TABLE 6 | All beta and standard error values for the four models (including parameters for covariates).

	Mediator at 8 years old		Reading achievement at 7 years old		Reading achievement at 10 years old		TV viewing time at 6 years old		Mediator at 7 years old	
Model with inattention as mediator	**Beta**	**SE**	**Beta**	**SE**	**Beta**	**SE**	**Beta**	**SE**	**Beta**	**SE**
TV viewing time at 6 years old	0.03	0.03	−0.06	0.03	0.01	0.03			0.03	0.03
Inattention at 7 years old	0.55*	0.03			−0.03	0.05				
Inattention at 8 years old					−0.25*	0.04				
Gender[a]	−0.11*	0.03	0.03	0.02	0.07*	0.03	−0.03	0.03		
PMK diploma	−0.03	0.03	0.19*	0.03	0.03	0.03	−0.18*	0.02	−0.15*	0.03
IQ at 6 years old	−0.04	0.03	0.34*	0.03	0.19*	0.03			−0.25*	0.03
Talk about school activities at 7 years old	0.01	0.03			−0.01	0.03				
parental valorization of grades at 7 years old	−0.00	0.03			0.03	0.03				
Reading achievement at 7 years old	−0.15*	0.04			0.38*	0.03				
Model with intrinsic motivation as mediator										
TV viewing time at 6 years old	−0.02	0.03	−0.07*	0.03	0.00	0.03			−0.04	0.03
Intrinsic motivation at 7 years old	0.32*	0.04			−0.04	0.04				
Intrinsic motivation at 8 years old					0.09*	0.03				
Gender	0.13*	0.03	0.09*	0.03	0.10*	0.03	−0.03	0.03		
PMK diploma	−0.03	0.03	0.19*	0.03	0.05	0.03	−0.18*	0.02	0.07*	0.03
IQ at 6 years old	0.08*	0.04	0.34*	0.03	0.20*	0.03			0.02	0.03
Talk about school activities 7 years old	0.09*	0.04			−0.02	0.03				
parental valorization of grades at 7 years old	0.09*	0.03			0.03	0.03				
Reading achievement at 7 years old	0.09*	0.04			0.50*	0.03				
Model with frequency of PMK-child interactions as mediator										
TV viewing time at 6 years old	0.04	0.04	−0.06*	0.03	−0.01	0.03				
PMK-child interactions at 6 years old	0.60*	0.05	0.01	0.04	−0.09	0.05				
PMK-child interactions at 8 years old					0.09	0.05				
Gender	0.00	0.03	0.12*	0.07	0.11*	0.03	−0.03	0.03	−0.03	0.04
PMK diploma	−0.01	0.04	0.19*	0.03	0.05	0.03	−0.18*	0.02	0.10*	0.04
IQ at 6 years old	0.10*	0.04	0.34*	0.03	0.20*	0.03				
Talk about school activities 7 years old	0.11*	0.04			−0.01	0.03				
parental valorization of grades at 7 years old	0.01	0.03			0.03	0.03				
Reading achievement at 7 years old	−0.03	0.04			0.51*	0.03				
Model with leisure reading as mediator										
TV viewing time at 6 years old	−0.07*	0.03	−0.05	0.03	0.07	0.03				
Leisure reading at 6 years old	0.19*	0.03	0.09*	0.03	0.01	0.03				
Leisure reading at 8 years old					0.07*	0.03				
Gender	0.17*	0.03	0.10*	0.03	0.09*	0.03	−0.03	0.03	0.18*	0.03
PMK diploma	−0.00	0.03	0.19*	0.03	0.05	0.03	−0.18*	0.02	0.06*	0.03
IQ at 6 years old	0.01	0.03	0.33*	0.03	0.21*	0.03				
Talk about school activities 7 years old	0.07*	0.03			−0.02	0.03				
parental valorization of grades at 7 years old	0.03	0.03			0.03	0.03				
Reading achievement at 7 years old	0.16*	0.04			0.49*	0.03				

**indicates a p-value less than 5%.*

[a]0 = boy and 1 = girl.

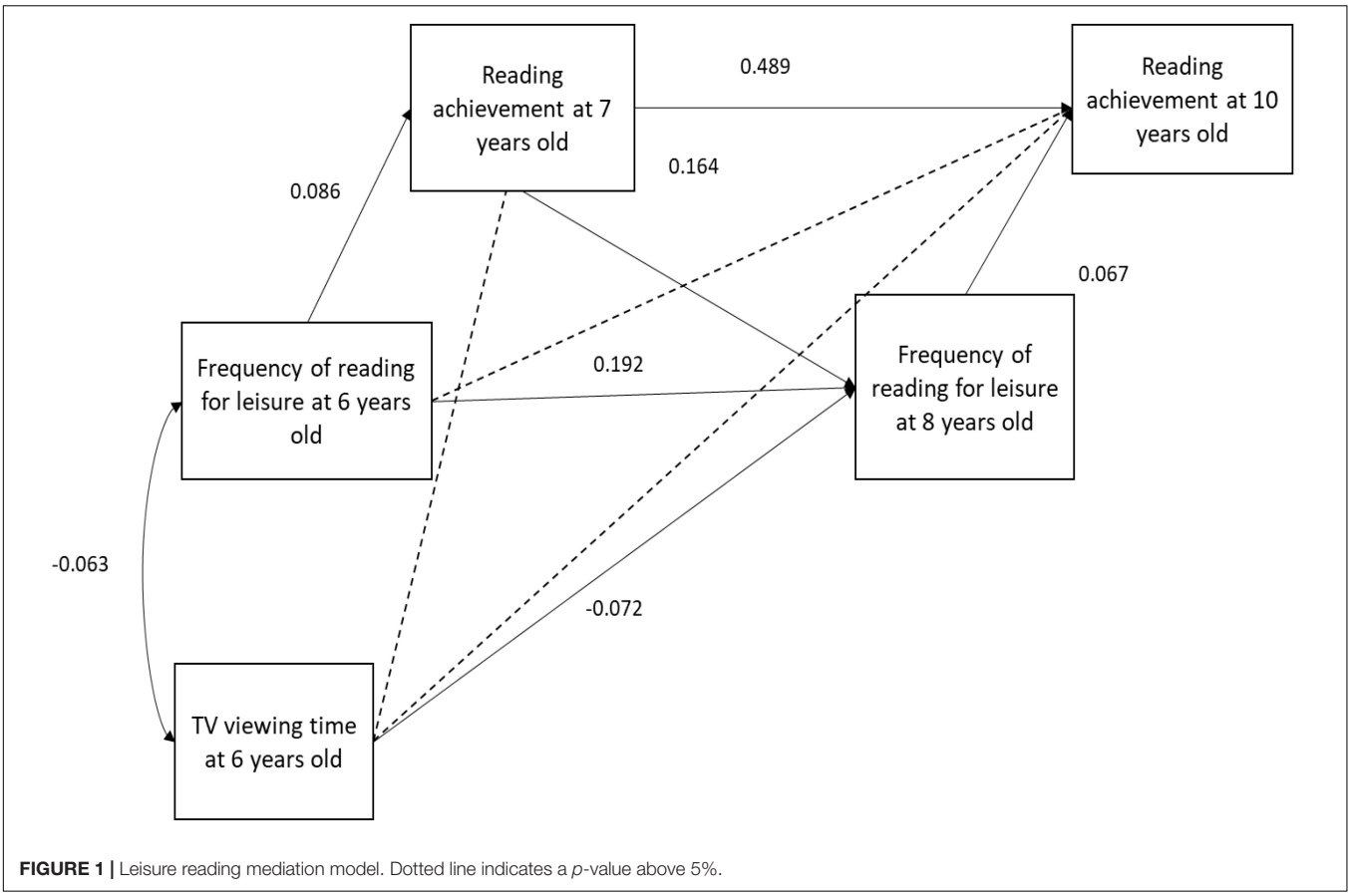

FIGURE 1 | Leisure reading mediation model. Dotted line indicates a *p*-value above 5%.

