Interpreting mathematics performance in PISA: taking account of reading performance

Abstract

This study examines the importance of reading performance in explaining mathematics performance in the Programme for International Student Assessment (PISA), and analyses how the relationship is present for different reading subareas. Data of Fangshan District of Beijing in PISA 2009 China Trial were used. Multilevel modelling analyses reveal that: (1) reading performance can explain a considerable proportion of the variance in mathematics performance, and moderates the gender gap favouring males in mathematics performance; (2) specific reading subareas significantly associated with mathematics performance. These findings suggest that taking into consideration students' performance in reading, especially some specific reading subareas, is important when interpreting mathematics performance. Implications for formulating policy based on PISA outcomes are made.

Keywords: PISA; Score interpretation; Mathematics; Reading; Reading subareas; Word problems

1. Introduction

1.1 The relationship between reading and mathematics

The relationships between reading and mathematics performance have been a topic discussed in a number of studies. The positive association between performance in these two domains is widely documented in a range of contexts. For example, Walker et al. (2008) found that on some mathematics problems, students with low reading ability are more likely to give incorrect answers even if they have the similar level of mathematical attainment. Similarly, by employing the Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) 2011 data of Italian 4th grade students, Caponera et al. (2016) suggest that good readers, regardless of their mathematics ability, are advantaged in solving mathematics problems. Chang and Ko (2012) found that mathematics achievement is best explained by reading

comprehension ability for the students with low mathematics achievement comparing with those with medium or high mathematics achievement. The predictive function of reading ability for the progress of mathematics achievement has been evidenced as well. Grimm's (2008) longitudinal study found that students who have better early reading ability are more likely to achieve greater improvement in mathematics.

By contrast, a relatively small body of literature shows rather than early reading ability, early mathematics ability is a stronger predictor of later achievements including that in reading (e.g. Claessens et al. 2009; Duncan et al. 2007). However, it is considered that, in these studies, the mathematics tests may capture language skills with the use of applied mathematics problems, meanwhile, the early reading test may be constructed inconsistently with the later reading test since it only involves letter sounds, word recognition and vocabulary (Claessens et al. 2009; Purpura et al. 2017).

Therefore, item types or measures of the specified mathematics construct employed in the mathematics test may be one of the possible reasons explaining the relationship between reading and mathematics performance. Indeed, it has been consistently reported that the relationship is stronger for mathematics word problems which have relatively high reading demands compared to pure computation problems (Doerr and Temple 2016; Helwig et al. 1999). For mathematics word problem solving, problem comprehension is theoretically and empirically suggested as a crucial component of the cognitive processes (Boonen et al. 2013; Boonen et al. 2016; Fuchs et al. 2016; Mayer 1986). Researchers argue that it is often in the phase of problem comprehension that students make errors on many mathematical problems (Lewis and Mayer 1987; Leiss et al. 2019; Schumacher and Fuchs 2012). It is found that relatively long text in mathematics word problems make poor mathematics achievers tend to be more disadvantaged in solving problems (Mullis et al. 2013; Walkington et al. 2017). It is even assumed that more ‘wordy’ mathematics problems may contribute to the narrowing gender differences favouring males in mathematics performance which are observed from large-scale assessments (Marks 2008).

Besides text length, linguistic- semantic factors may also hamper students' understanding in problems (LeFevre et al. 2010; van der Schoot et al. 2009). From an experiment examining 1st grade children's understanding in solving arithmetic world problems, Cummins et al. (1988) found that children are more likely to miscomprehend the problems which include abstract or ambiguous language. The experiment conducted by Lewis and Mayer (1987) found that for the problems including relational statements, the way the relational term is presented (e.g. 'more than' / 'less than') in the sentence is linked to students' comprehension of the problem. Reading ability is assumed to be helpful for handling linguistic-sematic characteristics and text complexity during the process of problem comprehension (Boonen et al. 2016).

Though strong correlation between reading and mathematics performance has been widely evidenced, it does not necessarily imply that one causes the other. It is suggested that there might be shared cognitive processes (e.g. working memory) or a general ability (e.g. general intelligence) between or above reading and mathematics, contributing to the mathematics performance (Ashkenazi et al. 2017). Reading ability may just act as a proxy for these unknown constructs if the relations between its multiple subareas and mathematics ability are consistent (Grimm 2008); yet, when the relation is only observed for specific reading subareas, shared commonalities between reading and mathematics abilities would be suggested (Grimm 2008; Purpura et al. 2017).

1.2 Reading demands in PISA mathematics problems

PISA, a triennial programme launched by the Organisation for Economic Co-operation and Development (OECD) in 1997, assesses how well students approaching the end of compulsory education are prepared and equipped to meet the challenges in their adult life by measuring students' achievement in reading, mathematics and science literacy (OECD, 1999). Its mathematics tasks typically employ context-embedded word problems in which mathematical objects and symbols are not explicitly presented to students (OECD 2010a). The use of word problems and the need to describe real-world contexts intrinsically brings relatively high reading demands to PISA mathematics test (Eivers 2010). Although the OECD (2010a, 2013a) claims that consideration of the appropriate level of reading required in mathematics problems is taken, it seems that they are still have relatively

high reading demands, at least in terms of the word counts (Wu 2010). The employment of non-continuous texts such as maps and graphs even makes PISA mathematics problems more complex to read (OECD 2010a), since more types of transformations among different representations are needed in problem comprehension (Duval 2006). The Appendix displays mathematics item examples which were used in PISA 2012 and were released by the OECD afterwards.

By regressing country means of PISA mathematics on country means of PISA reading, Wu (2010) found that country mean score in reading is a good predictor of country mean score in mathematics, since they have a very high correlation (r=0.95), and variance in reading scores accounts for 91% of the variance of mathematics scores. She argues that reading demands in items are one of the factors explaining the differential performance between PISA and TIMSS across countries, as many TIMSS items are context-free and have fewer words. By classifying PISA 2012 mathematics problems as 'low reading demand' and 'high reading demand', Ajello et al. (2018) found that Italian male students achieved higher in low-reading demand problems, while females performed better in high-reading demand problems. In the context of China, high correlation between mathematics performance and reading subareas in terms of text formats has been evidenced with PISA 2009 Shanghai data (Shen and Lu 2013). However, research on examining the relationships towards mathematics performance and reading cognitive processes is rarely seen. Moreover, it seems that the extent to which performance in reading and specific reading subareas account for mathematics performance is still under-researched.

Due to the strong overall relationship between mathematics and reading, researchers argue that construct validity in mathematics assessment in PISA can be obscured by reading differences. Rindermann and Baumeister (2015) examined the validity of PISA by rating its tasks on various scales (e.g. reading competence, math competence, problem solving, general knowledge). They suggest that the validity of literacies (e.g. reading, mathematics, science) measured in PISA is questionable, and also that understanding reading literacy is crucial for interpreting performance in PISA tasks (Rindermann and Baumeister 2015).

1.3 Research questions

Since students' performance in PISA has been playing a critical role in influencing educational policy or practice in participating jurisdictions (Breakspear 2012; Ertl 2006; Niemann et al. 2017; Nortvedt, 2018), appropriate interpretation of students' performance is important for informing good policymaking. However, as suggested by Pons (2012), "PISA knowledge for learning", that is, rigorous analysis of PISA data for understanding education quality and for informing policymaking is usually missing. Hence, it is argued that the reception and interpretation of PISA results in the policy field is usually superficial, without the awareness of the complexity underpinning the results (Gruber 2006; Mangez and Hilgers 2012). Although the high correlation between reading and mathematics performance is officially reported in PISA outputs (e.g. OECD 2012), it is only briefly displayed in PISA technical reports, rather than its results reports which are usually used by national policymakers for informing policy making, and the extent to which reading performance can explain the variance in mathematics performance is left unclear. Considering the relatively high reading demands of PISA mathematics problems (Wu 2010), taking students' reading ability into consideration in the interpretation of their mathematics performance is clearly important.