subsamples, we first computed a dichotomized score based on responses provided on TV viewing time, RA, IM, leisure reading, inattention and parent-child interaction. Those who had non-missing scores on these variables were assigned 0, whereas those who had missing scores were assigned 1. Thus, six dummy variables were computed. Then, we correlated these six dummies to the gender and educational degree of the PMK. Results (see **Table 4**) indicate that more missing values are observed among boys, r between –0.052 and –0.104 and children who live in a household where the PMK has a low level of education, r between –0.06 and –0.12. However, these correlations were quite low.

Thus, the estimation of our models was carried out using the FIML procedure (Muthén and Muthén, 2012), which is a more robust procedure than complete case analysis or imputation with the mean (Caron, 2019). In addition, several participants (29.8%, n = 644) presented missing data on more than one variable included in our models. We thus tested our models with and without these participants and we did not observe any meaningful difference in the results. For this reason, all participants were kept in our analyses.

Descriptive Statistics

Table 5 indicates that TV viewing time is statistically significantly associated with RA at 7 and 10 years old, as well as inattention, leisure reading frequency, and the frequency of parent-child interactions at 6 years old. Although statistically significant, all of these relationships were nonetheless very modest. **Table 2** shows that, at all measurement times, boys read less frequently for leisure than girls, that they have on average a lower level of IM to read as well as a lower RA. However, there were no gender differences in the average time spent watching TV at 6 years old.

Models Tested

Hypothesis Testing

Table 3 shows the results of the fit indices for the four structural equation models tested. These indices show a very good level of fit because the CFI and TLI values are above 0.95 and the RMSEA values are below 0.025.

Table 6 presents the beta and the standard error of paths from the four models. Among the covariates, IQ at 6 years old and PMK diploma are the most associated with the mediator variables and the RA variables. In addition, the PMK diploma is the most important predictor of TV viewing time. These results corroborate those of other studies (Koolstra et al., 1997; Ennemoser and Schneider, 2007).

Figures 1–4 present the results of the four models that test the substitution (**Figures 1**, **2**) and inhibition (**Figures 3**, **4**) hypotheses. Among these four models, the only mediator to be associated with TV viewing time was the frequency of leisure reading. More specifically, TV viewing time at 6 years old was negatively associated to leisure reading frequency at 8 years old ($\beta = -0.072$; SE = 0.033). However, this association was too small

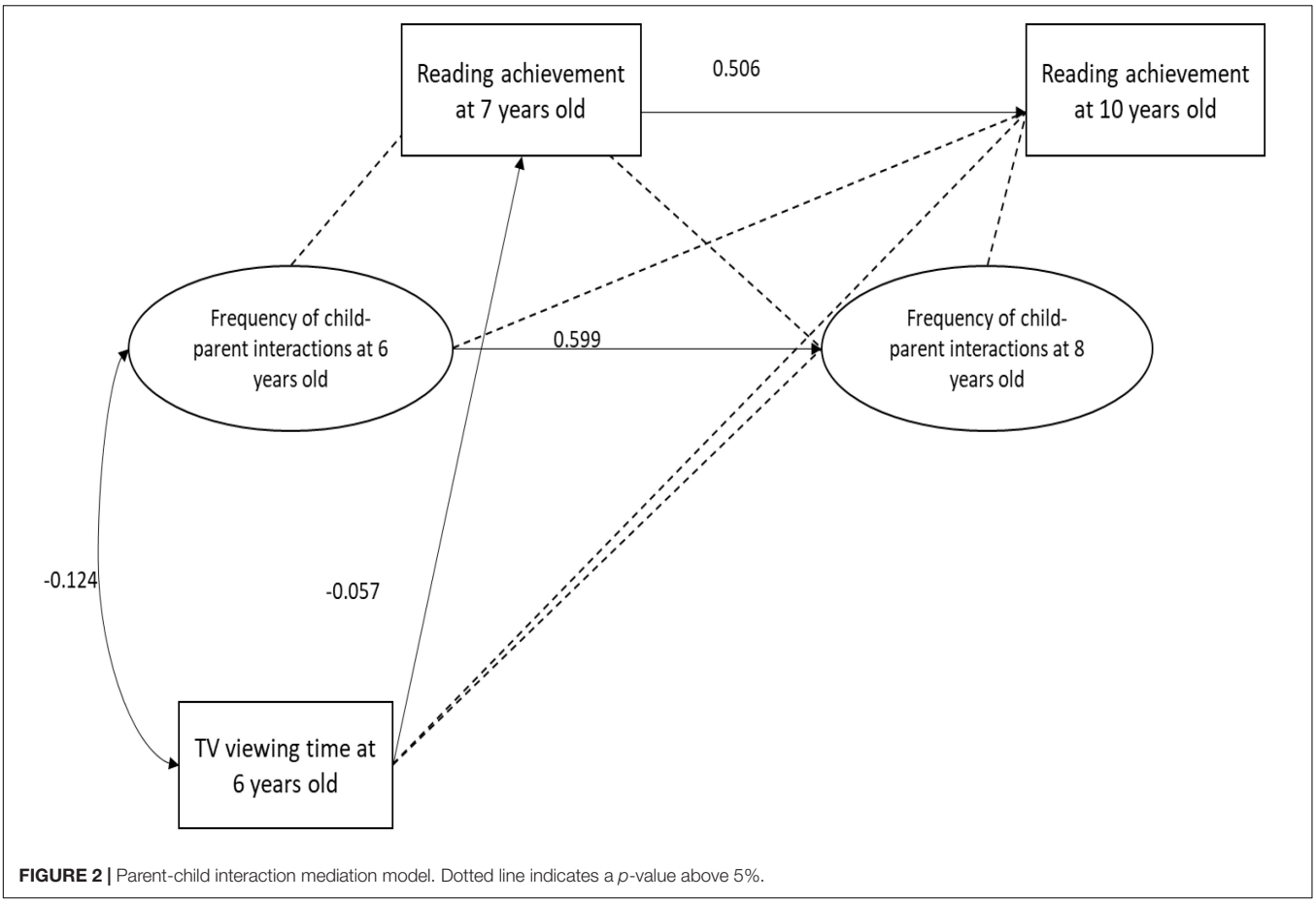

FIGURE 2 | Parent-child interaction mediation model. Dotted line indicates a *p*-value above 5%.

to produce indirect effects on RA at 10 years old (β = −0.005; $p = 0.117$).

Table 7, which presents the results of the indirect and total effects, indicates that, with the exception of the parent-child interactions model, the indirect effects of our models are statistically significant, but with small effect sizes (β between −0.036 and −0.032). First, these indirect effects of TV viewing time at 6 years old on RA at 10 years old were caused almost exclusively by the link between TV viewing time and RA at 7 years old. Because these indirect links were not caused by the mediating variables from which we have tested the substitution or inhibition effects, our results therefore did not support the inhibition and substitution hypotheses. Second, there was no statistically significant total effect in the 4 models. This result means that the sum of the direct and indirect effects of the 6 year olds TV viewing time toward the 10 years old RA was not statistically different from zero.

Invariance Analysis

The results of our invariance analysis indicated that there were no differences between the two groups concerning the associations between the time spent watching TV, our 4 mediators and RA. First, the results of models that tested parent-child attention and interaction as mediators yielded poorer model fit indices (see **Table 3**). Second, in the model where the IM is the mediator, our results indicate a drop in the acceptable level of the adjustment indices when residual errors are constrained (**Table 3**, A a q line 5). We therefore removed some of these constraints (see model 5a) and we then constrained the correlated uniquenesses (model 6), the variances and covariances (model 7) as well as the paths (model 8). For model 8, results indicate a little drop in CFI and NNFI values. Thus, we relaxed these constraints and we calculated the differences between genders for the regression coefficients. Our results show small differences between gender for association between IM at 7 years old and motivation at 8 years old, and between TV viewing time at 6 years old and RA at 7 years old. However, there is no difference between genders for associations between TV viewing time at 6 years old and IM at 8 years old and between the TV viewing time at 6 years old and RA at 10 years old (**Table 8**). Third, in the model where leisure reading is the mediator, our results indicate a decrease in the acceptable level of adjustment indices for the Residual Invariance Model (**Table 3**, model 2). However, our results do not indicate a further decline in these adjustment indices when the path coefficients are constrained to equality (**Table 3**, model 7). Thus, in this model with leisure reading as a mediator, the differences between genders were only found on the variances/covariances and not on the relationship between the time spent watching TV, leisure reading, and RA. In sum, the invariance analysis performed on the four models did not corroborate our second

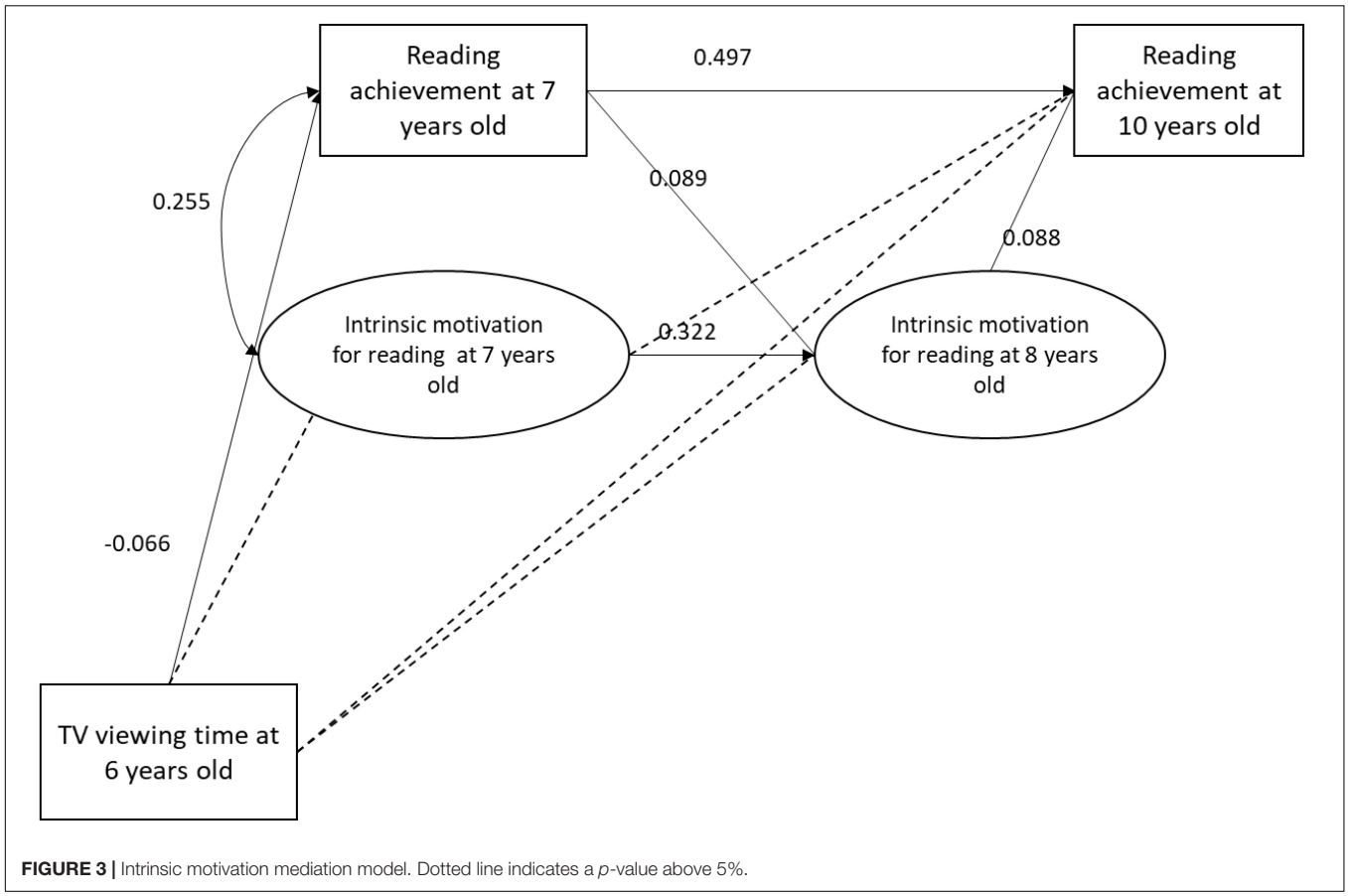

FIGURE 3 | Intrinsic motivation mediation model. Dotted line indicates a *p*-value above 5%.

hypothesis, which proposed that boys are more exposed than girls to the effects of substitution and inhibition.

DISCUSSION

The goal of this study was to estimate the contribution of four potential mediating variables (leisure reading, child-parent interaction frequency, IM to read and the level of inattention) explaining the relationship between TV viewing time and RA as a function of gender. These four mediators were chosen to test the substitution and inhibition hypotheses. In addition, we hypothesized that the size of the substitution and inhibition effects would be greater for boys. Overall, our results did not support these assumptions. First, only the leisure reading frequency was negatively associated with TV viewing time at 6 years old, however, this negative association was very small and had no indirect effect on RA at 10 years old. Second, relationships between TV viewing time at 6 years old and our mediating variables at age 9 and RA at age 10 did not vary across genders. Therefore, our results do not corroborate the substitution and inhibition hypotheses, nor do they corroborate our hypothesis proposing that TV viewing would be more detrimental to boys' RA than to the one of girls.

The substitution hypothesis assumes that the time children spend on activities that are favorable to their RA and the time they spend watching TV is organized as a zero-sum game (Peaucelle, 1969). However, our results show, that activities that are favorable to RA are not, or only slightly, replaced by the time that students spend watching TV. One explanation for these results is that children watch TV during times when they probably would not have chosen to perform activities more favorable to their RA. Furthermore, our results also suggest that children practice activities favorable to their RA when they are not permitted to watch TV. For example, children who read at night before sleep when their parents make sure that they cannot watch TV would not see time they spend on this activity decrease if they are more exposed to TV in afternoon. Thus, children who watch more TV are not doing this activity at the expense of the time they spend leisure reading or interacting with their parents.