In addition, it is also worthwhile to analyse whether different reading subareas are differentially associated with mathematics ability, from which one can obtain evidence suggesting whether reading ability is only a proxy of some other constructs that actually influence mathematics performance or reading ability itself is directly connected with mathematics ability through common components between these two domains. Previous studies examining the relationship between reading and mathematics most employ a single measure for reading ability (Harlaar et al. 2012), while identifying reading subareas associated with mathematics performance is suggested but not yet commonly conducted (Grimm 2008; van der Schoot et al. 2009). Clarifying these reading subareas could provide further insights into the interpretation of mathematics performance in terms of describing how the relationship between reading and mathematics performance is present.

The current study specifically investigates what differences in reading performance imply for the interpretation of mathematics performance, and whether some cognitive aspects of reading ability or

specific reading text formats have stronger association with mathematics performance than others. Possible effects of gender and family social and economic status background are taken into consideration. Hence, this study focusses on addressing the following two research questions:

(1) To what extent, does students' overall reading performance explain their mathematics performance after controlling for student background factors?

(2) Are there differences in the strengths of the relationships between students' performance in reading subareas and in mathematics?

This paper aims to raise the awareness of policy-makers and other consumers of PISA outcomes with regard to the potential importance of reading in interpreting PISA mathematics performance; and to begin the process of developing a stronger evidence base on the proper interpretation of mathematical outcomes from international large scale assessments.

2. Methods

This section contains two parts. Firstly, data involved in this study are described in terms of the variables used, weights, the approach to addressing missing values, and interactions between variables. Secondly, data analyses methods and procedures are introduced.

2.1 Data

This study uses data of PISA 2009 in which reading was the majority domain, since data of students' performance in reading subareas are available in this cycle in addition to performance in overall reading literacy and mathematics literacy.

Data of Fangshan District of Beijing in PISA 2009 China Trial were employed. China conducted three cycles of PISA China Trial respectively in 2006, 2009, and 2012 with the aim to inform the reforms of domestic educational assessment (Wang 2007, 2009). PISA China Trials were administrated in alignment with PISA technical standards, although the data are not released into public domain (Wang and Jing 2013). Fangshan District of Beijing has been involved in PISA since its participation in PISA 2009 China Trial (Wang et al. 2017). Not only has Fansghan published its results, but also local policymakers explicitly claimed that its local PISA scores and PISA assessment

ideas have been actively used in motivating a number of initiatives for improving teaching and learning in mathematics, science and reading in its local context (Wu 2015). Interpretation and utilisation of Fangshan results in PISA 2009 China Trial was also part of the local-level teacher education content for teachers across the whole local area (Guo et al. 2015). In this case, appropriate interpretation of students' performance in PISA seems especially important considering the high engagement with PISA outcomes in Fangshan local educational practices. Data of PISA China Trials are managed by the National Education Examinations Authority (NEEA) of China and are not yet released into public domain. Approval of using Fangshan data was obtained from the NEEA through a written application.

In each cycle of PISA, a two-stage sample design is typically used, in which schools having 15-year-old students who are enrolled at 7th grade or higher are selected, and then 15-year-olds are selected within the sampled schools (OECD 2012). Jurisdictions also have the option to use a three-stage design in which regions are first selected before sampling schools to obtain accurate estimates of regional results (OECD 2012). PISA China Trials used three-stage design, however, they only targeted academic school students (i.e. students who take the academic track) and do not include vocational schools (Wang and Jing 2013) – this limits the extent to which it makes sense to compare results from these assessments with those from the main PISA outcomes which are intended to include all types of schools having PISA eligible students. According to PISA 2009 sampling technical standards, 35 eligible students were selected from each sampled school, and all eligible students were selected in the case that schools had students fewer than 35 (OECD 2012). During the academic year 2008-2009, there were 52 secondary academic schools in Fangshan (Fangshan Bureau of Statistics 2009), 25 of which were sampled in PISA 2009 China Trial. 610 students from these 25 schools, representing student population in secondary academic schools in this local area, participated in this assessment. Data of them are employed in this study. The summary statistics of the sample are shown in Table 1 below.

Table 1
Summary statistics of the sample.

Total sample (N=610)		
	N	%
Gender		
Female	292	47.9
Male	318	52.1
Educational level		
Lower-secondary	231	37.9
Upper-secondary	379	62.1
	Mean	SD
Age in years	15.73	0.30

2.1.1 Variables

Students' performance in mathematics, reading, reading subareas, and students' gender as well as their family economic, social and cultural status were variables for analyses. We discuss each of these in turn.

Performance in mathematics. The definitions of the typical domains (i.e. mathematical literacy, reading literacy, and scientific literacy) assessed in PISA are usually developed over cycles to echo the changes in the wider field (OECD 2013a). Since PISA 2009 data were used in this current study, definitions employed in this cycle are used through this work. According to PISA 2009 framework (OECD 2010a, p. 14), mathematical literacy is defined as following:

> An individual's capacity to identify and understand the role that mathematics plays in the world, to make well-founded judgements and to use and engage with mathematics in ways that meet the needs of that individual's life as a constructive, concerned and reflective citizen.

Specifically, PISA assesses students' ability "to analyse, reason and communicate mathematical ideas effectively as they pose, formulate, solve and interpret mathematical problems in a variety of situations" (OECD 2010a, p. 105). This is further classified as formulating, employing, and interpreting mathematics in PISA 2012 (OECD 2013a).

In PISA 2009, each student's performance in mathematics is indicated by five plausible values

(PVs) (Mislevy 1991) which are generated by Item Response Theory (IRT) modelling (OECD 2012). The mean and standard deviation (SD) for the average mathematical performance of OECD countries was set as 500 and 100 respectively in PISA 2003 in which mathematics was the major domain for the first time (OECD 2012). On this previously established scale, mathematics performance is reported in follow-up PISA cycles (OECD 2012). The set of five PVs in mathematics were employed as the dependent variable in this study and were dealt with accordingly (see Analyses section).

Overall performance in reading and performance in reading subareas. As the major domain in PISA 2009, reading literacy referred to "an individual's capacity to: understand, use, reflect on and engage with written texts, in order to achieve one's goals, to develop one's knowledge and potential, and to participate in society" (OECD 2010a, p. 14).

Reading literacy has a broader meaning than decoding (i.e. word recognition), as it also contains other cognitive components such as linguistic knowledge (e.g. words, grammar, textual structures and features) and knowledge about the world (OECD 2010a, p. 23).

Reading literacy was classified into three 'aspects' (i.e. cognitive processes): *Access and Retrieve*, *Integrate and Interpret*, and *Reflect and Evaluate* (OECD 2010a). In the tasks of *Access and Retrieve*, students are required to access and locate details from the information explicitly specified in questions, while *Integrate and Interpret* requires students to make sense of meaning from information that is not explicitly stated (OECD 2010a). In the tasks of *Integrate and Interpret*, students have to identify assumptions and implications in a text, and understand the text as coherent by recognising the relationships between pieces of information (OECD 2010a). *Reflect and Evaluate* requires students to connect the information within a text with their own experience and knowledge of the world by providing evidence or arguments, drawing comparisons, or assessing the claims (OECD 2010a).