The inhibition hypothesis is also not corroborated. Specifically, results indicate that there is no association between TV viewing time and IM to read and inattention. A first explanation could be that principles on which the inhibition hypothesis is based are wrong. Specifically, it may be inadequate to propose that TV viewing is a leisure that does not require effort and attention and thus could induce in children "mental laziness". A second possibility is that the negative influence of TV is not important enough to translate into a measurable drop in IM to read and to an increase in inattention symptoms.

Gender moderation analyses did not indicate a difference between the two groups, which does not support our hypothesis.

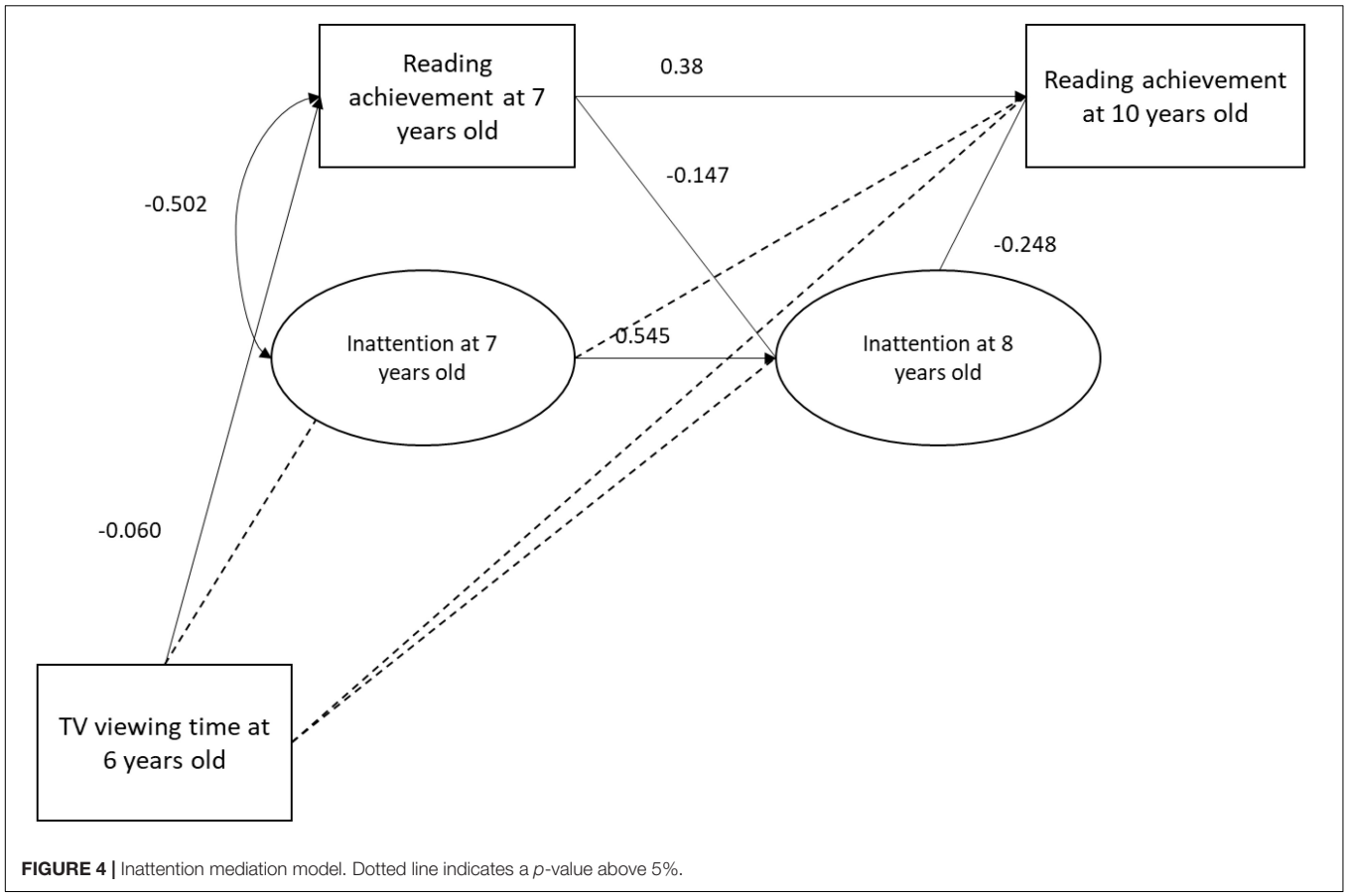

FIGURE 4 | Inattention mediation model. Dotted line indicates a *p*-value above 5%.

TABLE 7 | Indirect and total effects.

Model	Total effect	*p*-value	Indirect effect	*p*-value
Reading	−0.03	0.41	−0.03	0.03
Parental practices (PP)	−0.03	0.00	−0.03	0.10
Motivation	−0.03	0.29	−0.03	0.02
Inattention	−0.03	0.34	−0.04	0.03

TABLE 8 | Results of gender differences.

	β	Standard error	*p*-value
Intrinsic motivation at 8 years old	0.06	0.05	0.22
RA at 10 years old	−0.07	0.06	0.23

This hypothesis is based on the fact that boys and girls have a relationship with reading and TV that differs quantitatively and qualitatively. Our results indicate that, on the contrary, gender does not affect the relationship between TV viewing time, the mediating variables and the RA. As we have suggested, there may be substitution and inhibition moderators that have not been studied yet, such as social and economic status, age of children or the type of content children are watching. In this sense, if gender is not the most relevant moderator, the choice of another moderator should be considered in future studies.

LIMITS

A first limitation of our study concerns the TV viewing time measure. Indeed, several researchers questioned the accuracy and validity of this measure particularly with regard to the level of measurement error it contains and its relation with social desirability (see Bryant et al., 2007 for a systematic review of these studies from 1985 to 2006; Atkin et al., 2012; Cabanas-Sánchez et al., 2018; Byrne et al., 2021). However, the impact of this shortcoming seems to be trivial for several reasons. Indeed, since the 1980s, the amount of TV viewing time obtained via a self-report measure has been compared to the time derived from objective measures (video or direct observation; see Bryant et al., 2007 for a systematic review of these studies from 1985 to 2006; Clark et al., 2009; Atkin et al., 2012; Cabanas-Sánchez et al., 2018; Aunger and Wagnild, 2020; Byrne et al., 2021). Researchers conclude that self-report questionnaires have an acceptable level of validity (Gorin et al., 2006; Otten et al., 2010; Dwyer et al., 2011; Foley et al., 2012; Wijndaele et al., 2014; Cabanas-Sánchez et al., 2018; Aunger and Wagnild, 2020; De Moraes et al., 2020). Second, the QLSCD comprised a social desirability scale that we used to calculate the correlation with the TV viewing time variable. This correlation is -0.04. Thus, TV viewing time does not seem to be affected by the degree of social desirability of the participants. Third, Munasib and Bhattacharya (2010) and Nakamuro et al. (2013) measured the impact that measurement error can have

on the estimation of the relationship between TV viewing time and RA. These authors concluded that measurement error has no impact on results. For these reasons, although it is important to consider that the questionnaire has limitations in measuring TV viewing time, this aspect does not seem to invalidate our results.

Third, the mediating and dependent variables associated with the TV variable are spaced by an interval of two years between each time point. This time interval is imposed by QLSCD sampling. However, it is unknown if and how the duration of time between TV viewing, mediating variables and RA affects effect sizes. To our knowledge, this question of temporality on the link between TV and RA has not yet been studied. Thus, it would be relevant to address this question in further studies.

Fourth, if we have taken into account several important confounding variables, other sources of bias might nevertheless operate. A potential source of bias could come from parents who use TV as a means of reward and punishment (Huang and Lee, 2010). Indeed, this practice consists of increasing the time spent watching TV when children have good grades and decreasing it when children have poor ones. Such contingent use of TV would result in a positive association between TV and RA, that is not mainly attributable to the real effects of TV. However, it seems to us that this risk of bias is relatively low since our TV variable measures the viewing time of children before they enter primary school. Thus, the children in our sample are not subject to a school evaluation that parents could use to regulate their time spent watching TV.

Finally, TV viewing time was measured with 6 years old children. Our results are therefore limited to young children and do not seem to be replicable to an older population such as adolescents. In this regard, no study has tested the substitution and inhibition hypotheses jointly in a population of adolescents. It would therefore be interesting to test these two hypotheses with this population.

CONCLUSION

The main concerns and criticisms linked to TV viewing are that it replaces reading in children's leisure time, reduces their interest in this activity and increases their inattention, which would harm the development of their competencies at school. However, our results indicate that watching TV is not associated with lower RA and that the drop in the amount of time spent leisure reading is not enough to affect RA. On the social level, our results therefore provide useful input to the debate on TV. Our results do not support the substitution and inhibition hypotheses while controlling for important covariates. However, it seems wrong to consider that these results completely invalidate these two hypotheses for three reasons.

First, the research that has tested these two hypotheses presents mixed results. If some studies obtained results similar to ours, indicating that the time spent watching TV is very weakly and negatively associated with the time spent leisure reading (Koolstra et al., 1996; Ennemoser and Schneider, 2007) and that it is not associated with IM to read and inattention (Ansari and Crosnoe, 2016), other studies have shown different results (Ritchie et al., 1987; Koolstra et al., 1996; Anderson et al., 2001; Wright et al., 2001; Zimmerman et al., 2007). Our results should therefore be interpreted with caution.

Second, there is an increasing presence of new types of screens such as digital tablets, telephones or laptops (Kostyrka-Allchorne et al., 2017). These screens present major differences when compared to TV. Unlike a fixed screen, they allow the viewers to access a large variety of content easily and quickly, right in the palm of their hands, anywhere, anytime. Thus, it seems important to test the substitution and inhibition hypotheses in this context of new screens. Considering that watching TV shows is one of the main activities that children perform with these screens (Rideout, 2016) and considering that there is still little research on our subject, it therefore seems socially and scientifically important to emphasize the need to undertake additional studies in order to have a more substantiated knowledge on the relationship between exposure to TV or streaming programs, children's RA and the mediators and moderators likely to explain that relation.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://www.jesuisjeserai.stat.gouv.qc.ca/default_an.htm.

AUTHOR CONTRIBUTIONS

WS conducted the study process, performed data analysis, and drafted the manuscript. FG supervised the design of the study. FG and DT contributed to the data analysis and results interpretation, critically revised the article, and approved this version. All authors provided important contributions to the work.

REFERENCES

Acevedo-Polakovich, I. D., Lorch, E. P., Milich, R., and Ashby, R. D. (2006). Disentangling the relation between television viewing and cognitive processes in children with attention-deficit/hyperactivity disorder and comparison children. *Arch. Pediatr. Adolesc. Med.* 160, 354–360.

Anderson, D., Huston, A., Schmitt, K., Linebarger, D., Wright, J., and Larson, R. (2001). Early childhood television viewing and adolescent behavior: the Recontact Study. *Monogr. Soc. Res. Child Dev.* 66, I–154. doi: 10.1111/1540-5834.00124

Ansari, A., and Crosnoe, R. (2016). Children's hyperactivity, television viewing, and the potential for child effects. *Child. Youth Serv. Rev.* 61, 135–140. doi: 10.1016/j.childyouth.2015.12.018

Atkin, A. J., Gorely, T., Clemes, S. A., Yates, T., Edwardson, C., Brage, S., et al. (2012). Methods of measurement in epidemiology: sedentary behaviour. *Int. J. Epidemiol.* 41, 1460–1471. doi: 10.1093/ije/dys118

Aunger, J., and Wagnild, J. (2020). Objective and subjective measurement of sedentary behavior in human adults: a toolkit. *Am. J. Hum. Biol.* e23546. doi: 10.1002/ajhb.23546 [Epub ahead of print].

Barr, R., Lauricella, A., Zack, E., and Calvert, S. L. (2010). Infant and early childhood exposure to adult-directed and child-directed television

programming: relations with cognitive skills at age four. *Merrill Palmer Q.* 56, 21–48. doi: 10.1353/mpq.0.0038

Beentjes, J. W., and Van der Voort, T. H. (1988). Television's impact on children's reading. *Read. Res. Q.* 23, 389–413. doi: 10.2307/747640

Bergin, D. A. (1992). Leisure activity, motivation, and academic achievement in high school students. *J. Leis. Res.* 24, 225–239. doi: 10.1080/00222216.1992.11969890

Bittman, M., Rutherford, L., Brown, J., and Unsworth, L. (2011). Digital natives? New and Old Media and Children's Outcomes. *Aust. J. Educ.* 55, 161–175. doi: 10.1177/000494411105500206

Blankson, A. N., O'Brien, M., Leerkes, E. M., Calkins, S. D., and Marcovitch, S. (2015). Do hours spent viewing television at ages 3 and 4 predict vocabulary and executive functioning at age 5? *Merrill Palmer Q.* 61, 264–289. doi: 10.13110/merrpalmquar1982.61.2.0264

Boivin, M., Pérusse, D., Saysset, V., Tremblay, N., and Tremblay, R. (2000). *Conduites Parentales et Relations Familiales, Section I-Les Cognitions et les Conduites Parentales. Étude Longitudinale du Développement des Enfants du Québec (ÉLDEQ 1998-2002)*. Available online at: https://bdso.gouv.qc.ca/docs-ken/multimedia/PB01671FR_ELDEQ_vol1no10_2001H00F02.pdf (accessed October 01, 2021).