In addition to the classification based on cognitive processes, this construct was also classified separately as reading *Continuous texts* (e.g. sentences and paragraphs, see item example 1 in Appendix), and reading *Non-continuous texts* (e.g. forms, graphs, figures, etc., see item example 2 in Appendix) based on text formats (OECD 2010a).

PVs in reading and PVs in reading subareas were employed as explanatory variables in this study. As with mathematics, PVs in reading and reading subareas have mean=500 and SD=100 for OECD average on PISA international reading scale which was established in PISA 2000 in which reading was the major domain for the first time (OECD 2012).

Gender. Gender was included in this study to examine the relationship between mathematics and reading after adjusting for possible gender differences in performance in these domains. Traditionally there is a stereotype that females usually outperform males in reading (Ehrtmann and Wolter 2018; Nowicki and Lopata 2017), and in PISA 2009 females achieved higher than males in all participating jurisdictions (OECD 2010b). It is reasonable to assume that gender differences in reading performance may moderate gender differences in mathematics, if reading performance does have a significant effect on mathematics performance. In PISA 2009 student questionnaire database, variable ST04Q01 indicates student gender, with values 1 and 2 refer to female and male respectively. To more conveniently interpret the gender effect, the dummy variable 'male' was created based on this variable, with 0 represents female and 1 represents male.

ESCS. Students' family socioeconomic status and its relationships with their academic achievement have been widely discussed (e.g. Berkowitz et al. 2017; Marks 2006; Sirin 2005; White 1982). The index ESCS (PISA index of economic, social and cultural status) is included in PISA. It is based on three other indices which reflect students' background in terms of home possessions, parents' occupations and parents' educational levels (OECD 2012). ESCS is traditionally used in PISA reports for adjusting for the socioeconomic status of students as well as schools (OECD 2012). It has a mean =0 and SD=1 for OECD countries' average on PISA international scale (OECD 2012).

In this current study, ESCS values were centred on its mean of the student sample involved in this study, so that ESCS=0 representing the average ESCS background of the involved student sample.

Beyond individual student ESCS, it is known that the average ESCS within schools has significant effect on students' performance, and its effect is even stronger than that of student individual ESCS (OECD 2013b; Sirin 2005; White 1982). School-level ESCS is considered as an indicator of

socioeconomic segregation among schools (Perry and McConney 2010). The variable SCH_ESCS representing school ESCS was therefore created and also included in analyses.

2.1.2 Weights

PISA uses sampling weights to address sampling error and allow for making inferences of the population (OECD 2012). To address the measurement error brought by the generation of PVs, PISA adopts the approach of replicating estimation of parameters with replicate weights (OECD 2012). In this study, student final weights (i.e. sampling weights) and all replicate weights were employed when applicable (see Analyses section).

2.1.3 Missing values

ESCS has three missing values in the sample, but all other variables were complete. The three students who had missing values in ESCS were excluded from analyses, making the student sample size 607 across 25 schools.

2.1.4 Interactions between explanatory variables

Interactions between variables (i.e. reading×male, reading×ESCS, reading×SCH_ESCS) were included in modelling but none were found to be statistically significant and so are not reported on further in this paper.

2.2 Analyses

First, descriptive analysis for the above mentioned variables was carried out. Then correlation analysis was conducted to look at the bivariate relationships between mathematics performance and student background measures as well as reading performance, in addition to the inter-correlations between reading subareas. Following that, two-level multilevel modelling (MLM) analyses were conducted with the consideration that students were nested in schools. With the 'bottom-up' modelling strategy (Hox 2010), MLM analyses starts from the null model (M0) which does not include any explanatory variables to investigate the between-school variance in students' mathematics performance. The intraclass correlation coefficient (ICC), indicating the variance explained by schools (Hox 2010), was also calculated. After that, Model 1 (M1) employing student level (i.e. level 1) background variables male, ESCS, and school level (i.e. level 2) variable SCH_ESCS explores to

what extent these variables account for students' mathematics performance. With the control for background variables, in M2, one of the variables of the main interest of this study, that is, PVs in reading were added to investigate the account of reading performance in terms of explaining the variance of mathematics performance.

In M3 and M4, PVs in three reading aspects and PVs in two reading text formats were added respectively to identify whether some reading subareas had stronger effect comparing with other subareas for interpreting mathematics performance. Table 2 outlines all the two-level models.

Table 2
Outline of models.

Explanatory variable		Model				
		M0	M1	M2	M3	M4
Level 1						
Male			√	√	√	√
ESCS			√	√	√	√
Reading				√		
Reading subareas-aspects	• Access and Retrieve				√	
	• Integrate and Interpret				√	
	• Reflect and Evaluate				√	
Reading subareas-text formats	• Continuous Texts					√
	• Non-continuous Texts					√
Level 2						
SCH_ESCS			√	√	√	√

The percentage of between-school variance explained and the percentage of within-school variance explained were calculated respectively for M1, M2, M3, and M4 to investigate the contribution of adding explanatory variables into the modelling.

Although reading subareas scores were all reported on the international scale, the SDs of them for the sample involved in this present study were not the same (see Table 3 in Results section). Therefore, it is problematic to discuss their effect sizes by comparing their coefficients directly (Lorah 2018). Hence, the regression coefficients of reading subareas were then standardised with the method suggested by Snijders and Bosker (2012) to allow the comparability among them.

In conjunction with SPSS 23, IEA IDB Analyzer 4.0 (IEA 2018) which allows for incorporating PVs, sampling weights and replicate weights was used for data descriptive analysis and correlation analysis. The MIXED procedure of Stata 13 (StataCorp 2013a) was used for multilevel modelling (MLM) analyses. In MLM analyses, the full set of five PVs in mathematics, reading and reading subareas were also employed. The normalised student final weights were adopted (OECD 2009). The

MI procedure of Stata 13 was used to pool the results of the five datasets as per the standard methodological guidance (StataCorp 2013b).

3. Results

In this section, first, the descriptive statistics of the variables of interest followed by the correlation between mathematics performance and explanatory variables are presented. Then, multilevel modelling results are presented for the null model in addition to the models involving explanatory variables.

3.1 Descriptive statistics

Descriptive statistics of mathematics achievement as well as all explanatory variables are shown in Table 3.

Table 3
Descriptive statistics of variables.

Variable	Description	Mean (SD)	Mean (SD)		Gender difference (Male-Female)
			Female	Male	
Level 1(n=607)					
ESCS		-0.02 (0.84)			
Male	Dummy variable of ST04Q01 0=female, 1=male	0.55 (0.50)			
Mathematics		536.87 (87.77)	532.81 (89.33)	540.19 (86.31)	7.38
Reading		472.53 (77.97)	487.27 (79.60)	460.47 (74.46)	-26.80*
Reading1	*Access and Retrieve*	477.09 (101.98)	493.11 (105.80)	463.98 (96.72)	-29.12*
Reading2	*Integrate and Interpret*	465.36 (81.08)	477.24 (81.94)	455.64 (79.05)	-21.60*
Reading3	*Reflect and Evaluate*	487.15 (77.04)	503.48 (76.25)	473.78 (75.07)	-29.71*
Reading4	*Continuous texts*	475.47 (81.77)	491.32 (82.70)	462.50 (78.60)	-28.81*
Reading5	*Non-continuous texts*	466.82 (76.28)	479.84 (76.26)	456.17 (74.53)	-23.67*
Level 2 (n=25)					
SCH_ESCS	Student mean ESCS within schools	-0.02 (0.35)			

Note. * Difference is statistically significant at the .05 level

Using OECD average (500) on the international scale as the benchmark[1], Table 3 shows that students' mean performance in mathematics was above the OECD average, while the performance in reading and all the reading subareas lagged behind. Regarding gender difference, though on average males scored 7.38 points higher than females in mathematics achievement, this difference was not statistically significant (p>0.05). However, significant gender differences in favour of girls were observed in the overall reading performance as well as in reading subareas performance. Figure 1 further presents the distribution of males' and females' performance in mathematics and reading.