Boyle, M. H., Offord, D. R., Racine, Y., Sanford, M., Szatmari, P., and Fleming, J. E. (1993). Evaluation of the original Ontario child health study scales. *Can. J. Psychiatry* 38, 397–405. doi: 10.1177/070674379303800605

Bryant, M. J., Lucove, J. C., Evenson, K. R., and Marshall, S. (2007). Measurement of television viewing in children and adolescents: a systematic review. *Obes. Rev.* 8, 197–209. doi: 10.1111/j.1467-789X.2006.00295.x

Byrne, R., Terranova, C. O., and Trost, S. G. (2021). Measurement of screen time among young children aged 0–6 years: a systematic review. *Obes. Rev.* 22:e13260. doi: 10.1111/obr.13260

Cabanas-Sánchez, V., Martínez-Gómez, D., Esteban-Cornejo, I., Castro-Piñero, J., Conde-Caveda, J., and Veiga, Ó. L. (2018). Reliability and validity of the youth leisure-time sedentary behavior questionnaire (YLSBQ). *J. Sci. Med. Sport* 21, 69–74. doi: 10.1016/j.jsams.2017.10.031

Cardin, J.-F., Desrosiers, H., Belleau, L., Giguère, C., and Boivin, M. (2011). *Les Symptômes D'Hyperactivité et d'Inattention chez les Enfants de la Période Préscolaire à la Deuxième Année du Primaire. Portraits et trajectoires. Série Étude longitudinale du développement des enfants du Québec-ÉLDEQ*. Available online at: https://bdso.gouv.qc.ca/docs-ken/multimedia/PB01671FR_hyperactivite2010H00F00.pdf (accessed October 01, 2021).

Caron, P.-O. (2019). *La Modélisation par Équations Structurelles avec Mplus*. Québec: PUQ. doi: 10.2307/j.ctvt1sh9g

Cheng, S., Maeda, T., Yoichi, S., Yamagata, Z., Tomiwa, K., and Group, J. C. S. (2010). Early television exposure and children's behavioral and social outcomes at age 30 months. *J. Epidemiol.* 20, S482–S489. doi: 10.2188/jea.JE20090179

Childers, J. S., Durham, T. W., and Wilson, S. (1994). Relation of Performance on the Kaufman Brief Intelligence Test with the Peabody Picture Vocabulary Test—Revised among Preschool Children. *Percept. Mot. Skills* 79, 1195–1199. doi: 10.2466/pms.1994.79.3.1195

Christakis, D. A., Zimmerman, F. J., DiGiuseppe, D. L., and McCarty, C. A. (2004). Early television exposure and subsequent attentional problems in children. *Pediatrics* 113, 708–713. doi: 10.1542/peds.113.4.708

Clark, B. K., Sugiyama, T., Healy, G. N., Salmon, J., Dunstan, D. W., and Owen, N. (2009). Validity and reliability of measures of television viewing time and other non-occupational sedentary behaviour of adults: a review. *Obes. Rev.* 10, 7–16. doi: 10.1111/j.1467-789X.2008.00508.x

Clark, C. (2012). *Children's and Young People's Reading Today: Findings from the 2011 National Literacy Trust's Annual Survey*. London: National Literacy Trust.

Coyne, S. M., Linder, J. R., Rasmussen, E. E., Nelson, D. A., and Collier, K. M. (2014). It's a bird! It's a plane! It's a gender stereotype!: longitudinal associations between superhero viewing and gender stereotyped play. *Sex Roles* 70, 416–430. doi: 10.1007/s11199-014-0374-8

De Moraes, A. C. F., Nascimento-Ferreira, M. V., de Moraes Forjaz, C. L., RAistizabal, J. C., Azzaretti, L., Junior, W. V. N., et al. (2020). Reliability and validity of a sedentary behavior questionnaire for South American pediatric population: SAYCARE study. *BMC Med. Res. Methodol.* 20:5. doi: 10.1186/s12874-019-0893-7

Desmurget, M. (2011). *Tv Lobotomie: La Vérité Scientifique sur les Effets de la Télévision*. Paris: Max Milo.

Duch, H., Fisher, E. M., Ensari, I., Font, M., Harrington, A., Taromino, C., et al. (2013). Association of Screen Time Use and Language Development in Hispanic Toddlers: A Cross-Sectional and Longitudinal Study. *Clin. Pediatr.* 52, 857–865. doi: 10.1177/0009922813492881

Dunn, L. M., Dunn, L. M., and Thériault-Whalen, C. M. (1993). *Échelle de Vocabulaire en Images Peabody: EVIP*. Toronto: Psycan.

Dwyer, G. M., Hardy, L. L., Peat, J. K., and Baur, L. A. (2011). The validity and reliability of a home environment preschool-age physical activity questionnaire (Pre-PAQ). *Int. J. Behav. Nutr. Phys. Act.* 8, 1–13. doi: 10.1186/1479-5868-8-86

Ennemoser, M., and Schneider, W. (2007). Relations of television viewing and reading: findings from a 4-year longitudinal study. *J. Educ. Psychol.* 99, 349–368. doi: 10.1037/0022-0663.99.2.349

Foley, L., Maddison, R., Olds, T., and Ridley, K. (2012). Self-report use-of-time tools for the assessment of physical activity and sedentary behaviour in young people: systematic review. *Obes. Rev.* 13, 711–722. doi: 10.1111/j.1467-789X.2012.00993.x

Forget-Dubois, N., Lemelin, J.-P., Boivin, M., Dionne, G., Séguin, J. R., Vitaro, F., et al. (2007). Predicting early school achievement with the EDI: a longitudinal population-based study. *Early Educ. Dev.* 18, 405–426. doi: 10.1080/10409280701610796

Foster, E. M., and Watkins, S. (2010). The value of reanalysis: TV viewing and attention problems. *Child Dev.* 81, 368–375. doi: 10.1111/j.1467-8624.2009.01400.x

Frazier, P. A., Tix, A. P., and Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *J. Couns. Psychol.* 51, 115–134. doi: 10.1037/0022-0167.51.1.115

Gentzkow, M., and Shapiro, J. M. (2008). Preschool television viewing and adolescent test scores: historical evidence from the Coleman study. *Q. J. Econ.* 123, 279–323. doi: 10.1162/qjec.2008.123.1.279

Gorin, A., Raynor, H., Chula-Maguire, K., and Wing, R. (2006). Decreasing household television time: a pilot study of a combined behavioral and environmental intervention. *Behav. Interv.* 21, 273–280. doi: 10.1002/bin.221

Guay, F., Chanal, J., Ratelle, C. F., Marsh, H. W., Larose, S., and Boivin, M. (2010). Intrinsic, identified, and controlled types of motivation for school subjects in young elementary school children. *Br. J. Educ. Psychol.* 80, 711–735. doi: 10.1348/000709910X499084

Guay, F., Larose, S., Ratelle, C., Sénécal, C., Vallerand, R. J., and Vitaro, F. (2007). *Mes Amis, mes Parents et mes Professeurs: Une Analyse Comparée de leurs Effets Respectifs sur la Motivation, la Réussite, L'orientation et la Persévérance Scolaires*. Québec: Université de Laval.

Guthrie, J. T., Wigfield, A., and You, W. (2012). "Instructional contexts for engagement and achievement in reading," in *Handbook of Research on Student Engagement*, eds S. J. Christenson, A. L. Reschly and C. Wylie (Boston, MA: Springer), 601–634. doi: 10.1007/978-1-4614-2018-7_29

Hamer, M., Stamatakis, E., and Mishra, G. D. (2010). Television-and screen-based activity and mental well-being in adults. *Am. J. Prev. Med.* 38, 375–380. doi: 10.1016/j.amepre.2009.12.030

Himmelweit, H. T., Oppenheim, A. N., and Vince, P. (1958). *Television and the Child*. Oxford: Oxford University Press.

Hornik, R. (1981). Out-of-school television and schooling: hypotheses and methods. *Rev. Educ. Res.* 51, 193–214. doi: 10.3102/00346543051002193

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Huang, F., and Lee, M. J. (2010). Dynamic treatment effect analysis of TV effects on child cognitive development. *J. Appl. Econ.* 25, 392–419. doi: 10.1002/jae.1165

Jetté, M., and Des Groseilliers, L. (2000). *L'enquête: Description et Méthodologie dans Étude Longitudinale du Développement des Enfants du Québec (ÉLDEQ 1998–2002)*. Québec: Institut de la statistique du Québec.

Johnson, J. G., Cohen, P., Kasen, S., and Brook, J. S. (2007). Extensive television viewing and the development of attention and learning difficulties during adolescence. *Arch. Pediatr. Adolesc. Med.* 161, 480–486. doi: 10.1001/archpedi.161.5.480

Koolstra, C. M., Voort, T. H., and Kamp, L. J. T. (1997). Television's Impact on Children's Reading Comprehension and Decoding Skills: a 3-Year Panel Study. *Read. Res. Q.* 32, 128–152. doi: 10.1598/RRQ.32.2.1

Koolstra, C. M., Voort, T. H., and Voort, D. (1996). Longitudinal effects of television on Children's Leisure-time reading a test of three explanatory models. *Hum. Commun. Res.* 23, 4–35.

Kostyrka-Allchorne, K., Cooper, N. R., and Simpson, A. (2017). The relationship between television exposure and children's cognition and behaviour: a systematic review. *Dev. Rev.* 44, 19–58. doi: 10.1016/j.dr.2016.12.002

Landhuis, C. E., Poulton, R., Welch, D., and Hancox, R. J. (2007). Does childhood television viewing lead to attention problems in adolescence? Results from a prospective longitudinal study. *Pediatrics* 120, 532–537. doi: 10.1542/peds.2007-0978

Linebarger, D. L., and Walker, D. (2005). Infants' and toddlers' television viewing and language outcomes. *Am. Behav. Sci.* 48, 624–645. doi: 10.1177/0002764204271505

Lonner, W. J., Thorndike, R. M., Forbes, N. E., and Ashworth, C. (1985). The influence of television on measured cognitive abilities: a study with Native Alaskan children. *J. Cross Cult. Psychol.* 16, 355–380. doi: 10.1177/0022002185016003006

Maass, E. E., Hahlweg, K., Heinrichs, N., Kuschel, A., and Doepfner, M. (2010). Screen media in preschool age: on the relationship between media use, behavior problems, and ADHD. *Eur. J. Health Psychol.* 18, 55–68. doi: 10.1026/0943-8149/a000009

Malanchini, M., Wang, Z., Voronin, I., Schenker, V. J., Plomin, R., Petrill, S. A., et al. (2017). Reading self-perceived ability, enjoyment and achievement: a genetically informative study of their reciprocal links over time. *Dev. Psychol.* 53, 698–712. doi: 10.1037/dev0000209

Manu, M., Torppa, M., Eklund, K., Poikkeus, A. M., Lerkkanen, M. K., and Niemi, P. (2021). Kindergarten pre-reading skills predict Grade 9 reading comprehension (PISA Reading) but fail to explain gender difference. *Read. Writ.* 34, 753–771. doi: 10.1007/s11145-020-10090-w

Marsh, H. W., and Hau, K.-T. (1996). Assessing goodness of fit: Is parsimony always desirable? *J. Exp. Educ.* 64, 364–390. doi: 10.1080/00220973.1996.10806604

Mehmet-Radji, O. (2004). Early television exposure and subsequent attentional problems in children. *Child Care Health Dev.* 30, 559–560. doi: 10.1111/j.1365-2214.2004.00456_4.x

Miller, C. J., Marks, D. J., Miller, S. R., Berwid, O. G., Kera, E. C., Santra, A., et al. (2007). Brief Report: television viewing and risk for attention problems in preschool children. *J. Pediatr. Psychol.* 32, 448–452. doi: 10.1093/jpepsy/jsl035

Morgan, P. L., and Fuchs, D. (2007). Is There a Bidirectional Relationship between Children's Reading Skills and Reading Motivation? *Except. Child.* 73, 165–183. doi: 10.1177/001440290707300203

Morin, M.-F. (2014). *Portrait du Jeune Lecteur Québécois de la 1re Année du Primaire à la 5e Année du Secondaire. Rapport Final.* Québec: Université de Sherbrooke.

Munasib, A., and Bhattacharya, S. (2010). Is the 'Idiot's Box' raising idiocy? Early and middle childhood television watching and child cognitive outcome. *Econ. Educ. Rev.* 29, 873–883. doi: 10.1016/j.econedurev.2010.03.005

Muthén, L. K., and Muthén, B. O. (2012). *Mplus Version 7 User's Guide*. Los Angeles, CA: Muthén & Muthén.

Nakamuro, M., Matsuoka, R., and Inui, T. (2013). *More Time Spent on Television and Video Games, Less Time Spent Studying? Discussion Papers (by fiscal year), 2012, 2011*. Tokyo: Research Institute of Economy, Trade and Industry.

Nanhou, V., Desrosiers, H., Tétreault, K., and Guay, F. (2016). *La Motivation en Lecture Durant L'enfance et le Rendement Dans la Langue D'enseignement à 15 ans*. Québec: Institut de la statistique du Québec.

Otten, J. J., Littenberg, B., and Harvey-Berino, J. R. (2010). Relationship between self-report and an objective measure of television-viewing time in adults. *Obesity* 18, 1273–1275. doi: 10.1038/oby.2009.371

Pagani, L. S., Fitzpatrick, C., and Barnett, T. A. (2013). Early childhood television viewing and kindergarten entry readiness. *Pediatr. Res.* 74, 350–355. doi: 10.1038/pr.2013.105

Paik, H. (2000). Television viewing and high school mathematics achievement: a neural network analysis. *Qual. Quant.* 34, 1–15. doi: 10.1023/A:1004795407624

Parkes, A., Sweeting, H., Wight, D., and Henderson, M. (2013). Do television and electronic games predict children's psychosocial adjustment? Longitudinal research using the UK Millennium Cohort Study. *Arch. Dis. Child.* 98, 341–348. doi: 10.1136/archdischild-2011-301508

Peaucelle, J.-L. (1969). Théorie des jeux et sociologie des organisations. *Sociol. Travail* 11, 22–43. doi: 10.3406/sotra.1969.1410

Pingault, J.-B., Côté, S. M., Galéra, C., Genolini, C., Falissard, B., Vitaro, F., et al. (2013). Childhood trajectories of inattention, hyperactivity and oppositional behaviors and prediction of substance abuse/dependence: a 15-year longitudinal population-based study. *Mol. Psychiatry* 18:806812. doi: 10.1038/mp.2012.87

Popper, K. R., Condry, J., Orsoni, C., Bosetti, G., and Baudouin, J. (1995). La télévision, Un Danger Pour la Démocratie. *Réseaux* 13, 201–204.

Postman, N. (1986). *Amusing Ourselves to Death: Public Discourse in the Age of Show Business*. New York, NY: Penguin Books.