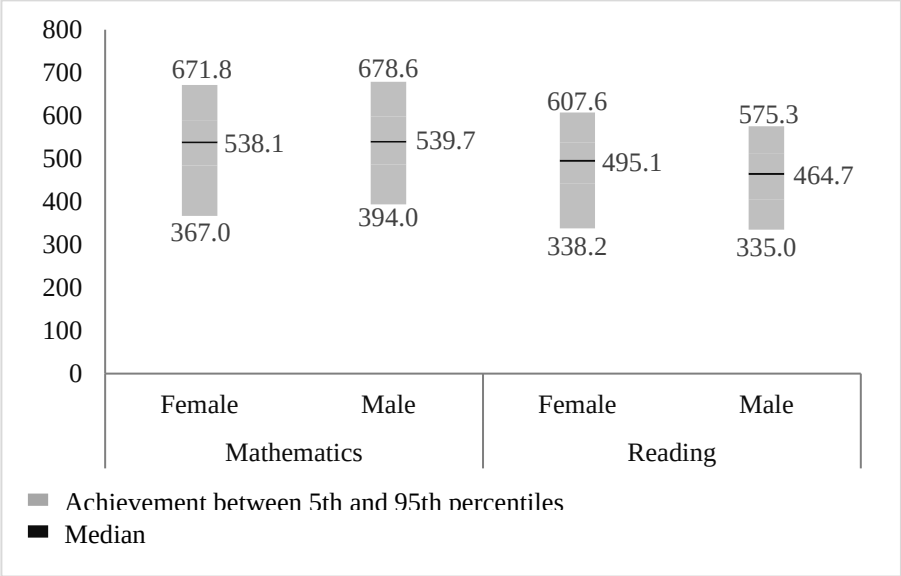


Fig. 1. Students' performance in mathematics and reading, by gender.

As shown in Figure 1, for mathematics, males and females' median performance were similar, however, low male mathematics achievers (i.e. the 5th percentile) scored about 27 points higher than the low female mathematics achievers. Differential distributions between males and females in reading performance are also evidenced in Figure 1. Male' median performance was about 30 points lower than females, and high male reading achievers (i.e. the 95th percentile) scored about 32 points lower than high female reading achievers.

[1] Theoretically, the performance of Fangshan students in PISA 2009 China Trial could not be compared with the OECD average directly, since the target population in PISA 2009 China Trial did not include the students enrolled in vocational education. Here, the OECD average was only used as an indicative reference score.

3.2 Correlation analysis

The bivariate correlation analysis was conducted to identify the relationships between mathematics and explanatory variables of interest. Results are shown in Table 4.

Table 4
Bivariate correlation between students' mathematics performance, background and reading performance.

	ESCS	SCH_ESCS	male	Reading	Reading1 (*Access and Retrieve*)	Reading2 (*Integrate and Interpret*)	Reading3 (*Reflect and Evaluate*)	Reading4 (*Continuous texts*)	Reading5 (*Non-continuous texts*)
Mathematics	0.14*	0.35*	0.04	0.77*	0.61*	0.66*	0.63*	0.67*	0.68*

Note. * Correlation coefficient is statistically significant at the .05 level

As Table 4 presents above, all the explanatory variables of interest except 'male' had significant positive relationships with students' mathematics performance (p<0.05). With regard to the background variables, comparing with SCH_ESCS (r=0.35), ESCS had a relatively weak relationship (r=0.14) with mathematics performance. The relationship between reading and mathematics performance was very high with r=0.77, while the correlation coefficients of reading subareas ranged from 0.61 to 0.68. The inter-correlations between reading subareas were high, with the coefficients ranging from 0.73 to 0.84, as shown in Table 5 below.

Table 5
Inter-correlations between students' performance in reading subareas.

	Reading1 (*Access and Retrieve*)	Reading2 (*Integrate and Interpret*)	Reading3 (*Reflect and Evaluate*)
Reading1 (*Access and Retrieve*)	1.00		
Reading2 (*Integrate and Interpret*)	0.84	1.00	
Reading3 (*Reflect and Evaluate*)	0.73	0.82	1.00

	Reading4 (*Continuous texts*)	Reading5 (*Non-continuous texts*)
Reading4 (*Continuous texts*)	1.00	
Reading5 (*Non-continuous texts*)	0.84	1.00

Note. * Correlation coefficient is statistically significant at the .05 level

The correlation analysis results confirm the importance of employing these explanatory variables in MLM analyses. Though correlation analysis did not demonstrate a significant relationship between gender and mathematics performance, male was still employed in the following MLM analyses

considering the differential distributions of mathematics performance between males and females, and gender difference in reading performance.

3.3 Multilevel modelling

Firstly, a null model, M0, was run to identify the between-school variance and within-school variance of mathematics performance.

Table 6
Analysis of null model.

	M0		
Fixed-effects	**Coefficient (SE)**	***t***	***p>t***
Intercept	538.175 (10.555)	50.99	<0.001
Variance components	**Estimate (SE)**	**[95% Conf. Interval]**	
Between-school variance	2345.902 (581.713)	[1442.033, 3816.317]	
Within-school variance	5128.053 (414.786)	[4371.581, 6015.427]	
ICC			
31%			

As shown in Table 6, $p<0.001$, suggesting that mathematics performance varied significantly among schools. The value of ICC indicates that 31% of variance of Fangshan students' mathematics performance in PISA 2009 China Trial lay between schools. The result of the null model supports the need to take into consideration the multilevel data structure in analyses. The estimates of variance components in the null model were then used as the basis for calculating the variance reduced by the more complex models presented in Table 7.

Table 7
Analysis of multilevel models with explanatory variables.

Model	**M1**		**M2**		**M3**		**M4**	
Fixed effects	**Coefficient (SE)**	***p***	**Coefficient (SE)**	***p***	**Coefficient (SE)**	***p***	**Coefficient (SE)**	***p***
Intercept	531.361 (8.876)	<0.001	126.176 (20.819)	<0.001	177.775 (29.125)	<0.001	164.177 (21.944)	<0.001
Level 1								
ESCS	-0.947 (3.903)	0.809	-5.085 (3.104)	0.103	-4.975 (3.535)	0.165	-4.006 (3.818)	0.300
male	16.739 (6.486)	<0.05	33.564 (4.882)	<0.001	31.247 (5.518)	<0.001	31.708 (7.483)	<0.01
reading			0.833 (0.038)	<0.001				
Reading1					0.074 (.061)	0.253		
Reading2					0.342 (0.104)	<0.01		
Reading3					0.307 (0.084)	<0.01		
Reading4							0.323 (0.065)	<0.001
Reading5							0.436 (0.071)	<0.001
Level 2								
SCH_ESCS	91.684 (22.473)	<0.001	41.755 (11.931)	<0.001	48.235 (15.858)	<0.01	43.523 (12.555)	<0.01
Variance components	**Estimate (SE)**	**[95% Conf. Interval]**	**Estimate (SE)**	**[95% Conf. Interval]**	**Estimate (SE)**	**[95% Conf. Interval]**	**Estimate (SE)**	**[95% Conf. Interval]**
Between-school variance	1404.934 (392.897)	[809.804, 2437.427]	320.191 (177.349)	[97.535, 1051.137]	443.702 (232.222)	[144.7462, 1360.117]	399.645 (220.561)	[117.481, 1359.5]
Within-school variance	5052.545 (414.988)	[4297.056, 5940.862]	2461.436 (242.732)	[1992.056, 3041.414]	3177.129 (300.965)	[2605.818, 3873.697]	3064.199 (244.508)	[2619.659, 3584.174]
% of between-school variance explained	40		86		81		83	
% of within-school variance explained	1		52		38		40	

In M1 which involved only level 1 (student) background variables ESCS and male, and level 2 (school) background variable SCH_ESCS, student ESCS had no significant association with mathematics performance (p>0.05), while both SCH_ESCS and male had (p<0.001and p<0.05 respectively). For females (male =0), an increase of 92 score points (0.92 SD of PISA international scale) in their mathematics performance was associated with a one-unit increase on SCH_ESCS. The significant effect of gender on mathematics performance indicates that for students from the average ESCS background who were enrolled in the schools with the same school average ESCS, males would perform 17 score points (0.17 SD of PISA international scale) higher than females. M1 explained 40% of between-school variance and 1% of within-school variance.

On the basis of M1, in M2 students' reading performance was added. As shown in Table 7, with background variables controlled, reading performance had significant effect on mathematics performance ($\beta_{reading}$= 0.833, p<0.001). For the students from the same socio-economic background, those who achieved 100 score points higher in reading than their peers would achieve 83 higher in mathematics. On the contrary, students who had relatively low reading literacy tended to have relatively poor mathematics performance in PISA 2009 China Trial. Another finding was that the change of the effect sizes of gender and school average ESCS was observed after adding reading performance as an explanatory variable. The coefficient of variable male β_{male} changed from 17 to 34, and it was significant at 0.001 level rather than 0.05 level. With ESCS and SCH_ESCS controlled for, males achieved on average 34 higher than females, supposing that they had same reading performance. On the contrary, the effect size of SCH_ESCS was weakened with β_{SCH_ESCS} turning from 92 to 42, though it was still significant (p<0.001). As Table 7 indicates, comparing with M1, M2 (which took into consideration reading performance) explained 46% more of between-school variance and 51% more of within-school variance.

In M3 which involved the three cognitive aspects of PISA reading, reading 1 denoting *Access and Retrieve* had no significant effect on mathematics (p>0.05). However, the other two aspects reading 2 (*Integrate and Interpret*) and reading 3 (*Reflect and Evaluate*) both had significant relationships with mathematics performance (both had p<0.01). In M4 involving reading 4 (*Continuous texts*) and

reading 5 (*Non-continuous texts*), both of these two variables were significantly associated with mathematics performance with $p<0.001$.

The standardised regression coefficients of reading subareas are shown in Table 8.

Table 8
Standardised regression coefficients of reading subareas.

	M3			M4	
	Reading 1 (*Access and Retrieve*)	Reading 2 (*Integrate and Interpret*)	Reading 3 (*Reflect and Evaluate*)	Reading 4 (*Continuous texts*)	Reading 5 (*Non-continuous texts*)
Standardised regression coefficients	0.086	0.315	0.269	0.301	0.379

Reading 2, representing the cognitive aspect *Integrate and Interpret*, had a marginally larger effect than reading 3 on mathematics performance. In terms of the ability of reading different formats of texts, reading 5 denoting non-continuous texts had a slightly larger effect size than reading 4 on mathematics performance.

4. Discussion and Conclusions

This study investigated and estimated the importance of taking reading performance into account in interpreting mathematics performance, and examined whether particular reading subareas have a stronger association with mathematics performance. In the sections below, key findings and policy implications are discussed, followed by the discussion on limitations and future research.

4.1 The importance of reading performance in interpreting mathematics performance

In line with the findings of previous studies (e.g. Wu 2010), the results of the current study indicate that reading performance had significant effect on mathematics performance. We have found that students with high reading performance were more likely to also achieve high in mathematics ($\beta_{reading}$= 0.833, see M2 in Table 7). As suggested by Helwig et al. (1999) and Doerr and Temple (2016), the relationship between reading and mathematics achievement is relatively high for mathematics word problems. This study has confirmed this finding with data of PISA which employs mathematics word problems with high reading demands. The significant effect of reading

performance and the magnitude of variance explained by including reading as an explanatory variable supports that it is crucial to take reading performance into account in interpreting students' performance (including mathematics performance) in PISA (Rindermann and Baumeister 2015).

With regard to reading subareas, all reading subareas performance except *Access and Retrieve* significantly explained mathematics performance (see M3 and M4 in Table 7). This supports that specific components of reading literacy significantly account for mathematic performance. To add into Grimm's (2008) finding that reading comprehension has differential impacts on different aspects of mathematics, we argue that different aspects of reading could also have varying effects on interpreting mathematics performance. To our knowledge, no studies have had discussion on the differences in the strengths of the effect sizes of performance in reading subareas for interpreting mathematics performance in PISA (Table 8). Regarding the nonsignificant effect of *Access and Retrieve* (M3, Table 7), it may because, for this subarea, students need to locate details from explicitly specified information in the problem (OECD 2010a). This process is not generally required in PISA mathematics problems, in which the underlying mathematics is within the texts (OECD 2010a). The stronger effect size of non-continuous texts comparing with that of continuous texts (Table 8) suggests that higher ability of reading texts such as graphs and tables would benefit students more in terms of their performance in mathematics.

Taking into account reading performance in interpreting mathematics performance also provides insights into understanding the effects of gender and socioeconomic status on mathematics performance. As shown in Table 6, the effect of gender became larger after controlling for students' reading performance, suggesting that to some extent reading performance moderated the gender difference in mathematics performance. Regarding socioeconomic status, the considerably reduced effect size of the average ESCS within schools after reading was added as an explanatory variable suggests that socioeconomic segregation among schools may influence students' mathematics performance in part through the differences in reading performance.

4.2 Implications

This study contributes to a more nuanced understanding of mathematics performance in PISA from the perspective of students reading performance. This perspective could shed light on future studies on the interpretation of mathematics performance in PISA and other assessments in which mathematics items have high reading demands. Specifically, this study extends the existing literature on discussing the relationship between reading and mathematics performance by identifying the significance of reading performance for interpreting mathematics performance in terms of the magnitude of its effect size and the variance in PISA mathematics performance it can explain. Moreover, by identifying that strong association with mathematics performance is only present for specific reading subareas, this study provides initial evidence supporting that, rather than just act as a proxy of other unknown constructs, reading literacy (as defined in PISA) itself may at least have some influence on mathematics performance in PISA. The specific subareas significantly associated with mathematics performance may suggest the commonalities between reading literacy and mathematics literacy which have impact on mathematics performance in PISA.

Hence, rather than adding to the debate on the construct validity issue of PISA (Rindermann and Baumeister 2015), we would argue that it might be the way that PISA reading literacy and mathematics literacy are constructed with some overlaps between each other, which to some extent is implicitly suggested in their definitions (see Variables section). We suggest that for those who would use PISA scores for informing educational policymaking or other educational practices, taking into consideration what PISA means by literacies, key item/problem characteristics in terms of reading demands, and associated reading performance, when interpreting students' mathematics performance is absolutely essential. It is also necessary to identify whether students' weakness lies in those overlaps or the abilities unique to mathematics literacy before taking policy initiatives to improve students' mathematics performance.

Taking into account reading performance also allows for deeper understanding of subgroups' differences in mathematics performance. The analyses of gender differences in mathematics implies that to compare subgroups, for example, gender difference in mathematics performance, using

average scores of the interested domain (e.g. mathematics) as the sole measure may not reveal the full relationships in the data, and caution is required when assessments results are simply interpreted in this way (see Table 3). By simply comparing, say, the mean mathematics performance between males and females, the gender difference was not significant (see Table 3). This is consistent with the previous study which employed Shanghai data in PISA 2009, PISA 2012, and China (B-S-J-G) data in PISA 2015 (Guan 2017). According to Guan (2017), males generally made up a higher proportion high achievers than did females. This current study also finds 'gender distributional imbalance' (Zhou et al. 2017) (see Figure 1). Multilevel regression modelling offers evidence that gender effects favouring males still exist (Zhu et al. 2018). Adding to Zhu et al. (2018), this current study found that gender difference is more notable after adding reading performance into modelling (see M1and M2 as shown in Table 7). Within the context of this study, our findings suggest that to fully develop students' potential in mathematics and reduce gender difference, the reading ability of male students and the mathematics ability of female students merit further attention. In recent year, shrinking gender difference in mathematics has been shown in PISA (OECD 2010b, 2014, 2016) and other assessment such as TIMSS (Martin and Mullis 2013). Therefore, researchers had begun to argue gender similarity (Hyde 2005). However, based on the findings of this study, we would argue that further investigation of the mediation effect of reading ability in gender difference in mathematics performance may be needed before making such claim.

The importance of taking into account reading performance when interpreting mathematics performance within and between jurisdictions also underlines the innate complexity of PISA results, and supports the argument that crudely borrowing policies from high-ranking jurisdictions in PISA is highly problematic (Oates 2011). As one of the most influential international large-scale assessments, PISA has brought the globalized phenomenon of 'policy borrowing' where jurisdictions seek 'best practices' from education systems elsewhere (Kamens 2013). Researchers argue that context matters in translating others' policies (Auld and Morris 2016; Oates 2011). To add to the literature which discuss policy borrowing from the contextual view, this study suggests that, the rich data available in PISA itself, for example, reading performance, should also not be neglected in order to better

understand the education system concerned, and when considering 'borrowing' educational features/policies from high performing jurisdictions.

4.3 Limitations and future research

The design of the current study does not allow for claims of causal inference (Gustafsson 2013) to be made with regard to mathematics and reading performance, although our findings are highly suggestive of a causal link between specific reading subareas and mathematics in the context of PISA. To better establish the causal inference, future research investigating mathematics teachers' perceptions of the influence of students' reading ability in these subareas on their performance on mathematics word problems, or conducting intervention on mathematics teaching by highlighting mathematical reading in these subareas may provide additional such evidence.

It should be noted that this current study only focused on a specific context which is Fangshan District of Beijing, and only secondary academic school students were involved. Hence, the findings of this study may not necessarily hold in other contexts and age groups. However, the wider literature (Shen and Lu 2013; Wu 2010) suggests that the general findings of this study are likely to replicate, to some degree, in other educational settings.

Understanding the mathematics performance by taking into consideration reading performance is still far from telling the whole story of mathematics performance in PISA. As the extent of variance not explained by M2 (shown in Table 7) suggests, other factors that might also explain mathematics performance in PISA might be expected – these might include, for example, affective measures like mathematics self-efficacy and anxiety (Lee 2009), and ability of communicating mathematical solutions or thoughts through writing (Adu-Gyamfiet al 2010). Further studies may investigate these by making use of additional PISA data and/or sources beyond of PISA.

Acknowledgements

We would like to thank Innocent Tasara for his comments on an early version of the manuscript. We would also like to thank the NEEA for providing the access to the data used in this study.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

Adu-Gyamfi, K., Bossé, M. J., & Faulconer, J. (2010). Assessing understanding through reading and writing in mathematics. *International Journal for Mathematics Teaching & Learning*, *11*(5), 1–22.

Ajello, A. M., Caponera, E., & Palmerio, L. (2018). Italian students' results in the PISA mathematics test: does reading competence matter? *European Journal of Psychology of Education*, *33*(3), 505–520. https://doi.org/10.1007/s10212-018-0385-x.

Ashkenazi, S., Rubinsten, O., & De Smedt, B. (2017). Editorial: Associations between reading and mathematics: Genetic, brain Imaging, cognitive and educational perspectives. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00600.

Auld, E., & Morris, P. (2016). PISA, policy and persuasion: translating complex conditions into education 'best practice.' *Comparative Education*, *52*(2), 202–229. https://doi.org/10.1080/03050068.2016.1143278.

Berkowitz, R., Moore, H., Astor, R. A., & Benbenishty, R. (2017). A research synthesis of the associations between socioeconomic background, inequality, school climate, and academic achievement. *Review of Educational Research*, *87*(2), 425–469. https://doi.org/10.3102/0034654316669821.

Boonen, A. J. H., de Koning, B. B., Jolles, J., & van der Schoot, M. (2016). Word problem solving in contemporary math education: A plea for reading comprehension skills training. *Frontiers in Psychology*, *7*, 1–10. https://doi.org/10.3389/fpsyg.2016.00191.

Boonen, A. J. H., van der Schoot, M., van Wesel, F., de Vries, M. H., & Jolles, J. (2013). What underlies successful word problem solving? A path analysis in sixth grade students. *Contemporary Educational Psychology*, *38*(3), 271–279. https://doi.org/10.1016/j.cedpsych.2013.05.001.

Breakspear, S. (2012). The policy impact of PISA: an exploration of the normative effects of international benchmarking in school system performance. *OECD Education Working Papers*, No. 71. http://dx.doi.org/10.1787/5k9fdfqffr28-en.

Caponera, E., Sestito, P., & Russo, P. M. (2016). The influence of reading literacy on mathematics and science achievement. *The Journal of Educational Research*, *109*(2), 197–204. https://doi.org/10.1080/00220671.2014.936998.

Chang C.-Y., & Ko H.-W. (2012). The relationship between mathematics achievement and reading comprehension: TIMSS 2003 and PIRLS 2006 test items as measuring instruments. *Bulletin of Educational Psychology*, *44*(1). Retrieved from http://ojs.lib.ntnu.edu.tw/index.php/bep/article/view/1319.

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review*, *28*(4), 415–427. https://doi.org/10.1016/j.econedurev.2008.09.003.

Cummins, D. D., Kintsch, W., Reusser, K., & Weimer, R. (1988). The role of understanding in solving word problems. *Cognitive Psychology*, *20*(4), 405–438. https://doi.org/10.1016/0010-0285(88)90011-4.

Doerr, H. M., & Temple, C. (2016). "It's a different kind of reading": Two middle-grade teachers' evolving perspectives on reading in mathematics. *Journal of Literacy Research*, *48*(1), 5–38. https://doi.org/10.1177/1086296X16637180.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., … Japel, C. (2007). School readiness and later achievement. *Developmental Psychology*, *43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428.

Duval, R. (2006). A cognitive analysis of problems of comprehension in a learning of mathematics. *Educational Studies in Mathematics*, *61*(1), 103–131. https://doi.org/10.1007/s10649-006-0400-z.

Ehrtmann, L., & Wolter, I. (2018). The impact of students' gender-role orientation on competence development in mathematics and reading in secondary school. *Learning and Individual Differences*, *61*, 256–264. https://doi.org/10.1016/j.lindif.2018.01.004.

Eivers, E. (2010). PISA: Issues in implementation and interpretation. *The Irish Journal of Education / Iris Eireannach an Oideachais*, *38*, 94–118.

Ertl, H. (2006). Educational standards and the changing discourse on education: the reception and consequences of the PISA study in Germany. *Oxford Review of Education*, *32*(5), 619–634. https://doi.org/10.1080/03054980600976320.

Fangshan Bureau of Statistics. (2009). *Fangshan District of Beijing Statistical Yearbook 2009*. Retrieved from http://tjj.bjfsh.gov.cn/tjsj/ndsj/index.htm.

Fuchs, L. S., Gilbert, J. K., Powell, S. R., Cirino, P. T., Fuchs, D., Hamlett, C. L., et al. (2016). The role of cognitive processes, foundational math skill, and calculation accuracy and fluency in word-problem solving versus pre-algebraic knowledge. *Developmental Psychology*, *52*(12), 2085–2098. https://doi.org/10.1037/dev0000227.

Grimm, K. J. (2008). Longitudinal associations between reading and mathematics achievement. *Developmental Neuropsychology*, *33*(3), 410–426. https://doi.org/10.1080/87565640801982486.

Gruber, K. H. (2006). The German 'PISA-shock': some aspects of the extraordinary impact of the OECD's PISA study on the German eduation system. In H. Ertl (Ed.), Cross-national Attraction in Education: Accounts from England and Germany (pp.195–208). Oxford: Symposium Books.

Guan, D. (2017). Woguo zhongxuesheng shuxue chengji de xingbie chayi yanjiu: Jiyu PISA 2009, 2012 and 2015. *Shuxue Tongbao*, *56*(9), 47–51.

Guo, Z., Wu, Y., & Zhang, J. (2015). Jiyu PISA shijian de quyu jiaoyu zhiliang tisheng. *Teachers' Journal (China)*, *5*, 8–12.

Gustafsson, J.-E. (2013). Causal inference in educational effectiveness research: a comparison of three methods to investigate effects of homework on student achievement. *School Effectiveness and School Improvement*, *24*(3), 275–295. https://doi.org/10.1080/09243453.2013.806334.

Harlaar, N., Kovas, Y., Dale, P. S., Petrill, S. A., & Plomin, R. (2012). Mathematics is differentially related to reading comprehension and word decoding: Evidence from a genetically-sensitive design. *Journal of Educational Psychology*, *104*(3). https://doi.org/10.1037/a0027646.

Helwig, R., Rozek-tedesco, M. A., Tindal, G., Heath, B., & Almond, P. J. (1999). Reading as an access to mathematics problem solving on multiple-choice tests for sixth-grade students. *The Journal of Educational Research*, *93*(2), 113–125. https://doi.org/10.1080/00220679909597635.

Hox, J.J. (2010). *Multilevel Analysis* (2nd ed.). New York: Routledge.

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*(6), 581–592.

IEA. 2018. *Help Manual for the IEA IDB Analyzer (Version 4.0)*. Hamburg, Germany. Retrieved from www.iea.nl/data.html

Kamens, D. H. (2013). Globalization and the emergence of an audit culture: PISA and the search for 'best practices' and magic bullets. In H. Meyer & A. Benavot (Eds.), *PISA, Power, and Policy: The Emergence of Global Educational Governance* (pp. 117–139). Oxford: Symposium Books.

Lee, J. (2009). Universals and specifics of math self-concept, math self-efficacy, and math anxiety across 41 PISA 2003 participating countries. *Learning and Individual Differences*, *19*(3), 355–365. https://doi.org/10.1016/j.lindif.2008.10.009.

LeFevre, J.-A., Fast, L., Skwarchuk, S.-L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, *81*(6), 1753–1767. https://doi.org/10.1111/j.1467-8624.2010.01508.x.

Leiss, D., Plath, J., & Schwippert, K. (2019). Language and mathematics- Key factors influencing the comprehension process in reality based tasks. *Mathematical Thinking and Learning*, *21*(2), 131–153. https://doi.org/10.1080/10986065.2019.1570835.

Lewis, A. B., & Mayer, R. E. (1987). Students' miscomprehension of relational statements in arithmetic word problems. *Journal of Educational Psychology*, *79*(4), 363–371.

Lorah, J. (2018). Effect size measures for multilevel models: definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education*, *6*(1), 8. https://doi.org/10.1186/s40536-018-0061-2.

Mangez, E., & Hilgers, M. (2012). The field of knowledge and the policy field in education: PISA and the production of knowledge for policy. *European Educational Research Journal*, *11*(2).189–205. http://dx.doi.org/10.2304/eerj.2012.11.2.189.

Marks, G. N. (2006). Are between- and within-school differences in student performance largely due to socio-economic background? Evidence from 30 countries. *Educational Research*, *48*(1), 21–40. https://doi.org/10.1080/00131880500498396.

Marks, G. N. (2008). Accounting for the gender gaps in student performance in reading and mathematics: evidence from 31 countries. *Oxford Review of Education*, *34*(1), 89–109. https://doi.org/10.1080/03054980701565279.

Mayer, R. E. (1986). Mathematics. In R. F. Dillon & R. J. Sternberg (Eds.), *Cognition and Instruction* (pp. 127–154). London: Academic Press INC.

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

Mullis, I. V., Martin, M. O., & Foy, P. (2013). The impact of reading ability on TIMSS mathematics and science achievement at the fourth grade: An analysis by item reading demands. In M. O. Martin & I. V. S. Mullis (Eds.), *TIMSS and PIRLS 2011: Relationships among reading, mathematics, and science achievement at the fourth grade—Implications for early learning* (pp. 67–108). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Niemann, D., Martens, K., & Teltemann, J. (2017). PISA and its consequences: shaping education policies through international comparisons. *European Journal of Education*, *52*(2), 175–183. https://doi.org/10.1111/ejed.12220.

Nortvedt, G.A. (2018). Policy impact of PISA on mathematics education: the case of Norway. *European Journal of Psychology of Education*, *33*(3), 427–444. https://doi.org/10.1007/s10212-018-0378-9.

Nowicki, E. A., & Lopata, J. (2017). Children's implicit and explicit gender stereotypes about mathematics and reading ability. *Social Psychology of Education*, *20*(2), 329–345. https://doi.org/10.1007/s11218-015-9313-y.

Oates, T. (2011). Could do better: using international comparisons to refine the National Curriculum in England. *The Curriculum Journal*, *22*(2), 121–150. https://doi.org/10.1080/09585176.2011.578908.

OECD. (1999). *Measuring student knowledge and skills: a new framework for assessment*. Paris: OECD Publishing.

OECD. (2009). *PISA Data Analysis Manual: SPSS, Second Edition*. Paris: OECD Publishing. https://doi.org/10.1787/9789264056275-en.

OECD. (2010a). *PISA 2009 Assessment and Framework: Key Competencies in Reading, Mathematics and Science.* Paris: OECD Publishing. https://doi.org/10.1787/19963777.

OECD. (2010b). *PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science (Volume I)*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264091450-en.

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing. https://doi.org/10.1787/9789264167872-en.

OECD. (2013a). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy.* Paris: OECD Publishing. http://doi.org/10.1787/9789264190511-en.

OECD. (2013b). *PISA 2012 Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (Volume II)*. Paris: OECD Publishing. https://doi.org/10.1787/9789264201132-en

OECD. (2014). *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I)*. Paris: OECD Publishing. http://dx.doi.org/10.1787/9789264208780-en.

OECD. (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing. http://doi.org/10.1787/9789264266490-en.

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, *112*(4), 1137–1162.

Pons, X. (2012). Going beyond the 'PISA shock' discourse: an analysis of the cognitive reception of PISA in six European countries, 2001-2008. *European Educational Research Journal*, *11*(2), 206–226. https://doi.org/10.2304/eerj.2012.11.2.206.

Purpura, D. J., Logan, J. A. R., Hassinger-Das, B., & Napoli, A. R. (2017). Why do early mathematics skills predict later reading? The role of mathematical language. *Developmental Psychology*, *53*(9), 1633–1642. https://doi.org/10.1037/dev0000375.

Rindermann, H., & Baumeister, A. E. E. (2015). Validating the interpretations of PISA and TIMSS tasks: A rating study. *International Journal of Testing*, *15*(1), 1–22. https://doi.org/10.1080/15305058.2014.966911.

Schumacher, R. F., & Fuchs, L. S. (2012). Does understanding relational terminology mediate effects of intervention on compare word problems? *Journal of Experimental Child Psychology*, *111*(4), 607-628. https://doi.org/10.1016/j.jecp.2011.12.001.

Shen, X., & Lu, J. (2013). Cong PISA shujukan kuaxueke yuwen suyang de zhongyaoxing. *Shanghai Research on Education, 7*, 5–9+58.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research*, *75*(3), 417–453. https://doi.org/10.3102/00346543075003417.

Snijders, T. A. B., & Bosker. R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks: Sage Publishing.

StataCorp. (2013a). *Stata Statistical Software: Release 13*. College Station, TX: StataCorp LP.

StataCorp. (2013b). *Stata 13 Base Reference Manual*. College Station, TX: Stata Press.

van der Schoot, M., Bakker Arkema, A. H., Horsley, T. M., & van Lieshout, E. C. D. M. (2009). The consistency effect depends on markedness in less successful but not successful problem solvers: An eye movement study in primary school children. *Contemporary Educational Psychology*, *34*(1), 58–66. https://doi.org/10.1016/j.cedpsych.2008.07.002 .

Walker, C. M., Zhang, B., & Surber, J. (2008). Using a multidimensional differential item functioning framework to determine if reading ability affects student performance in mathematics. *Applied Measurement in Education*, *21*(2), 162–181. https://doi.org/10.1080/08957340801926201

Walkington, C., Clinton, V., & Shivraj, P. (2017). How readability factors are differentially associated with performance for students of different backgrounds when solving mathematics word problems. *American Educational Research Journal*. https://doi.org/10.3102/0002831217737028

Wang, L. (2007). Woguo daguimo jiaoyu pingjia xiangmu tanjiu yu shijian. *Educational Science Research*, *11*, 25–28.

Wang, L. (2009). China large-scale education assessment reform: Lessons learned from PISA China Trial research. *China Examinations*, *5*, 17–25.

Wang, L., & Jing, A. (2013). Women cong PISA xuedaole shenme: jiyu PISA zhongguo shice de yanjiu. *Peking University Education Review*, *11* (1), 172–180. http://doi:10.19355/j.cnki.1671-9468.2013.01.013.

Wang, L., Jing, A., & Tong, W. (2017). Transition from PISA China Trial to education evaluation system with Chinese characteristics: based on PISA2009 China Trial. *Educational Research (China)*, *38* (1), 114–123.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin*, *91*(3), 461–481. https://doi.org/10.1037/0033-2909.91.3.461

Wu, M. (2010). *Comparing the Similarities and Differences of PISA 2003 and TIMSS*. https://doi.org/10.1787/5km4psnm13nx-en.

Wu, Y. (2015). Beijing Fangshan: Jiyu PISA ceshi de quyu guanli biange. *Zhongxiaoxue Guanli*, *8*, 13–16.

Zhou, Y., Fan, X., Wei, X., & Tai, R. H. (2017). Gender gap among high achievers in math and implications for STEM Pipeline. *The Asia-Pacific Education Researcher*, *26*(5), 259–269. https://doi.org/10.1007/s40299-017-0346-1.

Zhu, Y., Kaiser, G., & Cai, J. (2018). Gender equity in mathematical achievement: the case of China. *Educational Studies in Mathematics*, *99*(3), 245–260. https://doi.org/10.1007/s10649-018-9846-z.

Appendix: Examples of released PISA mathematics problems

Example 1: PENGUINS

The animal photographer Jean Baptiste went on a year-long expedition and took numerous photos of penguins and their chicks.

He was particularly interested in the growth in the size of different penguin colonies.

Question 1: PENGUINS *(PM921Q01)*

Normally, a penguin couple produces two eggs every year. Usually the chick from the larger of the two eggs is the only one that survives.

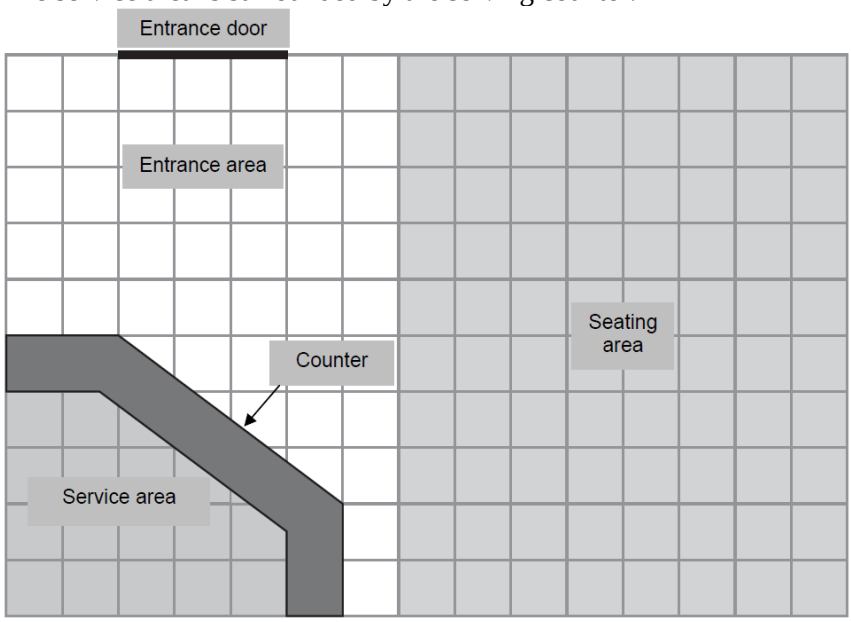
With rockhopper penguins, the first egg weighs approximately 78 g and the second egg weighs approximately 110 g.

By approximately how many percent is the second egg heavier than the first egg?

A. 29%
B. 32%
C. 41%
D. 71%

Example 2: ICE-CREAM SHOP

This is the floor plan for Mari's Ice-cream Shop. She is renovating the shop.
The service area is surrounded by the serving counter.



Note. Each square on the grid represents 0.5 metres × 0.5 metres.

Question 1: ICE-CREAM SHOP (*PM00LQ01)*

Mari wants to put new edging along the outer edge of the counter. What is the total length of edging she needs? Show your work.

.....
.....
.....
.....