Rabiner, D., Coie, J. D., and Group, C. P. P. R. (2000). Early attention problems and children's reading achievement: a longitudinal investigation. *J. Am. Acad. Child Adolesc. Psychiatry* 39, 859–867. doi: 10.1097/00004583-200007000-00014

Razel, M. (2001). The complex model of television viewing and educational achievement. *J. Educ. Res.* 94, 371–379. doi: 10.1080/00220670109598774

Rideout, V. (2016). Measuring time spent with media: the Common Sense census of media use by US 8-to 18-year-olds. *J. Child. Media* 10, 138–144. doi: 10.1080/17482798.2016.1129808

Ritchie, D., Price, V., and Roberts, D. F. (1987). Television, reading, and reading achievement a reappraisal. *Commun. Res.* 14, 292–315. doi: 10.1177/009365087014003002

Romano, E., Tremblay, R. E., Farhat, A., and Côté, S. (2006). Development and prediction of hyperactive symptoms from 2 to 7 years in a population-based sample. *Pediatrics* 117, 2101–2110. doi: 10.1542/peds.2005-0651

Ruangdaraganon, N., Chuthapisith, J., Mo-suwan, L., Kriweradechachai, S., Udomsubpayakul, U., and Choprapawon, C. (2009). Television viewing in Thai infants and toddlers: impacts to language development and parental perceptions. *BMC Pediatr.* 9:34. doi: 10.1186/1471-2431-9-34

Ryan, R. M., and Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am. Psychol.* 55, 68–78. doi: 10.1037/0003-066X.55.1.68

Salla, J., Michel, G., Pingault, J. B., Lacourse, E., Paquin, S., Galéra, C., et al. (2016). Childhood trajectories of inattention-hyperactivity and academic achievement at 12 years. *Eur. Child Adolesc. Psychiatry* 25, 1195–1206. doi: 10.1007/s00787-016-0843-4

Schmidt, M. E., Rich, M., Rifas-Shiman, S. L., Oken, E., and Taveras, E. M. (2009). Television viewing in infancy and child cognition at 3 years of age in a US Cohort. *Pediatrics* 123, e370–e375. doi: 10.1542/peds.2008-3221

Schmiedeler, S., Niklas, F., and Schneider, W. (2014). Symptoms of attention-deficit hyperactivity disorder (ADHD) and home learning environment (HLE): findings from a longitudinal study. *Eur. J. Psychol. Educ.* 29, 467–482. doi: 10.1007/s10212-013-0208-z

Schramm, W. (1961). *Television in the Lives of Our Children*. Palo Alto, CA: Stanford University Press.

Sharif, I., Wills, T. A., and Sargent, J. D. (2010). Effect of visual media use on school performance: a prospective study. *J. Adolesc. Health* 46, 52–61. doi: 10.1016/j.jadohealth.2009.05.012

Shin, N. (2004). Exploring pathways from television viewing to academic achievement in school age children. *J. Genet. Psychol.* 165, 367–382. doi: 10.3200/GNTP.165.4.367-382

Sisson, S. B., Shay, C. M., Broyles, S. T., and Leyva, M. (2012). Television-viewing time and dietary quality among US children and adults. *Am. J. Prev. Med.* 43, 196–200. doi: 10.1016/j.amepre.2012.04.016

Stevens, T., Barnard-Brak, L., and To, Y. (2009). Television viewing and symptoms of inattention and hyperactivity across time: the importance of research questions. *J. Early Interv.* 31, 215–226. doi: 10.1177/1053815109338562

Stevens, T., and Mulsow, M. (2006). There is no meaningful relationship between television exposure and symptoms of attention-deficit/hyperactivity disorder. *Pediatrics* 117, 665–672. doi: 10.1542/peds.2005-0863

Strayhorn, J. M., and Weidman, C. S. (1988). A parent practices scale and its relation to parent and child mental health. *J. Am. Acad. Child Adolesc. Psychiatry* 27, 613–618. doi: 10.1097/00004583-198809000-00016

Supper, W., Talbot, D., and Guay, F. (2021). Association entre le temps d'écoute de la télévision et le rendement scolaire des enfants et des adolescents: recension

systématique et méta-analyse des études longitudinales réalisées à ce jour. *Can. J. Behav. Sci*. doi: 10.1037/cbs0000275

Swing, E. L., Gentile, D. A., Anderson, C. A., and Walsh, D. A. (2010). Television and video game exposure and the development of attention problems. *Pediatrics* 126, 214–221. doi: 10.1542/peds.2009-1508

Tétreault, K., and Desrosiers, H. (2014). *Les Facteurs liés à la Réussite à L'épreuve Obligatoire de Mathématique en Sixième Année du Primaire: Un tour D'horizon. Observatoire de la Culture et des Communications*. Québec: Institut de la statistique du Québec.

Tomopoulos, S., Dreyer, B. P., Berkule, S., Fierman, A. H., Brockmeyer, C., and Mendelsohn, A. L. (2010). Infant media exposure and toddler development. *Arch. Pediatr. Adolesc. Med.* 164, 1105–1111. doi: 10.1001/archpediatrics.2010.235

Torppa, M., Niemi, P., Vasalampi, K., Lerkkanen, M. K., Tolvanen, A., and Poikkeus, A. M. (2020). Leisure reading (but not any kind) and reading comprehension support each other—A longitudinal study across grades 1 and 9. *Child Dev.* 91, 876–900. doi: 10.1111/cdev.13241

Verhoeven, M., Junger, M., van Aken, C., Deković, M., and van Aken, M. A. (2010). Parenting and children's externalizing behavior: bidirectionality during toddlerhood. *J. Appl. Dev. Psychol.* 31, 93–105. doi: 10.1016/j.appdev.2009.09.002

Wijndaele, K., De Bourdeaudhuij, I., Godino, J. G., Lynch, B. M., Griffin, S. J., Westgate, K., et al. (2014). Reliability and validity of a domain-specific last 7-d sedentary time questionnaire. *Med. Sci. Sports Exerc.* 46, 1248–1260. doi: 10.1249/MSS.0000000000000214

Winn, M. (1977). *The Plug-in Drug*. New York, NY: Viking Penguin.

Wright, J. C., Huston, A. C., Murphy, K. C., St Peters, M., Piñon, M., Scantlin, R., et al. (2001). The relations of early television viewing to school readiness and vocabulary of children from low-income families: the early window project. *Child Dev.* 72, 1347–1366. doi: 10.1111/1467-8624.t01-1-00352

Zavodny, M. (2006). Does watching television rot your mind? Estimates of the effect on test scores. *Econ. Educ. Rev.* 25, 565–573. doi: 10.1016/j.econedurev.2005.08.003

Zimmerman, F. J., and Christakis, D. A. (2005). Children's television viewing and cognitive outcomes: a longitudinal analysis of national data. *Arch. Pediatr. Adolesc. Med.* 159, 619–625. doi: 10.1001/archpedi.159.7.619

Zimmerman, F. J., and Christakis, D. A. (2007). Associations between content types of early media exposure and subsequent attentional problems. *Pediatrics* 120, 986–992. doi: 10.1542/peds.2006-3322

Zimmerman, F. J., Christakis, D. A., and Meltzoff, A. N. (2007). Associations between media viewing and language development in children under age 2 years. *J. Pediatr.* 151, 364–368. doi: 10.1016/j.jpeds.2007.04.071

Zimmerman, F. J., Gilkerson, J., Richards, J. A., Christakis, D. A., Xu, D., Gray, S., et al. (2009). Teaching by listening: the importance of adult-child conversations to language development. *Pediatrics* 124, 342–349. doi: 10.1542/peds.2008-2267

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership