ASSOCIATION OF LANGUAGE PROFICIENCY AND PERFORMANCE ON MATHEMATICS AND SCIENCE ASSESSMENTS: THE CASE OF PISA 2015

by

JENILYN A. LEDESMA

B.Sc., The University of Santo Tomas, 1990

M.Phil., The University of Hong Kong, 1997

M.Ed., The University of Western Australia, 2002

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Cross-Faculty Inquiry in Education)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

April 2021

© Jenilyn A. Ledesma, 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Association of language proficiency and performance on mathematics and science assessments: The case of PISA 2015

submitted by Jenilyn A. Ledesma in partial fulfillment of the requirements for

the degree of DOCTOR OF PHILOSOPHY

in Cross-Faculty Inquiry in Education

Examining Committee:

Kadriye Ercikan, Professor Emerita, Educational and Counselling Psychology and Special Education, Faculty of Education, University of British Columbia

Supervisor

Jim Anderson, Language and Literacy Education, University of British Columbia

Supervisory Committee Member

Kim Koh, Werklund School of Education, University of Calgary

Supervisory Committee Member

Cynthia Nicol, Curriculum and Pedagogy, University of British Columbia

University Examiner

Susan Gerofsky, Curriculum and Pedagogy, University of British Columbia

University Examiner

ABSTRACT

The purpose of this study is to examine the degree of association between reading proficiency and performance on mathematics and science assessments for students who speak the test language at home (TLH) and those who speak a different language at home (NTLH), in Canada, Singapore and Indonesia. The study also investigates whether this association varies with gender and socio-economic status (ESCS). The findings reveal significant differences in TLH and NTLH performance scores in all three countries; with the TLH group tending to outperform the NTLH group, except for Mathematics in Canada and Science in Indonesia. When statistical adjustments are made for reading proficiency, the adjusted and unadjusted mean scores for both mathematics and science show different patterns of differences for TLH and NTLH. The NTLH adjusted means are higher in both mathematics and science in all three countries compared to the TLH adjusted means. One explanation could be an underestimation of the scores for NTLH due to their lower reading abilities. Further analyses on the interaction between language group and gender reveal that, in Indonesia, female NTLH scores are higher for both subjects, while male NTLH scores are higher only in science, compared to TLH students. These findings point to disparities in score meaning in mathematics and science assessments, as well as limitations in comparing performances of TLH and NTLH students within the three selected countries. Given that numerous studies on youth comparisons across the globe often use the PISA reading literacy test as its basis, this study adds more evidence to the importance of considering cultural and linguistic diversity of students for appropriate comparisons within and across countries.

LAY SUMMARY

PISA assesses students on reading, mathematics, and science. Compared to students who speak the test language at home (TLH), students who do not speak the test language at home (NTLH) can be disadvantaged because of limitations in language proficiency. This study investigates the association of reading proficiency with performance on mathematics and science assessments of TLH and NTLH. The study also examines whether the association varies for gender and socio-economic groups. The findings indicate that even though TLH tended to have higher performance levels on mathematics and science assessments in almost all comparisons, except for math in Canada and science in Indonesia, when students' language backgrounds are considered patterns of differences on mathematics and science assessments are reversed. The results highlight the necessity of looking at possible impact of language competency on assessment performance, and the importance of comparability when assessment results of students from different language backgrounds are compared.

PREFACE

This dissertation is an original, unpublished, independent work by the author, Jenilyn A. Ledesma, under the guidance of Professor Kadriye Ercikan. The research conducted for this dissertation did not require ethics approval.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AERA	American Education Research Association
ANCOVA	Analysis of Covariance
ANOVA	Analysis of Variance
APA	American Psychological Association
BICS	Basic Interpersonal Communicative Skills
BSNP	National Education Standardization Board
CALP	Cognitive Academic Language Proficiency
CEFR	Common European Framework of Reference for Languages
CMEC	Council of Ministers of Education Canada
CV	Covariate
DIF	Differential Item Functioning
DSA	Direct School Admission
DV	Dependent Variable
ELL	English Language Learner
ESCS	Economic, Social and Cultural Status
ESEA	Elementary and Secondary Education Act
ESSA	Every Student Succeeds Act
GDP	Gross Domestic Product
HEI	Higher Education Institution
ILSA	International Large-Scale Assessments
IRT	Item Response Theory

ITC	International Test Commission
IV	Independent Variable
LEP	Limited English Proficient
MOE	Ministry of Education
MT	Madrasah Tsanawiyah
MTL	Mother Tongue Language
NCLB	No Child Left Behind
NCME	National Council on Measurement in Education
NMAP	National Mathematics Advisory Panel
NTLH	Non-Test Language Speakers at Home
OECD	Organization for Economic Cooperation and Development
PIRLS	Progress in International Reading and Literacy Study
PISA	Programme for International Student Assessment
PSLE	Primary School Leaving Examination
PV	Plausible Values
QCA	Qualifications and Curriculum Authority
SES	Socio-economic Status
SMP	Sekolah Menengah Pertama
STA	Standards and Testing Agency
SVR	Simple View of Reading
TIMSS	Trends in International Mathematics and Science Study
TLH	Test Language Speakers at Home

UN	Ujian Nasional
UN	United Nations
UNESCO	United Nations Educational, Scientific, and Cultural Organization
UNDP	United Nations Development Program

ACKNOWLEDGEMENTS

I would like to express my deep appreciation and gratitude to my supervisor,

Professor Kadriye Ercikan,

for her mentorship, guidance, and encouragement.

I am also thankful to my committee members,

Professor Jim Anderson and Professor Kim Koh,

for their thoughtful and insightful feedback.

And

None of this would have been possible without the love and patience of my children,

Kevin, Isabelle, and Jason.

THANK YOU FOR EVERYTHING.

CHAPTER 1

INTRODUCTION

1.1 Research Background

Increasing cultural and linguistic diversity due to globalization, transnationalism, and internationalism has given rise to questions about the ways in which linguistic and cultural differences are taken into account on international assessment measures for adolescent youth (Asil & Brown, 2016; Hill & Parry, 2014; Smith, Cheema, Kumi-Yeboah, Warrican, & Alleyne, 2018). This global phenomenon is prevalent across both Organization for Economic Cooperation and Development (OECD) and non-OECD countries.

Such challenge demanded the need to address linguistic and cultural diversity in math, science, and literacy assessment. However, the cultures and literacies of OECD countries tend to be privileged on international testing; in fact, dominant languages and language ideologies have rendered the cultural and linguistic qualities of non-OECD youth nonexistent (Brown, 2014; Hill & Parry, 2014; Shohamy, 2013). On the one hand, students considered as “elective multilinguals” (e.g. English as a Foreign Language learner) in many OECD countries (e.g. the United States, Australia) are expected to master second or foreign standard languages (Ellis, 2004; European Commission [EC], 2012), and tend to have their languages reflected in international literacy measures. On the other hand, students residing in non-OECD countries (e.g. Asia, Africa, India, Latin America, the Caribbean) function as “circumstantial multilinguals” (e.g. English as a Second Language learner) and only need to demonstrate proficiency on standard language literacy measures (Ellis, 2004; Wei & Su, 2012), hence their languages are not often taken into account in these measures (Droop & Verhoeven, 2003).

Considering that youth who bring their distinct languages and cultures to these new countries accounts for approximately 38.1 million (14%) of all international migrants (United Nations [UN], 2016, 2017) across the globe, understanding the possible role of societal contexts is important when determining the impact of linguistic and cultural differences on international literacy measures.

1.2 Assessment Policies

Many countries require students from different language backgrounds to be tested in key subject areas such as mathematics, science, and reading (e.g., No Child Left Behind [NCLB], 2002; Qualifications and Curriculum Authority [QCA], 2010) under the same assessment conditions and following standardized procedures. For example, in the United States, the reauthorization of Title 1 of the Elementary and Secondary Education Act (ESEA) of 1965, known as the *No Child Left Behind* Act of 2001 (NCLB, 2002), required increased accountability from schools, primarily measured by standardized content assessments in English. The NCLB mandated that English language learners (ELLs) and limited English proficient (LEP) students be tested in the same content areas and in the same testing procedure (Menken, 2009). The latest reauthorization of the 1965 Elementary and Secondary Education Act, the *Every Student Succeeds Act* (ESSA) (US Government of Education, 2017), also mandated that all students, including English Language Learners (ELLs), graduate high school equip for college or ready for a career. To measure progress and maintain a critical focus on educational equity and excellence for all, ESSA regulated that states must administer annual statewide assessments to all students, including ELLs, in reading/language arts and mathematics in grades 3-8, and once in high school, as well as assessments once in each grade level in science (US Government of Education, 2017). States were also required to conduct annual English language proficiency assessment to all ELLs in grades K – 12, who are held to the same high standard as native speakers but with appropriate accommodations provided when needed.

Similarly, the United Kingdom's *Education Act* of 1997, 2005, and 2011 also required the strengthening of the accountability framework for schools, by producing similar testing procedures and accountability standards where ELLs and native speakers of English are required to exhibit a certain degree of English proficiency (QCA, 2010). The 2019 handbook on the *National Curriculum Assessments Code of Practice* produced by the Standards and Testing Agency (STA) also explained how the tests of the national curriculum are designed, developed, and delivered. The popularity and relevance of the English language has led to endeavours that require students from different language background to demonstrate proficiency in it, regardless of whether they are native speakers or not (Dixon et al., 2012; Grabe & Stoller, 2013; Graddol, 2006).

In Canada, education is the exclusive purview of the provincial/territorial governments. Each province or territory has its own system of education that reflects the geographic, linguistic, and cultural landscape (Council of Ministers of Education Canada [CMEC], 2016). Large-scale assessments of achievement in Canada have many similarities to those in the neighboring United States but differ with respect to their roles and functions (Ercikan, Oliveri, & Sandilands, 2013). As in the United States, most jurisdictions in Canada implement initiatives aimed at assessing the literacy and numeracy skills of students in both French and English programs. Each jurisdiction uses its own proficiency evaluation scales and performance criteria for various grade levels. For instance, British Columbia uses standards defined by the Common European Framework of Reference for Languages (CEFR) to assess student performances, both native and English Language learners alike. Although the context of test administration varies across jurisdictions, educational achievement tests in Canada are primarily low-stakes. Unlike tests used in the United States or in some other countries, results from these tests are not used for high-stakes decisions, such as evaluating schools, teaching quality, or for allocating resources to school districts or jurisdictions (Oliveri, Ercikan, & Simon, 2015). Instead, the tests are used for curriculum planning or program monitoring, and the data from the assessments are used to understand variables associated with learning and achievement (Oliveri, Ercikan, & Simon, 2015).

1.3 Statement of the Problem

Large-scale, international assessments tend to be constructed in the official languages of the country where they are administered, and sometimes do not take into account the special needs of linguistic minority students; students who come from multiple language backgrounds are being tested in the country's official languages (often a single language). Such construct-irrelevant demands create additional linguistic cognitive demands not targeted by the assessment, and possibly add inaccuracies to measurement of what these students know and can accomplish (Abedi, 2001; Abedi & Gandara, 2006; Abedi, Hofstetter, Baker, & Lord, 2001). In order to make appropriate interpretations of educational assessment results, it is necessary to understand the factors that impact the performance of native speakers and language minority students on standardized assessments. Given the increasing prevalence of language diversity and multiculturism across OECD and non-OECD contexts, international assessment measures must consider the undue disadvantage language minority students undergo when they are administered tests in a language that they do not typically use at home (Abedi, Hofstetter, & Lord, 2004; Smith et al., 2018).

In the US context, language minority students have been lagging behind their language majority peers in all content areas, particularly academic subjects that are high in language demand, such as Science and Math (Abedi, 2008; Abedi & Gandara, 2006). The research suggests that several factors account for this achievement gap, including the challenge of second language acquisition (Abedi, Leon, & Mirocha, 2003; Hakuta, Butler, & Witt, 2000), parent education level and socio-economic status (Elosua & De Boeck, 2020; Moore, Glei, Driscoll, Zaxlow, & Redd, 2002), inequitable schooling conditions (Abedi, Herman, Courtney, Leon, & Kao, 2004; Gandara, Rumberger, Maxwell-Jolly, & Callahan, 2003), and measurement tools that are not accurately assessing students' skills and abilities (Abedi & Gandara, 2006; Abedi, Leon, & Mirocha, 2001). Although many of these measurement instruments are designed to measure content areas, such as science and math, they may also have inherent language demands that might affect the performance of ELLs on these assessments (Abedi, 2008; Abedi, Leon, & Mirocha, 2003; August & Hakuta, 1997; Menken, 2008; Solano-Flores, 2006).

Several studies have shown that limited English proficiency has significant implications on students' success in mathematics and science assessments (Abedi, 2004; Abedi & Lord, 2001; Ercikan et al., 2015; Solano-Flores & Trumbull, 2003). According to Martiniello (2008, 2009), ELLs confront substantial challenges because of linguistically- and culturally-dependent content and representations in assessments. For example, in the German PISA 2009 sample, language use at home was seen as having an effect on student performance (Marx et al., 2015). An achievement disadvantage of a quarter standard deviation on the composite reading literacy scale was linked with not speaking the test language at home, even when the students are comparable in terms of social and educational background statistics (Marx et al., 2015; Organization for Economic Cooperation and Development [OECD], 2010a, 2010b).

The language challenges built into test items can impede students' opportunities to demonstrate what they know. Learning mathematics, for instance, is complex and requires students to know and use a variety of knowledge, both language and non-language ways of representing mathematical ideas (Ni Riordain & O'Donoghue, 2009). In Science, while symbolism describes the patterns and/or relationships between the entities, language provides the contextual information about the situation. It is not just a question of manipulating symbols, but also about making meaning as to how different systems interact (Schleppegrell, 2007). Findings from several studies (Abedi & Lord, 2001; Alt, Arizmendi, & Beal, 2014; Martiniello, 2008) reported that the presence of both linguistic and content biases impacted ELL students' performances on word-based math assessment items, hence preventing valid assessments of math skills of students from diverse cultural and linguistic backgrounds. Linguistic and content biases associated with assessments raised the question of whether poor performance by culturally and linguistically diverse students reflect academic difficulty or linguistic difference (Newkirk-Turner & Johnson, 2018). These studies highlighted the need for greater attention to assessment procedures used when measuring performance of school-age children in content areas such as mathematics and science.

Although linguistic abilities of students seem to affect their performance on assessments, other studies suggested that the process may be far more complicated than just their language competencies (Solano-Flores, 2006, 2008). A possibility exists for academic performance to be interrelated with other factors, including academic literacy level, race/ethnicity, socioeconomic status, and school and family support, and thus produces confounding effects (Karoly & Cannon, 2007). Of these interrelated factors, English proficiency level has been proven to be one of the contributing factors that affect student academic performance (Abedi, 2001). For instance, Cocking and Chipman's (1988) study demonstrated that students who speak Spanish as their primary language scored higher on the Spanish version of the test than on the English version of the same test. Abedi (2001) interpreted the results in such a way that, if English proficiency was controlled, the academic gap between the students who typically speak fluent English and/or coming from affluent families and the students who generally have a limited English proficiency and coming from low socio-economic status (SES) families might be different.

Most research done on language minority students has used samples from primarily English-speaking countries, e.g., the United States, Canada, Australia, and the United Kingdom. Hence, the focus has been mostly on English Language Learners, as they are commonly known in these countries. Only very few studies (Fleckenstein, Leucht, Pant, & Koller, 2016; Hopfenbeck et al., 2018) looked at language minority students in countries in which English is not the major academic and assessment language, especially across multiple nations with various primary academic and assessment languages (OECD, 2010a).

To fill this gap, this study investigated if there is differential association of reading ability on performance in math and science assessments of linguistic minority students who do not speak the test language at home (Non-Test Language Speaker at Home or NTLH) and students who speak the test language at home (Test Language Speaker at Home or TLH) in Indonesia. This association was contrasted for those of TLH and NTLH students in Canada, Indonesia, and Singapore. The Programme for International Student Assessment (PISA) 2015 was the data source for the study; student achievement was measured using their mathematics, science and reading scores.

1.4 Researcher Positionality

As an individual who grew up in a multi-lingual family and speaks 6 languages, the author recognizes the complexity of having to master several home languages, and is aware of the roles home language plays in one's academic journey. The author has spent most of her adult life in Asia, and has travelled extensively for work to and have friends from Singapore and Indonesia. She has also lived in Canada for the past 10 years, and has been involved in providing free library-organized language classes to new immigrants as well as leading the Children Ministry at her church for several years. Her multigenerational roots in Asia add credibility to her work; however, she is also aware that pure unbiased observation is not possible. As she sought to present an evidence-based portrayal of student performance in Reading, Mathematics, and Science from the three selected countries, she made explicit the assumptions and biases that she would bring to the research. She also addressed potential subjectivities in her research, for example, how her passion for language learning might have shaped her approach to conducting this research.

1.5 Country Selection for the Study

Prior to the main research, a trial investigation was conducted to select countries for the study. Since approximately 5,000-25,000 15-year-old students from at least 150 schools per country did the PISA survey, from the 72 participating countries/economies, those with *final* valid sample sizes of less than 5,000 were excluded. This step yields a total of 60 remaining countries. And to ensure that the study included a good mix of students who speak the test language at home and those who do not, only countries that have at least 20% of students coming from one of the two language groups were selected, i.e. 80% test language speakers at home and 20% non-test language speakers at home or vice versa. This led to a final number of 9 countries, namely Switzerland, Cyprus, Canada, United States, Qatar, United Arab Emirates, Indonesia, Singapore, and Malaysia.

In the last step of the trial investigation, the number of students who did both Reading and Math (Reading + Math group) and both Reading and Science (Reading + Science group) were obtained. Only countries with a valid sample size of at least 200 students or more per group were included; the remaining countries were Canada, Indonesia, Singapore, and United States. Although Canada and United States have striking similarities in terms of language, economy, population size, and education system, Canada has been performing well on the PISA surveys. Hence, Canada was selected as an example of a high-quality, Western education system that is comparable to some of the high performing East Asian countries such as Singapore. And given the positionality of the researcher in terms of her personal connections to and experiences with Singapore and Indonesia, both countries were selected, forming the sample for the present study.

1.6 Research Aims, Objectives and Questions

The study aimed to explore the impact of reading ability of students on their math and science performance, based on the PISA 2015 data derived from the following three countries: Canada, Indonesia, and Singapore. These countries were selected for the following reasons.

Canada is a country with significant numbers of students from immigrant backgrounds; these students face challenges different from students whose home languages are one of the two official languages – English and French. Canada was selected because English is one of the official languages used in the country, and most test-takers who did the English version of the PISA 2015 assessment were TLH. Students who took the French version of the test were excluded in the present study. Indonesia was also chosen because although Bahasa Indonesian is the official language used in the country, more than half of the students who completed the Bahasa Indonesian version of the PISA 2015 assessment did not speak the test language at home; they were considered as NTLH because their home language is not the same as the language of assessment. Singapore was also selected because there was an equal proportion approximately of TLH and NTLH who did the English version of the PISA 2015 assessment. Likewise, education in Singapore is bilingual with English as one of the main media of instruction, and therefore all these students can logically be expected to have good command of the test language.

Given the well-documented performance differences between gender and socio-economic groups (Baker, Goesling, & Letendre, 2002; Chiu & McBride-Chang, 2010; Ercikan, McCreith, & Lapointe, 2005a, 2005b; OECD, 2010a, 2010b), the study also examined if the relationship between reading ability and mathematics and science performance varies for gender and socio-economic groups.

More specifically, the study investigated the following research questions:

RQ1: To what degree do students' reading abilities (PISA Reading Score) account for their math performance (PISA Math Score)?

RQ2: To what degree do students' reading abilities (PISA Reading Score) account for their science performance (PISA Science Score)?

RQ3: To what degree does the relationship between reading ability and mathematics and science performance vary for gender and socio-economic groups?

1.7 Significance of the Study

The study is significant for two reasons. Ercikan and associates (2015) demonstrated that ELLs in the United States confront substantial challenges with mathematics and science assessments because of linguistically and culturally dependent content and representations in assessments. There is also growing evidence that limited language proficiency has significant implications on students' success in mathematics and science assessment (Abedi, 2004; Abedi et al., 2004; Abedi & Herman, 2010; Abedi & Linquanti, 2012; Noble, Risebery, Suarez, Warren, & O'Connor, 2014; Solano-Flores & Trumbull, 2003). Hence, investigating whether reading ability has a moderating effect on math and science performances of students from different language backgrounds in the three selected countries can offer explanations on, and contribute information to, the comparability and consistency of score meaning for these student groups and is central to the validity of score interpretations. Likewise, using variables like gender and socio-economic status may also reveal whether the language effects vary for these groups.

1.8 Limitations of the Study

As with most academic research, this study has its limitations. First, although the study aims to explore the impact of the language ability of the students on their performance in the PISA survey, only the reading ability was examined. A comprehensive testing of the multi-dimensionality of language competence (Jang & Roussos, 2007), i.e. listening, writing, or speaking, was not included in this study. One reason is that reading has been considered a major prerequisite for educational success (Fleckenstein et al., 2016), and therefore can be regarded as the most relevant language modality in large-scale assessments. Advanced level of reading comprehension has also been reported to facilitate the acquisition of knowledge and new ideas (Chall, 1996). Another reason is data availability. Reading ability is the only measure PISA used to indicate reading proficiency of students.

1.9 Dissertation Organization

To address the research questions, this dissertation is organized as follows. The first chapter introduces the research background, statement of the problem, researcher positionality, research aims and objectives, research questions, as well as the significance and limitations of the study. Chapter Two reviews relevant literature in the field of large-scale assessment, multiculturalism and multilinguality, and differential impact of language ability on math and science performance; special attention is given to students who speak the test language at home, and students who do not speak the test language at home. In Chapter Three, the detailed research design, such as sampling method of PISA, participants, instrumentation, measures, and data collection methods are described. The data analysis plan is presented as well. Chapter Four reports the results and interpretations of research questions. Discussion and conclusion based on the findings, as well as implications and recommendations for future research are reported in Chapter Five.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

International large-scale assessments (ILSA) have become a truly international phenomenon (Broadfoot, 2007), and have been on the rise in the past two decades (Fischman & Topper, 2017). In fact, every country in the world has its own system of formal examination and educational assessment. Examples of these initiatives include not only the goal-driven approach of the United Nations' *Sustainable Development Goals* (UN, 2017) and the test-driven comparisons of specific aspects of education such as literacy (UNESCO, 2006), but also the large-scale international surveys such as the Progress in International Reading and Literacy Study (PIRLS, for primary schooling), the Trends in International Mathematics and Science Study (TIMSS, covering both primary and secondary schooling), and the Programme for International Student Assessment (PISA, for secondary schooling). All these international, comparative testing and surveys of representative samples of pupils from both low-income and high-income nations are prominent features in educational policy evaluation (Wyse & Torgerson, 2017). However, the research designs used in these international surveys are intended to establish *correlations* between education policies and outcomes, and not necessarily to establish whether such policies *cause* the observed outcomes.

2.2 Programme for International Student Assessment (PISA)

The Programme for International Student Assessment (PISA) is an international, large-scale assessment carried out by the Organization for Economic Cooperation and Development (OECD) in member and non-member countries/economies. It is a triennial study that measures the performance of 15-year-old students in three domains: mathematics, reading, and science. Since its launch in 2000, PISA has been repeated every three years, with the most recent being PISA 2018. One of the assessed areas in each assessment cycle is usually given greater prominence and is chosen as the major domain, while the other two areas are assessed less thoroughly. The domains in focus in each cycle were reading in 2000, 2009 and 2018, mathematics in 2003 and 2012, and science in 2006 and 2015. PISA does not combine the three domain scores into an overall score. They each have their own performance scores. In addition to the three domains of reading, mathematics, and science, financial literacy and problem-solving have been added in some of the cycles, as well as the new domain of global competency introduced in 2018. Unlike other ILSA, which has a curriculum approach, the PISA study is not directly linked to a school curriculum. Rather, it emphasizes "how well 15-year-old students are able to *apply* the knowledge and skills they have learned by the time they finish their compulsory schooling to real-life situations" (OECD, 2014a, 2014b). The test content of PISA is therefore independent of the participating countries' school curricula (OECD, 2014a, 2014b).

The aim of PISA is to provide countries with comparable data aimed at improving their education policies and outcomes. Only students being educated at school are tested. Each country is required to draw a sample of at least 5,000 students. In smaller countries, an entire age cohort may be tested. Each student sits a two-hour test, part of which is multiple-choice and part of which requires short answers. The test items comprise both cognitive testing, as well as questions on students' background, such as their learning habits, motivation, and home/family characteristics. There are many assessment materials available, so any given student is not tested on all items. Because the students worked on different test materials, raw scores in each subject domain are scaled to an OECD average to enable meaningful comparisons. The overall achievement distribution is a mean of 500 and a standard deviation of 100. Achievement data from the latest test cycles are linked to the previous cycles using item response theory (IRT) scale-linking methods. Proficiency estimates are also generated using a latent regression extension of the Rasch model under IRT to provide “plausible values” which enable unbiased estimates of between-group differences.

2.3 Education System and Assessment in Canada

2.3.1 Education System

Canada is the world’s second largest country by total area (Statistics Canada, 2011), consisting of 13 jurisdictions (10 provinces and 3 territories). Canada is also one of the most educated countries in the world (Statistics Canada, 2018). According to Statistics Canada (2018), in 2016, 91 percent of people in Canada aged 25 to 64 had at least completed high school, and 66 percent had obtained a post-secondary educational credential. These figures are both above the OECD averages of 78 percent and 40 percent, respectively. Canada also has a strong and well-funded decentralized public education system. There is neither a ministry of education nor a federal education department; in fact, the country’s education system is largely managed provincially and can receive funding from the local, provincial, and federal governments (Klinger, De Luca, & Miller, 2008; Klinger, Maggi, & D’Angiulli, 2011). Consequently, since each jurisdiction/province oversees its own education system, the education system can vary between provinces. However, the standard of education still remains consistently high in Canada.

There is both a public and private education system in Canada. Education in Canada is heavily subsidized by the provinces, from kindergarten through to the post-secondary level, spending on average almost six percent of its GDP on education (OECD, 2020). This means Canada spends proportionately more on education than the average 4.9% among OECD countries (OECD, 2020). Access to education in Canada is available to children the year they turn five, except in Ontario and Quebec where children may start a year earlier. Kindergarten may also be optional depending on the province. Homeschooling is considered legal across Canada.

The education system in Canada consists of pre-primary or kindergarten, primary, secondary, and post-secondary schooling. Compulsory school attendance starts from 5 until the age of 16 in all provinces, except for Manitoba, Ontario, and New Brunswick where the required age is 18. Pre-primary or kindergarten is the first stage of education in Canada and are offered to children between the ages of four to five. This stage is mandatory in New Brunswick and Nova Scotia but is optional everywhere else in Canada. The next level is primary or elementary education. This level runs from grade 1, generally at the ages of 6 or 7, and goes until grade 7 (ages 11 to 12). The next level is secondary education or high school, which runs from grade 8 (ages 13-14) to grade 12 (ages 17-18), followed by post-secondary education. The school year normally runs from September until June. Canada has a wide selection of colleges and universities, located in both urban and rural regions of the country. Post-secondary schooling is split between colleges (usually for vocational and technical training) and degree-awarding universities (e.g. a bachelor's, master's, or doctorate degree).

Canada has two official languages: English and French. English is the main language in most provinces. However, French is also widely spoken throughout the country. Both languages have official status at the federal level, and the federal government provides financial resources to fund educational programs for each one. Both French or English as a second language is generally taught from an early age.

2.3.2 Assessment in Canada

Canada has a decentralized education system. Each jurisdiction has control over its own educational system, curricula, educational policies, and provinces can decide which assessments to administer and to which students (Klinger, De Luca, & Miller, 2008). Jurisdictions typically administer tests in key subject areas like mathematics, reading, and science. Currently, every province / territory (except Nunavut) has at least one large-scale assessment program (Klinger, De Luca, & Miller, 2008; Kocay, 2019). These assessment programs tend to be similar in form and function across the jurisdictions (Klinger, De Luca, & Miller, 2008). They contain both multiple-choice and constructed-response items. Most are criterion-referenced and are often developed by Canadian teachers under the supervision and guidance of provincial/territorial assessment offices (Ercikan, Oliveri, & Sandilands, 2013).

It is also common practice to conduct early student testing (e.g., Grade 4 Foundation Skills Assessments in BC) without any direct associations with students' grades, followed by secondary examinations that have relatively major impact on students' grades and graduation requirements (e.g., Graduation Numeracy and Literacy Assessments in BC). Assessment results comparing minority versus majority linguistic populations are often publicly available. It is worth noting that achievement may vary for different student groups, such as gender, language, or ethnicity groups. In one study, significant differences have been found when comparing students in Francophone minority and Francophone majority settings (Ercikan, McCreith, & Lapointe, 2005a, 2005b). Within this context, investigating comparability across language groups is therefore important.

2.4 Education System and Assessment in Indonesia

2.4.1 Education System

Indonesia has the fourth largest population in the world. Over the past few decades, Indonesia has made great success in improving access to education by having students attend school earlier and staying longer. However, the country's education system has been criticized for being a high-volume but low-quality enterprise that has fallen short of the internationally competitive system (United Nations Development Program [UNDP], 2014). In fact, Indonesia's education system has been less successful in ensuring that the students receive an education. Based on UNESCO's 2012 statistics (UNDP, 2014), 42 per cent of Indonesian 15-year-olds who did the PISA 2015 assessment failed to meet the minimum standards in *all* three areas: reading, mathematics, and science (Hidayat & Patras, 2013; Sukyadi & Mardiani, 2011). The reading interest index in Indonesia is only 0.001, which meant that for every 1,000 people, only one person has an interest in reading, a situation called as '*tragedi nol buku*' or 'tragedy of zero books' (Dharma, 2014).

Indonesia's education system has four levels of education: primary (grades 1-6), junior secondary (grades 7-9), senior secondary (grades 10-12), and higher education. The first two levels constitute 'basic education'. The country's primary and secondary education sector (covering grades 1 to 12) has more than 266,000 schools, where 45 million students are taught by 2.7 million teachers (Kurniawati, Suryadarma, Bima, & Yusrina, 2018). The state educational system is mostly non-sectarian, but it also includes some religious (typically Islamic) schools and higher education institutions (HEIs). Approximately 85 per cent of the students are enrolled in regular schools, which could be public, private non-religious, and private religious. The rest are in Islamic, largely private schools called *madrasahs*. The primary education (87%) and junior secondary education (75%) levels are overwhelmingly public, whereas the proportion is more balanced at the senior secondary level (58% public and 52% private) (Kurniawati et al., 2018). Generally, state educational institutions are considered of higher quality than private educational institutions, although there is great variation among both public and private institutions (Rosser, 2018).

Under the decades-long rule of President Suharto, the education system in Indonesia was highly centralized, with the Ministry of Education and Culture primarily responsible for managing all levels of the education system, and the Ministry of Religious Affairs responsible for funding state Islamic schools and HEIs and regulating matters related to religious education. However, since the end of the Suharto regime in 2001 and the subsequent democratization of Indonesia, responsibility for managing the education system has changed significantly. Over time, various functions of the government have been decentralized and authority was transferred over to locally elected district-level governments. This shift, however, did not extend to higher education. In early 2000s, legislations not only introduced free and mandatory basic education, it also gave schools greater administrative authority. The government made a commitment to spend 20 percent of the national budget on education; the country spent 3.6 percent of its GDP on education in 2015 (Dilas, Mackie, Huang, & Trines, 2019).

Indonesia is a highly multicultural and multilingual society; Bahasa Indonesia is the national language of the nation. It is widely spoken by majority of the Indonesians in 'high' or public functions as a means of cementing cultural identities (Lauder, 2008). Although English is not widely used in society, i.e. not used as a medium of communication in official domains such as the government and the education sector, it is still seen as the most important 'foreign' language to be taught in Indonesia (Yusny, 2013). However, despite being considered essential, English has never been recognized as an official 'second' language.

2.4.2 Assessment in Indonesia

Being the fourth most populous country in the world, Indonesia's education system is large (Kurniawati et al., 2018). Educational quality is a pressing concern in Indonesia. Promotion and graduation are largely based on external national examinations, *ujian nasional*, or UN, typically in a multiple-choice format at the end of each school year, and, since 2015, the introduction of continual assessment and school-based examinations (Dilas, Mackie, Huang, & Trines, 2019). Assessment of learning achievement is divided into two types. The first one is *ulangan* or formative assessment, in which there is an endeavor to continuously measure students' achievement in their learning and obtain relevant information for their learning development. The other one is *ijian*, a summative testing that is conducted to measure the required learning competencies achieved by the students.

All formative and summative assessments are conducted by teachers. Schools conduct the *ujian sekolah* (school tests), while the government conducts the national exams, administered by the National Education Standardization Board (BSNP). at schools. At the end of basic education in grade 9, students from public schools receive the certificate of completion, *sekolah menengah pertama,* or SMP, while students from private and religious schools are awarded the certificate of completion *madrasah tsanawiyah,* or MT.

Senior secondary education lasts for three years in grades 10 to 12; it is neither compulsory nor free. At the end of grade 12, students sit for school examination as well as final external UN examination. Upon successful completion of the program, students from regular schools receive a certificate of completion *ljazah* SMA, or *ljazah* MA for Islamic schools. Results in high-stake national exams determine students' acceptability in higher level of education.

2.5 Education System and Assessment in Singapore

2.5.1 Education System

Located at the southern tip of the Malaysian peninsula, Singapore, a small island nation, was once a British colony until gaining independence in 1965. Since independence, Singapore has capitalized on its strategic location and entrepreneurial approach to become financially and politically successful (Ng, 2008). The educational production of human capital has become a priority (Koh & Paris, 2011). Education in Singapore is centralized and managed by the Ministry of Education (MOE) (MOE, 2014). MOE controls the administration and development of state schools; it also has an advisory and supervisory role in private schools. The government spends about 20% of the national budget on education, and subsidizes public state education and government-assisted private education (MOE, Singapore, 2016a). Overall, Singapore's success in education has been influenced by the integrated planning system of the

Curriculum 2015 framework. The orientation of Curriculum 2015 aimed for every student to become: a confident person who thinks independently and critically; someone who communicates effectively; a self-directed learner; an active contributor, with innovation and initiative; and a concerned citizen who is informed about the world and local affairs (OECD, 2012a, 2012b). The close links between policy makers, researchers and educators, and commitment to equity and merit have become the key to Singapore's success (Suprapto, 2016; Tan, Koh & Choy, 2016).

Singapore's six-year compulsory primary schooling requires all children to start at age 7, giving them a strong educational foundation. Primary 6 students take the Primary School Leaving Examination (PSLE) to assess their suitability for secondary education; the test streams and places them in a secondary school course based on their learning pace, ability, and inclinations (OECD, 2016a). Alternatively, students can also seek admission to a secondary school based on their diverse strengths and interests through the Direct School Admission (DSA) exercise (MOE, Singapore, 2016a), where participating schools are given greater flexibility in selecting students. At the lower secondary levels, grades 7 and 8 students experience a broad-based education in the languages, the humanities and the arts, mathematics and sciences, design and technology, physical education as well as character and citizenship education. Then they progress to upper secondary levels, grades 9 and 10, where students learn two new languages, social studies and mathematics, and a wide range of elective subjects and programs (MOE Singapore, 2016b).

Singapore is also a multi-ethnic and multilingual society; its language policy has been described as a need to ensure the cohesion of its multi-ethnic fabric (e.g. Gupta, 1998; Wee, 2003). Kuo's (1980) account of the languages spoken within and across the ethnic groups showed Singapore to be linguistically diverse. Singapore's language policy initially stated that the four official languages – Mandarin Chinese, English, Malay, and Tamil – are used as the media of instruction. However, in the past 20 to 30 years, this has slowly shifted to English increasingly becoming the mother tongue for most Singaporeans, especially with the younger generations. What was once a linguistically heterogeneous group has now evolved into one with English as an inter-group and intra-group lingua franca amongst Singaporeans, replacing all other languages as a supra-ethnic language (Cavallaro & Ng, 2014). As a result, their ethnic languages became more like a second language. Although the Singaporeans are English dominant, they are still able to converse in at least two languages. Clearly, Singapore has moved from 'English knowing' to 'English dominant' (Cavallaro & Ng, 2014). In fact, in most international measures of literacy and numeracy skills, Singaporean children are amongst the most competent in the world (Goh & Gopinathan, 2008, Tan, 2015).

2.5.2 Assessment in Singapore

Teachers in Singapore perform continuous assessment of their students at all levels of education (MOE Singapore, 2016b). Students sit for national examinations at the end of primary, secondary, and post-secondary education. These exams serve as gateways to lower secondary, higher secondary, and tertiary education. Assessment in Singapore is high-stake, and has a relatively major impact in determining a student's path to higher education (Chee et al., 2016; Lee, Goh & Birger, 2008).

Schools place students into separate exam levels based on the subjects they took in year five and six of primary school. At the end of primary school (age 12), they take the Primary

School Leaving Examination (PSLE) and their exam scores will help determine which band in lower secondary education as well as which school they will attend. In short, PSLE ranking assesses suitability for secondary school and sorts students into the appropriate secondary school course of study. The Ministry allows some schools to practice Direct School Admission, admitting students based on other achievements before the PSLE results are released, to provide greater diversity in student talents and interests (MOE Singapore, 2016b). The Ministry also helps place students who are not accepted into their schools of choice.

Post-secondary education following four years of lower secondary school begins at age 16. All secondary schools in Singapore adopt the academic grading system of the Singapore-Cambridge GCE O-Level examination, which students sit at the end of four or five years of secondary education, taking at least 6 subjects or more. Students are admitted to post-secondary schools based on their Cambridge GCE "O" level exam results, or "N" level results if they are in the normal band. At the end of secondary education, students with the requisite exam results can attend universities, while those who do not have the required results are expected to study for two more years at a pre-university (or polytechnic) and pass their A levels before entering university in Singapore. The GCE "O", "N", and "A" levels determine which type of post-secondary education a student may attend (Poon et al., 2016).

2.6 Multiculturalism and Multilinguality

The concept of a "multicultural" society can be interpreted as a sociological fact, an ideology, or as a policy (Brosseau & Dewing, 2009). As a sociological fact, multiculturalism refers to the presence of people from diverse racial, indigenous, and ethnic backgrounds.

Ideologically, it consists of a coherent set of ideas and ideals on cultural diversity. At the policy level, it refers to the management of diversity through formal initiatives and policies.

"Multilinguality", or linguistic diversity, is an integral part of cultural diversity. It allows the transmission and sharing of knowledge from one culture to another, and from one generation to another. It also has a crucial role in individual and group identities, as it connects and reconcile the ideas and values of people from diverse cultural, social, economic, and ethnic backgrounds. It is a key contributor to dialogues, individual empowerment, and social transformations (UNESCO, 2015). With the rapid globalization, UNESCO (2015) is convinced that multiculturalism and multilinguality have key roles to play in fostering pluralistic, equitable and inclusive knowledge societies. In fact, the organization encourages its member states to develop comprehensive language-related policies, allocate resources, and use appropriate tools to promote and facilitate linguistic diversity and multilingualism.

International comparisons of student performance are becoming a new indicator of the quality of a nation's educational system (Lingard, 2010). Large-scale international assessments are increasingly being used as a way to evaluate and compare the quality of the future labor force across different countries (Schwippert & Lenkeit, 2012). However, these assessments tend to ignore the fact that theory, policy and practice may interface with the cultures of different host societies (Schwippert, 2007). These assessments reinforce the underlying premise that the world of education is not culturally diverse (Benoliel & Berkovich, 2017). It elicits an assumption that global variation in students' academic performance is attributable to national educational structures and policies (Feniger & Lefstein, 2014). Thus, policy makers in countries with contrasting cultures have tended to follow policy ‘blueprints’ adopted by others, with little consideration of its cultural fit.

Cultural context matters when it comes to educational achievement (Benoliel & Berkovich, 2018). Societies differ greatly in their educational philosophy and practices, and cultural orientations can underlie their societal structural arrangements. For example, some education systems express more hierarchical and conservative cultural orientations, such as Singapore, while other education systems express a more democratic and autonomous cultural dimension, such as Canada. All these culturally-shaped beliefs and forms of interactions may help explain the differences in the achievement levels of different countries in international assessments, emphasizing further the importance of considering the societal and cultural milieu of a country in international assessments of student performance (Bourdieu, 2011).

2.7 Effect of Language on Assessment

Linguistic complexity is defined by Schleppegrell (2004) as the amount of discourse (oral or written), the variety of grammatical structures, the organization and cohesion of ideas, and the use of text structures in specific genres. Hence, test items that consist of complicated sentences can potentially contribute to misunderstanding for some students (Abedi & Lord, 2001), making them unable to show their ability. In their studies on the impact of language characteristics in mathematics test items, Shaftel and associates (Shaftel, Belton-Kocher, Glasnapp, & Poggio, 2006) claimed that removing linguistic complexity in exam items have shown moderate increases in ELL scores compared to the original. They suspected that linguistic complexity led to item bias.

As the *Standards for Educational and Psychological Testing* stated, “in testing applications where the level of linguistic or reading proficiency is not part of the construct of interest, the linguistic or reading demands of the test should be kept to the minimum necessary for the valid assessment of the intended construct” (American Education Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 2014). Reducing linguistic complexity on test items has been strongly encouraged by researchers as a method to increase the validity of test scores (Abedi, 2004, 2008; Kopriva, 2000).

OECD’s definition of *literacy* deviates from traditional notions that emphasize “mastery of specific school [reading] curricula” (OECD, 2014a, para.3). The definition is largely consistent with current notions of literacy as dynamic, social, cultural, and multi-faceted (Cope & Kalantzis, 2012), which challenges the notion of literacy as a set of foundational or constrained skills (Paris, 2005). PISA operationalizes *literacy* as a “student’s ability to apply knowledge and skills in key subject areas, and to analyze, reason, and communicate effectively as they examine, interpret, and solve problems” (OECD, 2014a, para.1).

The PISA’s concept of literacy has been applied to all three subjects of Reading, Mathematics, and Science, and the correlations among the performance for the three subjects were unusually high, with coefficients reaching beyond $r=0.95$ (Soh, 2012, 2013, 2014). Thus, what can be said about PISA Science test is equally applicable to the PISA Mathematics and Reading tests. The proportion of TLH and NTLH also varies from country to country. When speakers and non-speakers of test language are pooled as a national sample, the country’s mean score could well be under-estimated since the non-speakers of the test language can logically be expected to have poorer command of the test language (Soh, 2014).

According to Soh (2014), “the language of an achievement test pre-conditions the level of performance of students who are assessed by the test via the test language” (p. 2). In the case of multiple-choice tests, students need to understand the item stems as well as the options. For open-ended questions, the students need to know what is being asked, and be able to write their answers correctly. Without adequate command of the test language, the students may not respond correctly even if they know the correct answers (Soh, 2014).

Using the argument of test language effect in terms of the students’ linguistic ability as the background context, this study was an attempt to show that the Reading score reported in PISA 2015 may not truly reflect the performance of the students from the three chosen countries. More specifically, the reading score may be underestimated due to the presence of sizable proportions of non-speakers of test language, who are by default second-language speakers where the test language is concerned since it is not the language they speak at home (Soh, 2014).

2.8 Validity of Inferences Made

With the increased use of large-scale assessments in recent years, it is becoming progressively necessary for test developers to prove to stakeholders that they are assessing the underlying skills and abilities that they are claiming to assess so correct interpretations can be made (Baumgartner & Steenkamp, 1998; Shaw, Crisp, & Johnson, 2012). The concept of validity has also evolved over time. Initially, it was defined in terms of whether a test measures what it is intended to measure with respect to its underlying construct, where a construct is defined as the underlying, and sometimes assumed, psychological process or personal trait rather than the test behavior or test scores (Cronbach & Meehl, 1955). Validity, as defined by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), now refers to the degree to which the collected evidence supports the interpretations made from the test scores based on the proposed use of the test. The process of validation encompasses the whole assessment process, from its purpose to its intended outcome, and cannot be isolated from its context (Messick, 1989b). Cronbach and Meehl (1955) also proposed three different types of validity evidence that can be collected for assessments: content-related, criterion-related, and construct-related. Since these different pieces of evidence are essentially complimentary to one another, a unified view of validity is now defined as a combined perspective of the test administration and the use of assessment scores, and not just the test itself (Shaw, Crisp, & Johnson, 2012).

One of the stated uses of the PISA results, besides measuring the *yield* of educational systems, is to use their scores to make comparisons between countries so they can serve as benchmarks for each other (OECD, 2010a). It is inferred that these scores show how much and what students in a nation can do (Ammermueller, 2013; Ammermueller & Pischke, 2009; Haertel, 1999), so the difference in scores can be used to compare countries. According to Messick (1989a), threats to the validity of the test score interpretation can occur from either (a) construct-irrelevant variance (measuring something other than the construct of interest) or (b) construct under-representation (incomplete measurement of the construct). Construct related evidence for validity of an assessment refers to the degree of association between the test score and what ability it is meant to describe or predict (Messick, 1989a). Haladyna and Downing

(2004) refer construct-irrelevant variance to systematic error (rather than random error) introduced into the assessment data variable that is unrelated to the construct being measured.

The PISA sampling frame ensures optimal representation of the 15-year-old population using criteria of eligibility and exclusion of students and schools within each country. The OECD has also set a high threshold of 85% response rate per country (Hopfenbeck et al., 2018). Using age as a criterion was perceived as problematic because the sample could include a sizeable proportion of students who had either repeated a class or skipped one (Prais, 2004). Goldstein and colleagues (2007) proposed the incorporation of longitudinal data to account for variability in grade retention and school starting age. Schuelka (2013) also commented on the exclusion of students with disability from participating in PISA, and warned that this procedure marginalized disabled students further and prevented them from taking part in any educational equity policy.

PISA applies a very rigorous procedure of double translation and adaptation of tests from two source versions - English and French (Grisay, de Jong, Gebhardt, Breezier, & Halleux-Monseur, 2007). Nevertheless, concerns as to the extent to which the adapted forms of PISA are comparable to the source versions are raised, as bias could emerge due to translation effects, differences in language, culture, curriculum coverage and so on (Grisay et al., 2007; Nardi, 2008). Such measurement invariance in different language versions of the same test could have serious implications on the validity of cross-lingual comparisons (e.g., Huang, Wilson, & Wang, 2016). A considerable number of studies examined measurement invariance and bias in PISA assessments, and most of the studies reported a substantial amount of differential item functioning (DIF) when comparing different language versions (Chen & Jiao, 2014; Huang, Wilson, & Wang, 2016; Oliveri & Ercikan, 2011; Oliveri, Olson, Ercikan, & Zumbo, 2012; Xie & Wilson, 2008).

2.9 Critique of PISA

PISA has been around for nearly two decades since its first survey in 2000, and now it has become one of the most influential international assessment programs worldwide (Zhao, 2020). The expanding influence of PISA is apparent by the increased participation of countries/economies in the survey, initially from 32 in 2000 to 72 in 2015 (OECD, 2016a). It is also evidenced by its expansion in the coverage of domains, from reading, mathematics, and science, to the addition of financial literacy and collaborative problem solving (OECD, 2017a). However, this growing influence has also been accompanied by growing criticisms from educational researchers all over the world. The criticisms revealed a wide range of problems with PISA, which can be summarized into four big areas.

The first criticism is the illusion of a unitary view of education, one that is primarily a Western view of societies. The PISA framework and its tests were first developed for member states of OECD, the majority of which includes the world's most advanced economies. Not surprisingly, using the framework as a benchmark for non-OECD countries can be a mismatch to the actual needs of these countries. PISA also claims that, using a set of universal indicators/instruments, it is possible to measure the quality of education across different countries and education systems, despite differences in political structure, social structure, traditions, culture, and ways of living (Sjoberg, 2015). However, in reality, there is more than one society in the world, and these societies are diverse and different from each other –

culturally, politically, economically and so on. Each one operates differently and present different challenges.

The second one is the illusion of science. Although PISA follows a scientific approach to developing test items and sampling methods, uses sophisticated psychometric theories and complex statistical modellings in data collection and analyses, and adopts a rigorous implementation procedures to arrive at and report the findings, PISA is still viewed as having inherent problems that threaten the quality and validity of its findings. The first concern involves sampling. Using age as a criterion instead of grade level was seen as problematic because the sample includes students at different grade levels with different exposure to the school curriculum. Another sampling problem is the issue of representativeness. For example, low participation rates can distort any generalization of PISA results to the entire 15-year-olds of a country. The third sampling problem is the exclusion of students with disabilities, an issue of equity which further marginalizes students with special needs (Schuelka, 2013). The second concern is bias in assessments. These biases include item formats, constructs, language, culture, and text types (e.g. Hopfenbeck et al., 2018). For instance, PISA reading assessment has been found to be girl-friendly and thus biased against boys (Solheim & Lundetrae, 2018). PISA instruments were also found to be more comparable across Western countries than they are across Middle Eastern or Asian countries for linguistic and cultural reasons (Grisay et al., 2007). Studies also reported a substantial amount of differential item functioning (DIF) when comparing different language versions (e.g. Ercikan, Oliveri, & Sandilands, 2013; Oliveri et al., 2012). These criticisms indicate that PISA may be measuring different abilities for students of different languages (Zhao, 2020).

The third criticism is the illusion of interpretation. PISA has been criticized for homogenizing education (Zhao & Gearin, 2016). Education systems and processes are becoming increasingly homogenous and standardized, for instance, implementing similar policies and practices regardless of the country's local needs and context. This is problematic because educational policies and practices are contextually and culturally sensitive, so what works in one context may not work or may even cause damage in another (Sjoberg, 2015). Such move also reduces diversity in educational approaches, values, processes, and talents.

The final criticism is the illusion of excellence. The PISA scores and standings on the league tables equate high-performing countries to educational quality and effectiveness of the education systems. However, there is no empirical evidence to justify or prove the claim that PISA indeed measures knowledge and skills essential for the future world (Hopmann, 2008). There is also no evidence of significant relationships between test scores and economic growth of a country (Komatsu & Rappleye, 2017).

It is clear from the criticisms raised against PISA that it has conceptual, technical, and interpretation issues, although users (e.g. education officials) of the outcomes normally take the PISA scores at face value (Soh, 2014). These criticisms in the PISA literature were considered as limitations in the present study. Care has been exercised when interpreting the findings to ensure that the actual performance of the country was reported as accurate as possible. These criticisms were also taken into account when making conclusions to explain and improve understanding about the uncertainties surrounding the results.

2.10 Chapter Summary

Education systems worldwide are faced with the same challenge of educating children who come from multiple languages and cultural backgrounds (Bailey & Butler, 2007; Bailey, Maher, & Wilkinson, 2018). Research has demonstrated that English language learners (ELLs) and language minority students typically confront substantial challenges in science and math assessments because of linguistically and culturally dependent content and representations in assessments (Abedi, 2001, 2004; Ercikan et al., 2015; Roth, Ercikan, Simon, & Fola, 2015; Soh, 2014). Typically, children from a different language and cultural background than the host country tend to have lower achievement levels on large-scale assessments (Ercikan et al., 2015).

The term *linguistic minority* includes a broad spectrum of students, from students who are schooled in a language other than their first language to students who live in a linguistic majority setting. In a testing situation, the students' language and cultural background interact in distinct ways. From a socio-cultural perspective, student performance on the assessment is the outcome of a complex interaction between their background (including language, practices, and culture) and the assessment properties (such as the targeted knowledge, the language being used in the test, and the format of the test items) (e.g., Pellegrino, Chudowsky, & Glaser, 2001; Solano-Flores, 2008). Duran (1989) described this complex relationship as follows:

> Contemporary cross-cultural research suggests that there are intimate connections among the ways people perceive the nature of problem-solving situations, problem-solving tasks, the language surrounding tasks, and socio-cultural experiences. Thus, analysis of issues affecting the testing of language minority persons is not totally reducible to

consideration solely of how lack of familiarity with a language affects test performance. (p. 574)

One of the advantages of PISA is that it enables comparisons among countries. The validity of comparisons of scores, however, depends a lot on whether the tests actually assess the same or similar knowledge and competencies across the groups and subgroups that are compared (Hambleton, Merenda, & Spielberger, 2005). The PISA's concept of *literacy* has been applied to all three subjects of Reading, Mathematics and Science. However, PISA involves students who speak the test language at home as well as those who do not. And when both speakers and non-speakers are pooled as a national sample, the country mean could well be an under-estimation of the actual milieu of the country (Soh, 2014).

Whenever students are not tested in their home language, a validity issue arises – how does language, rather than the targeted knowledge, affect their performance? While increasing efforts have been made to address concerns in this area, for example, by examining equivalence of language versions, translations, student language self-identification, sources of bias, scores from tests of similar content in different languages, and differences in test items across international examinee populations, to name a few (e.g. Aydin, Erdag, & Tas, 2011; Faulkner-Bond & Sireci, 2015; Oliveri, Ercikan, Zumbo, & Lawless, 2014; Solano-Flores, Backhoff, Contreras-Nino, & Vazquez-Munoz, 2015), there still continues to be a lag in test development and validation practices that embrace International Test Commission (ITC) guidelines concerning language for test development and evaluation (Anderson, Lin, Tregust, Ross & Yore, 2007; Rios & Sireci, 2014). As a result, patterns persist in international testing across the languages and cultures of young learners; these tests do not consider the effects of students' linguistic background and cultures on reading literacy assessment (Smith, Frazier, Lee, & Chang, (2018).. These tests also fail to acknowledge the need for evidence supporting more valid interpretations of assessments for linguistic minorities (Collier, 1987; Cronbach, 1989, Cummins, 1979).

The language complexity of the assessment instruments can be a factor affecting student performance on the test. Students must learn how to discern the meaning of a problem, which involves recognition of how the linguistic, cognitive, and social systems interact in any given test item. It demands that students comprehend the test items (including the language of presentation), formulate an approach to solving the problems, and apply the most appropriate knowledge to solve the problems correctly (Martiniello, 2009). The true competency of the student is undeniably affected by whether they have adequate language proficiency to read, understand, and write in the language in which they are tested (Abedi, 2004; Solano-Flores, 2008). When an assessment presents linguistic demands that are not targeted by the assessment, student competencies may be under-estimated.

Assessment of knowledge and competencies of English language learners (ELLs) has been researched extensively in the United States. It has been shown that the linguistic demands of assessments indeed lead to an under-estimation of competencies of ELLs (Abedi, 2004) and as much as about 40% measurement error can be attributed to language and its interaction with other factors (Solano-Flores & Trumbull, 2003). The purpose of this study is to investigate whether the language ability of TLH and NTLH students from the selected countries has a differential impact on their reading, mathematics, and science performance in the PISA 2015 assessment cycle. The findings can provide information as to how the language spoken at home can potentially affect their performance on PISA.

CHAPTER 3

METHODOLOGY

3.1 Research Background

PISA is a large-scale, international assessment that surveys 15-year-old students' performance in mathematics, reading, and science – the three literacy areas. In addition to assessment components, the PISA assessment also collects a wide variety of contextual data about students' demographic backgrounds, their learning attitudes and behaviours, their parents, and their schools. The OECD develops the PISA assessments in three-year cycles for OECD and non-OECD countries, with one of the three literacy areas (math, science, and reading) being the focus of the survey in any given year. PISA has been conducted every three years since 2000; the most recent cycle is PISA 2018.

As opposed to academic achievement, the aim of PISA is to assess the extent to which students can apply what they have learned during their years of compulsory schooling to real life problems as represented in the assessments (Smith, Frazier, Lee, & Chang, 2018). This differs, in part, from other international measures, such as the Progress in International Reading Literacy Study (PIRLS) and the Trends in International Mathematics and Science Study (TIMSS) which focus on school-based academic achievement. Another feature of PISA is that country level samples are nationally representative of their target populations, thus allowing statistical results to have wider generalizability (Smith et al., 2018).

In 2015, the sixth cycle of PISA was completed, with a major focus on the science domain. The science framework was originally developed for PISA 2006, but the framework has kept its essential features in 2015 to allow for trend analysis in performance over time. In addition, two major improvements for PISA 2015 were made. First, “knowledge about science” has been defined more clearly into two components – procedural knowledge and epistemic knowledge (OECD, 2015a). The 2015 science framework is discussed in more depth in section 3.2 below. The second change was the move from a paper-based to a computer-based assessment. Prior to 2015, PISA was implemented through a paper-based format. But recognizing the pervasiveness of computer-based tools in the 21st century, PISA 2015 assessed all subjects for the first time via computer, although paper-based assessment instruments were still provided for countries that chose not to do the test using computers (OECD, 2015a).

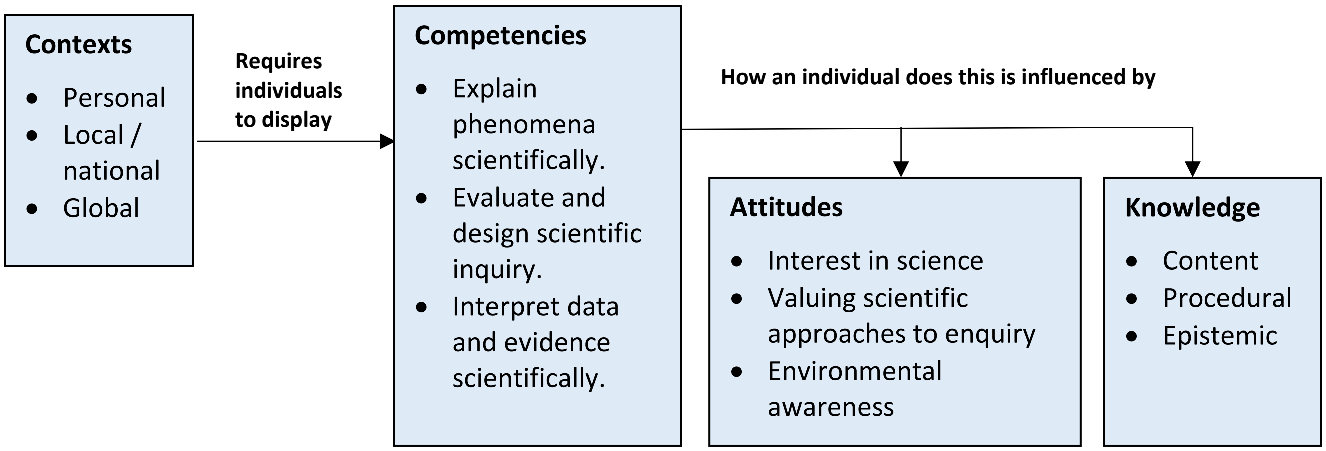
3.2 PISA 2015 Science Framework

The PISA 2015 Science framework is divided into three competencies, three knowledge types, and three areas of knowledge of science. In the PISA context, science refers to “scientific literacy”, which is defined as the ability to engage in reasoned discourse of scientific ideas and science-related issues as well as three “competency” subscales relating to: a) explaining phenomena scientifically, b) evaluating and designing scientific enquiry, and c) interpreting data and evidence scientifically (OECD, 2016a).

Student performance is based on their knowledge of science content as well as procedural and epistemic knowledge of science. “Content knowledge” refers to knowledge of facts, concepts, ideas, and theories about the natural world that science has established. “Procedural knowledge” refers to the knowledge of the practices and concepts on which empirical enquiry is based, while “epistemic knowledge” refers to an understanding of the role of specific constructs and defining features essential to the process of knowledge building in science. The main features of the PISA 2015 science framework are presented in Figure 3.1.

Figure 3.1

Main Features of the PISA 2015 Science Framework



Note: Source: PISA 2015 assessment and analytical framework (OECD, 2017a, p.25)

In addition, PISA 2015 also measures students' interest in and awareness of science and environmental issues as well as their perceived value of scientific approaches to enquiry. Three competencies are used to describe how a scientifically literate person engages in issues and ideas related to science. These competencies are described in Table 3.1.

Each of the scientific competencies requires some content knowledge (knowledge of theories, explanatory ideas, information, and facts), an understanding of how such knowledge has been derived (procedural knowledge), as well as the nature of that knowledge (epistemic knowledge). The three main areas of science knowledge are physical systems, living systems, and Earth and space systems (OECD, 2016b). This knowledge is classified according to the three broad content areas described in Table 3.2.

Table 3.1

Science Competencies

Competency	Description
Explain phenomena scientifically	being able to recognize, offer, and evaluate explanations for a range of natural and technological phenomena by demonstrating the ability to: • Recall and apply appropriate scientific knowledge. • Identify, use, and generate explanatory models and representations. • Make and justify appropriate predictions. • Offer explanatory hypotheses. • Explain the potential implications of scientific knowledge for society.
Evaluate and design scientific enquiry	being able to describe and appraise scientific investigations and propose ways of addressing questions scientifically by demonstrating the ability to: • Identify the question explored in a given scientific study. • Distinguish questions that could be investigated scientifically. • Propose a way of exploring a given question scientifically. • Evaluate ways of exploring a given question scientifically. • Describe and evaluate how scientists ensure the reliability of data, and the objectivity and generalizability of explanations.
Interpret data and evidence scientifically	being able to analyze and evaluate data, claims, and arguments in a variety of representations and draw appropriate scientific conclusions by demonstrating the ability to: • Transform data from one representation to another. • Analyze and interpret data and draw appropriate conclusions. • Identify the assumptions, evidence, and reasoning in science-related texts. • Distinguish between arguments that are based on scientific evidence and theory, and those based on other considerations. • Evaluate scientific arguments and evidence from different sources (e.g., newspapers, journals, the Internet).

Note. Source: PISA 2015 assessment and analytical framework (OECD, 2016a, p.27-28)

Table 3.2

Content Knowledge of Science

Content Knowledge	Description
Physical systems	This requires knowledge of the structure of matter, properties of matter, chemical changes of matter, motion and forces and action at a distance, energy and its transformation, and interactions between energy and matter.
Living systems	This requires knowledge of cells, the concept of an organism, humans, populations, ecosystems, and biosphere.
Earth and space systems	This requires knowledge of the structures of the Earth systems, energy in the Earth systems, change in Earth systems, Earth's history, Earth in space, and the history and scale of the universe and its history.

Note. Source: PISA 2015 assessment and analytical framework (OECD, 2016a, p.28)

3.3 PISA 2015 Science Achievement Results by Proficiency Level

PISA developed benchmarks that related a range of scores in science to levels of knowledge and skills that are measured by the assessment. Although these levels are not linked directly to any specific program of study in science, in the PISA context they provide an overall picture of students' accumulated understanding of science at age 15. PISA science literacy is expressed on a seven-level proficiency scale, in which tasks at the lower end of the scale (Level 1) are deemed easier and less complex than other tasks at the higher end of the scale (Level 6). This progression in task difficulty/complexity applies to both the overall science scale and for each competency and knowledge area (OECD, 2015a).

Table 3.3 provides a summary description of the tasks that students can do at the seven proficiency levels for overall science, along with the corresponding lower limit for each level.

It is assumed that students classified at a given proficiency level can engage with that task successfully at that level as well as those at the lower levels.

Table 3.3

PISA 2015 Science Proficiency Levels

Level	Lower score limit	Percentage [a]	Task characteristics
6	707.93	1.1%	Students at Level 6 of the PISA science assessment can successfully complete the most difficult PISA items. At Level 6, students can: • Draw on a range of interrelated specific ideas and concepts from the physical, life, Earth, and space sciences, link different information sources and representations, and move flexibly among them; • Use content, procedural, and epistemic knowledge to offer explanatory hypotheses of novel scientific phenomena, events, and processes or to make predictions; • Discriminate between relevant and irrelevant information and draw on knowledge external to the normal school curriculum when interpreting data and evidence; • Distinguish between arguments that are based on scientific evidence and theory and those based on other considerations; and • Evaluate competing designs of complex experiments, field studies, or simulations and justify their choices.
5	633.33	7.7%	At level 5, students can: • Use abstract scientific ideas or concepts to explain unfamiliar and more complex phenomena, events, and processes involving multiple causal links; • Apply more sophisticated epistemic knowledge to evaluate alternative experimental designs and justify their choices and use theoretical knowledge to interpret information or make predictions; and • Evaluate ways of exploring a given question scientifically and identify limitations in interpretations of data sets, including sources and the effects of uncertainty in scientific data.

Level	Lower score limit	Percentage [a]	Task characteristics
4	558.73	26.7%	At level 4, students can: • Use more complex or more abstract content knowledge, which is either provided or recalled, to construct explanation of more complex or less familiar events and processes; • Conduct experiments involving two or more independent variables in a constrained context; • Justify an experimental design, drawing on elements of procedural and epistemic knowledge; and • Interpret data drawn from a moderately complex data set or less familiar context, draw appropriate conclusions that go beyond the data, and provide justifications for their choices.
3	484.14	54.0%	At level 3, students can: • Draw upon moderately complex content knowledge to identify or construct explanations of familiar phenomena; • Construct explanations with relevant cueing or support in less familiar or more complex situations; • Draw on elements of procedural or epistemic knowledge to carry out a simple experiment in a constrained context; and • Distinguish between scientific and non-scientific issues and identify the evidence supporting a scientific claim.
2	409.54	78.8%	Level 2 is considered the baseline level of science proficiency that is required to participate fully in modern society. At level 2, students can: • Draw on everyday content knowledge and basic procedural knowledge to identify an appropriate scientific explanation, interpret data, and identify the question being addressed in a simple experimental design; • Use basic or everyday scientific knowledge to identify a valid conclusion from a simple data set; and • Demonstrate basic epistemic knowledge by being able to identify questions that could be investigated scientifically.

Level	Lower score limit	Percentage [a]	Task characteristics
1a [b]	334.94	94.5%	At level 1a, students can: • Use basic or everyday content and procedural knowledge to recognize or identify explanations of simple scientific phenomena; • Undertake structured scientific enquiries with no more than two variables with support; • Identify simple causal or correlational relationships and interpret graphical and visual data that require a low level of cognitive demand; and • Select the best scientific explanation for given data in familiar personal, local, and global contexts.
1b [b]	260.54	99.4%	At level 1b, students can: • Use basic or everyday scientific knowledge to recognize aspects of familiar or simple phenomenon; and • Identify simple patterns in data, recognize basic scientific terms, and follow explicit instructions to carry out a scientific procedure.

Note. Source: PISA 2015 results: Excellence and equity in education (OECD, 2016b)

[a] Percentage of students across the OECD who are able to perform task at this level or above.

[b] Level 1 is used interchangeably with Levels 1a and 1b.

3.4 PISA 2015 Reading and Mathematics

Since reading and mathematics were minor domains in PISA 2015, there were fewer assessment items in these two areas compared to the major domain of science. As minor domains, reading and mathematics were measured only as an overall performance instead of at a detailed level, and as such were not reported by performance levels or competency subscales.

3.4.1 PISA 2015 Reading and Mathematics Performance Scores

The scores for reading and mathematics among OECD countries are expressed on a scale with an average of 500 and a standard deviation of 100 (OECD, 2016a). This average was established when the domain was the focus of the assessment (i.e., in year 2000 and a re-examination in 2009 for reading; 2003 and further revision in 2012 for mathematics). The reading and mathematical literacy components of both the computer-based and paper-based instruments were composed of the same clusters of reading and mathematics trend items. But compared with previous PISA assessments, the number of trend items in PISA 2015 for reading and mathematics was increased for a better construct coverage while reducing the number of students responding to each question. This design was intended to reduce potential bias while stabilizing and improving trends measurement.

3.4.2 Reporting Proficiency in Reading

PISA 2015 reading literacy scale was based on PISA 2009, and is divided into seven levels (OECD, 2015a). Level 6 is the highest described level of proficiency (Level 5 was the highest level before PISA 2009). Level 1b is the bottom level of proficiency in PISA 2015. Since PISA 2009, Level 1 was re-labelled as Level 1a, and a new level, Level 1b, was added to describe students who were previously rated as "below level 1". Levels 2, 3, 4 and 5 remained the same in PISA 2015. A full description of the proficiency levels is included in Appendix B.

3.4.3 Reporting Proficiency in Mathematics

The six proficiency levels used in the PISA 2015 mathematics assessment are the same as those established for PISA 2003 and 2012 assessments, when mathematics was the major area of assessment (OECD, 2015a). Level 6 is the highest described level of proficiency, while 1b is the bottom level of proficiency. A description of the mathematical skills, knowledge and understanding that are required at each level of the scale is presented in Appendix C.

3.5 Research Design

Large-scale international assessments like Programme for International Student Assessment (PISA) (OECD, 2017a, 2017b) collect data by administering achievement tests to students of many different languages from participating countries. As a normal practice, tests are presented to students in their respective language of instruction – the language officially sanctioned as the country's medium of teaching, e.g., English in USA, German in Germany, Finnish in Finland (Soh, 2014). However, as Soh (2014) noted, the language used for teaching and assessment may not always be the same as the language spoken at home by all students of a participating country, thus the country's performance in international studies like PISA could be affected if the participating students are non-speakers of the test language. Although PISA has exercised care to ensure that the resultant test scores across countries are of a high degree of comparability and validity, the actual performance of the country may still be distorted to a certain extent due to the students' linguistic background.

This is an exploratory, quantitative study focusing on correlational relationships. The main methodology focused on a correlational analysis examining the relationship between language proficiency and performance in science and math assessments through an analysis of covariance. The study is designed to better understand whether language proficiency might be a factor that affects the math and science achievement scores when students are speakers or non-speakers of the test language. As noted earlier, the research done thus far on this topic was conducted in countries that are English-speaking, thus were solely focused on English language learners. To build upon this body of research, this study addresses possible impact of language on performance of students in Canada, Indonesia, and Singapore, and involves test administered in English in Canada and Singapore and Bahasa Indonesia in Indonesia. The data used in this study are drawn from the Organization for Economic Cooperation and Development's Programme for International Student Assessment 2015 (PISA) (OECD, 2017a, 2017b).

3.5.1 Assessment Domains

The PISA 2015 assessed three domains for all countries. These three core domains in the PISA assessment are *reading literacy, scientific literacy, and mathematical literacy.* They are defined by PISA (OECD, 2017a, p.15) as follows:

> ***Reading literacy*** (hereinafter referred to as reading) is an individual's capacity to understand, use, reflect on, and engage with written texts, in order to achieve one's goals, develop one's knowledge and potential, and participate in society.
>
> ***Scientific literacy*** (hereinafter referred to as science) is an individual's ability to engage with science-related issues, and with the ideas of science, as a reflective citizen. A scientifically literate person is willing to engage in reasoned discourse about science and technology, which requires the competencies to explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence scientifically.
>
> ***Mathematical literacy*** (hereinafter referred to as mathematics) is an individual's capacity to formulate, employ and interpret mathematics in a variety of contexts. It includes reasoning mathematically and using mathematical concepts, procedures, facts, and tools to describe, explain and predict phenomena. It assists individuals to recognize the role that mathematics play in the world and to make the well-founded judgments and decisions needed by constructive, engaged, and reflective citizens.

Each of the literacy domains assessed by the PISA 2015 encompasses construct definitions that stress key knowledge and skills needed to function in everyday tasks and be an active member of society. These definitions also incorporate tasks that require critical thinking, problem solving, decision making, evaluating, applying, and extrapolating students' knowledge to novel situations as they would beyond classroom setting (OECD, 2017a).

3.5.2 Reading Score

OECD's PISA definition of *literacy* deviates from traditional notions that emphasize "mastery of specific school [reading] curricula" (OECD, 2014a, p.25). PISA operationalizes *literacy* as a "student's ability to apply knowledge and skills in key subject areas, and to analyze, reason, and communicate effectively as they examine, interpret, and solve problems" (OECD, 2014a, p. 25). This definition challenges the notion of literacy as a set of foundational or constrained skills (Paris, 2005); instead, it suggests that literacy is dynamic, social, cultural, and multi-faceted (Cope & Kalantzis, 2012).

As opposed to an assessment of achievement, PISA is a literacy assessment that evaluates students' readiness for everyday reading-related tasks that they may face after leaving school. In this respect, PISA does not seek to measure reading-related skills that students have mastered at school. Rather, it seeks to measure students' capability to put their knowledge into practical use (OECD, 2017a, 2017b). PISA provides five "plausible" values for a student's reading score because each student completed only a subset of the items (OECD, 2015a). This use of plausible values as scores for each student is discussed in more depth in Section 3.7.

3.5.3 Socio-economic Status

Socio-economic status (SES) is typically defined by family income, level of poverty in the child's neighborhood, educational attainment by parents, and occupation of the heads of households (Clements & Sarama, 2008). Many studies found that SES is a strong predictor of student achievement and is associated with large differences in performance in most countries and economies that participated in PISA (OECD, 2014a). Researchers have examined the effect of SES on mathematical achievement (Frenette, 2007), and children from disadvantaged low-income families perform substantially worse in mathematics than their counterparts from higher income families (National Mathematics Advisory Panel [NMAP], 2008). Chiu (2007, 2009, 2010) also reported a strong association between family communication and academic performance. Their research revealed that children show higher academic performance when parents are more involved in their children's learning activities. Socio-economically advantaged students and schools also tend to outscore and outperform their disadvantaged peers by a large margin (Baker, Goesling, & Letendre, 2002; OECD, 2014a). Although having low SES does not necessarily equate to a negative effect on academic performance, it is considered a dominant trend that can be associated to unfavourable educational outcomes (Baker, Goesling, & Letendre, 2002).

3.5.4 Test Language

One context in which home language and test language matters is the PISA assessment. It is widely agreed that proficiency in the language in which a test is administered tends to affect examinees' performance on the test (Faulkner-Bond & Sireci, 2015). Regardless of the lack of intent to draw inferences for PISA about language spoken at home, increasingly, research demonstrates the effects of language on PISA mathematics and science literacy assessments (e.g., Ercikan et al., 2015; Geva, 2006; Gottardo & Mueller, 2009). In comparison, fewer studies have concentrated on the effects of language on the PISA reading literacy assessment; in fact, there is little information about the impact of linguistic differences from students' backgrounds on their achievement in PISA measures (e.g., Asil & Brown, 2016). Given that the PISA reading literacy test is used as a basis for numerous comparisons of youth across the globe, by determining the impact of students' language status on PISA reading literacy performance, evidence about the validity of students' performance on the assessment as indicator of knowledge, skills and competencies, and the interpretation of this performance can be obtained. Such findings can help determine consistency of findings across studies and arrive at more definitive and accurate conclusions (Makel & Plucker, 2014). Such studies also align with the International Test Commission's (ITC) guidelines concerning validity that recommend determining student proficiency in the language of test administration and aligning them with the appropriate language version of the test administered (International Test Commission [ITC], 2017, 2018).

3.6 PISA Sampling, Sampling Weights and Booklet Design

3.6.1 PISA Sampling

PISA employs a two-stage stratified sampling design. In the first stage, within each country individual schools are sampled using probability proportional to size sampling. A minimum of 150 schools are selected in each country; if a participating country has fewer than 150 schools then all schools participate. In the second stage, within each participating school a predetermined number of students are selected as the target cluster size or 'eligible students' –

usually 42 students in computer-based countries and 35 students in paper-based countries (OECD, 2016a, 2016b). Students within the sampled schools are randomly selected with equal probability. In schools with fewer than the number of target cluster size, then all students are selected. In total, a minimum sample size of 5250 students is needed in computer-based countries, and 4500 students in paper-based countries, or the entire population if it is less than the target cluster size.

3.6.2 Sampling Weights and Booklet Design

Sampling weights are used to make valid estimates and inferences of the target population and calculate appropriate estimates of sampling error. In terms of scaling, a matrix-sampling of items or rotated block design method is used, whereby participants are given only a small sample of all the items and their responses placed on a common scale to provide an overall picture. In such sampling design, assessment items are assigned to several blocks that are then combined in systematic ways into a set of booklets, with each student completing just one booklet.

Computer-based Format. The 2015 PISA computer-based assessment included six clusters from each of the trend domains of science, reading, and mathematics, and an additional six clusters of new science literacy test items, and three clusters of new collaborative problem-solving materials. The clusters were allocated in a rotated design using 66 test forms to create six groups (OECD, 2016a, 2016b). Each booklet was composed of various combinations of clusters. Some students did booklets containing items from all the three areas, while others took part solely in one (the area of focus, science) or two areas (the area of focus, science, and one from the other two areas, reading and mathematics). For this study, only the first twelve clusters on science, reading, and mathematics literacy were used; the three clusters on collaborative problem solving were excluded.

According to the design,

- 33 percent of students within each school were assigned to one of 12 science literacy and reading literacy test booklets;
- 33 percent were assigned to one of 12 science literacy and mathematics literacy test booklets;
- 22 percent were assigned to one of 6 science literacy and collaborative problem-solving test booklets;
- 4 percent were assigned to one of 12 science literacy, mathematics literacy, and collaborative problem-solving test booklets;
- 4 percent were assigned to one of 12 science literacy, reading literacy, and collaborative problem-solving test booklets; and
- 4 percent were assigned to one of 12 science literacy, reading literacy, and mathematics literacy test booklets.

In the design, approximately 65 percent of science items were multiple-choice formats and 35 percent were constructed responses. For reading and mathematics items, approximately 40 percent were multiple choice formats and 60 percent were constructed responses.

Paper-based format. For the 2015 PISA paper-based version assessment, only six clusters from each of the trend domains of science, reading, and mathematics literacy were included. The clusters were also allocated in a rotated design to create three groups of test booklets. According to the design,

- 44 percent of students within each school were assigned to one of 6 science literacy and reading literacy test booklets;
- 44 per cent were assigned to one of 6 science literacy and mathematics literacy test booklets;
- 12 percent were assigned to one of six science literacy, reading literacy, and mathematics literacy test booklets.

All students who took the paper-based assessment answered science items, but not everyone answered mathematics and/or reading items.

3.7 Plausible Values

Since students receive only a subset of all assessment items in PISA 2015, the measurement of individual proficiency is expected to have large measurement error due to small number of items administered to individual students (von Davier, Gonzalez, & Mislevy, 2009). In order to address this, the multiple imputation method, also referred to as the *plausible values approach* (Mislevy, 1991; Rubin, 1987), is used to reflect the estimates of population characteristics. Missing information that is present due to the matrix sampling design used in PISA is arranged into a non-monotone missing data pattern, where all variables are highly and positively correlated (e.g. Durrant, 2009; Horton & Kleinman, 2007; Wilson & Lueck, 2014). A set of scores referred to as plausible values (PVs) is generated for each student.

The PVs approach uses students' responses to the subject-based items collected from the cognitive questionnaire, together with other data collected from the student background questionnaire, parent questionnaire and school questionnaire, in order to directly estimate the characteristics of student populations and subpopulations. In this way, estimates of student performance may be obtained on the assessment even though each student responds only to just a subset of the assessment items. PVs are not considered as test scores for individual students, but rather as a measure of performance of the population. A more detailed description about plausible values can be found in Monseur and Adamas (2009).

Since 2000, student proficiency estimates in PISA survey were computed through *plausible values.* Commonly, five PVs are generated for each student on each performance scale. Statistical analyses are performed independently on each of the five plausible values and results are aggregated to obtain the final estimates of the statistics and their respective standard errors (OECD, 2017a). The plausible value methodology, combined with the replicates, requires that the parameter, such as a mean, a standard deviation, a percentage, or a correlation, has to be computed 85 times (i.e. 5 plausible values by one student final weights and 80 replicates) to obtain the final estimate of the parameter and its standard error (OECD, 2017a). Working with only one plausible value can provide unbiased estimate of population parameters, but it cannot estimate what five plausible values can do - the imputation error that reflects the influence of test unreliability for the parameter estimation. The smaller the sample size, the greater is the imputation error. Hence, using all five plausible values may provide a better picture of the substantial differences in student performance, especially for small samples. In the PISA 2015, ten PVs were generated for each student; all ten PVs were used in the data analyses of the present study.

3.8 Dataset

The present study is based on PISA 2015 *Student Questionnaire* data, which are well documented and freely downloadable from the official OECD website (www.pisa.oecd.org). Approximately 540,000 students participated in the PISA 2015, representing about 29 million 15-year-olds across 72 participating countries / economies (OECD, 2016a).

As briefly mentioned in Chapter 1, Canada, Indonesia, and Singapore were selected from the PISA 2015 dataset after a review of their test-takers' background information that exhibit the following characteristics in line with the purpose of this research:

- Canada – being a country where most test-takers took the English version of the PISA 2015 assessment, and with the majority of them being speakers of the test language (i.e. TLH). The PISA test taken by the students was computer-based.
- Indonesia – being a country where the test-takers took the PISA 2015 assessment using the Bahasa Indonesian version, but a majority were non-speakers of the test language (i.e. NTLH). The PISA test used by the students was the paper-based version because the country was not yet ready to transition to a computer-based assessment. The paper-based test comprised a subset of the tasks included in the computer-based version, all of which were developed in earlier cycles of PISA. The newly developed tasks for science and reading in PISA 2015 was not included in the paper-based test.

- Singapore – where an equal proportion of TLH and NTLH took the PISA 2015 assessment in English. Singaporean students took the computer-based PISA test.

These three countries were selected also because they represented different performance levels on PISA, with Singapore achieving the highest level of performance in Science, Reading and Mathematics, and outperforming both Canada, one of top performing countries, and Indonesia, a relatively low performing country. Furthermore, the three countries were also chosen for data availability; these three countries have sufficiently large sample sizes for TLH and NTLH groups analyses and comparison within these countries.

In total, seventy-two countries participated in PISA 2015; approximately 5,000 - 25,000 students aged 15 from at least 150 schools in each country did the PISA study. The PISA 2015 was administered in schools during regular school hours. The assessment was a two-hour paper/computer-based test. In addition, students also completed a 35-minute student background questionnaire providing information about themselves and their home. An overview of PISA 2015 (internationally and the three selected countries) is given in Table 3.4.

Table 3.4

Overview of PISA 2015 (International and the Three Selected Countries)

	International	**Canada**	**Indonesia**	**Singapore**
Participating countries	72	Whole	Whole	Whole
Population	Youth aged 15	Same	Same	Same
No. participating students	5,000-25,000 per country	20058	6513	6115
Domains	Major: Science	Same	Same	Same
	Minor: Reading and Math			
Language in which test was administered	47 languages	English, French	Bahasa Indonesian	English
Format of assessment	Computer/Paper	Computer	Paper	Computer

Note. Source: PISA 2015 results: Excellence and equity in education (OECD, 2016b)

3.9 Target Population

The data for this study were drawn from the publicly accessible Student Questionnaire data. A total of 32,686 students representing 15-year-old students who are native English speakers, English language learners, and language minority speakers living in Canada, Indonesia and Singapore formed the analysis population. Their performance in science, reading and mathematics in the 2015 PISA were also collected. Figure 3.2 and Table 3.5 show the research sample for the data.

Figure 3.2

Research Samples Using PISA 2015 Data

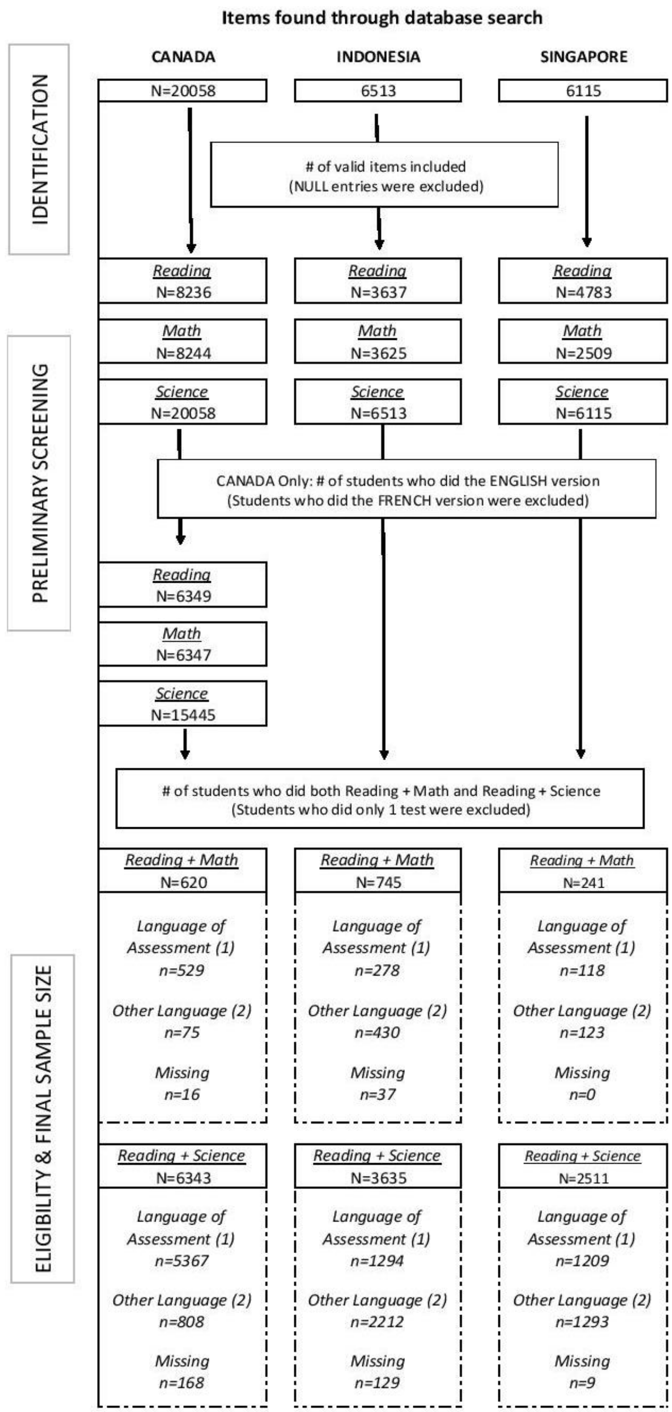


Table 3.5

Target Population

Selected Countries	Test Language	N All	N Valid	Language of test N	Language of test %	Other language N	Other language %
Canada	English & French	20058	19536	15902	81.40	3 551	18.18
Indonesia	Bahasa Indonesian	6513	6513	2254	34.61	4 013	61.63
Singapore	English	6115	6098	3087	50.63	3 005	49.29
Total		32686	32147	21243	66.08	10 569	32.88

Note. Source: PISA 2015 results: Excellence and equity in education (OECD, 2016b)

3.10 Samples for the Study

The target study samples for the three selected countries ranged between 6,115 students from Singapore to 20,058 students from Canada, as shown in Table 3.5. Students were further selected to include only those who did both reading and math (R+M group) and both reading and science (R+S group). After preliminary screening and following the eligibility criteria for the research samples described in Figure 3.2, a total of 12,489 students met the inclusion criteria and were included in the final sample group of the present study (Table 3.6).

Table 3.6

Final Sample Group for Canada, Indonesia, and Singapore

Countries	Domain	Total (valid)	Final Total	Reading + Math				Reading + Science			
				Lang Test	**Non-Lang Test**	**Mis sing**	**Total**	**Lang Test**	**Non-Lang Test**	**Mis sing**	**Total**
Canada [a]	Reading	8236	6349	529	75	16	**620**	5367	808	168	**6343**
	Mathematics	8244	6347								
	Science	20,058	15445								
Indonesia	Reading	3637	3637	278	430	37	**745**	1294	2212	129	**3635**
	Mathematics	3625	3625								
	Science	6513	6513								
Singapore	Reading	4783	4783	118	123	0	**241**	1209	1293	9	**2511**
	Mathematics	2509	2509								
	Science	6115	6115								
TOTAL				925	628		**1606**	7870	4313		**12489**

Note. Both Canada [124] and Singapore [702] used the computer version of the assessment. Indonesia [360] used the paper version of the assessment.

[a] The numbers only represent students who did the English version; French was excluded.

3.11 Language Spoken at Home

The study investigates the association of language proficiency and performance to Mathematics and Science assessments. The first step in the analyses was to identify the home language background of the students, which was a key defining variable for the language groups of all three countries included in the study.

In PISA 2015, students were asked to report what language they speak at home with a single item (International Language at Home (hereinafter referred to as Language at Home or Language Group), ST022Q01TA, “which language do you speak at home most of the time?”). Based on the responses, the students were divided into two language groups: students who speak the test language at home (TLH), and students who are non-speakers of the test language at home (NTLH). 7,870 of the students (63.0%) spoke the language of assessment (Test Language Speaker at Home, or TLH), while 4,313 (34.5%) spoke other languages at home (Non-Test Language Speaker at Home, or NTLH). Table 3.7 describes the language spoken at home.

Table 3.7

Language Spoken at Home

International Language at Home (ST022Q01TA)	Canada	Indonesia	Singapore	Total
Language of Test	5367 (84.6%)	1294 (35.6%)	1209 (48.2%)	7870 (63.0%)
Other Language	808 (12.8%)	2212 (60.8%)	1293 (51.5%)	4313 (34.5%)
Missing	168 (2.6%)	129 (3.6%)	9 (0.3%)	306 (2.5%)
TOTAL	6348 (100%)	3642 (100%)	2504 (100%)	12489 (100%)

3.12 Instrument

The PISA 2015 survey focused on science, with reading, mathematics and collaborative problem solving as minor areas of assessment. PISA 2015 also included an optional assessment of young people’s financial literacy. In general, PISA assesses not only whether students can reproduce knowledge, but also whether they can extrapolate from what they have learned and apply their knowledge in new situations. Specifically, it emphasizes the mastery of processes, the understanding of concepts, and the ability to function in various types of situations (OECD, 2015a).

The instruments for the PISA 2015 data collection that generated the sample for this study included a student questionnaire and cognitive booklets. Test items were a mixture of multiple-choice questions, and questions requiring students to construct their own responses. The items were organized in groups based on a passage setting out a real-life situation. About 810 minutes of test items for science, reading, mathematics and collaborative problem solving were covered, with different students taking different combinations of test items. Computer-based tests and paper-based tests were used, and the assessments lasted approximately two hours for each student.

Students also completed a 35-minute background questionnaire, which sought information about the students themselves, their homes, and their school and learning experiences. There were also optional questionnaires for parents, teachers (Science teacher and General teacher) and school principals. Such data collected were not included in the sample of the present study. The data for each of the three countries in the sample were analyzed separately, and findings are presented in Chapter 4 - Results.

3.13 Measures

The descriptions of the variables included in the analysis of this study are presented in Table 3.8. Each variable is described separately.

Table 3.8

Descriptions of Variables Included for Analysis

Factor	Description
Reading Score	Overall reading score
Mathematics Score	Overall mathematics score
Science Score	Overall science score
International Language at Home	International language of students at home. Binary variable (1 = Language of test, 2 = Other Language) Note: This is the variable used to separate students into TLH and NTLH groups in the present study
Gender	Binary variable (1 = Female, 2 = Male)
SES Index (ESCS)	PISA-developed mean-standardized scores from set of component variables: (a) Parental education (PARED) – highest education of parents in years, (b) Highest parental occupation (HISEI), (c) Home possessions (HOMEPOS) All three components are standardized to have a mean of zero and standard deviation of one after imputation

3.13.1 Covariate Variable (CV)

Reading Score. The PISA Reading test is made up of three subtests dealing with aspects of cognitive strategies, approaches, or purposes that participants use to negotiate their way into, around, and between texts (OECD, 2017a, p. 56). The first subtest, *Access and Retrieve,* assesses skills related to finding, selecting, and collecting information. The second subtest, *Integrate and Interpret,* evaluates processing of what is read to make sense of a text; this requires the understanding of relationships between different parts of a text. The third subtest, *Reflect and Evaluate,* involves drawing on knowledge, ideas, or values external to the text, reflecting on a text, and relating own experience or knowledge to the text. The reading score has an OECD mean of 500 and a standard deviation of 100 points. The total Reading Score, a single reading literacy scale which is an overall combination of the three subtest scores, was used as the covariate (CV) variable for this study. As explained in section 3.7, because students completed only a subset of the items in the test booklets, PISA provided ten plausible values (PVs) for a student's reading score; all ten PVs were used in the analyses.

3.13.2 Dependent Variables (DV)

Mathematics Score. A mathematics score, treated as a proxy for math ability, measures the individual's mathematics performance (OECD, 2017). In mathematics, PISA measures students' ability to activate their knowledge and skills to solve problems found in real-life situations. It centers around three major domains of assessment: *Mathematics Content Categories, Mathematics Contexts,* and *Mathematical Processes.* As mentioned before, the math score has an OECD mean of 500 and a standard deviation of 100 points.

Science Score. The science score reflects students' performance in science through questions related to contexts, knowledge, competencies, and attitudes. Students engage with science-related issues, explain phenomena scientifically, evaluate and design scientific enquiry, and interpret data and evidence competently. The science score also has an OECD mean of 500 and a standard deviation of 100 points.

3.13.3 Independent Variables (IV)

Home Language Status. The PISA 2015 home language status is based on students' responses to one item [ST022Q01TA], asking if the language spoken at home most of the time was the language of testing, the official national language, or another language. Based on the responses, students for this study are categorized into two groups: students who speak the test language at home (TLH), and students who do not speak the test language at home (NTLH). For the three sampled countries, the language in which the test was offered was also the language of instruction. Out of the total 12,494 students who formed the final study sample, as shown previously in Table 3.7, 63% of the students (n = 7,872) spoke the test language (TLH) and the remaining 34.5% (n = 4,316) spoke other languages at home (NTLH).

Gender. Gender is a basic demographic characteristic included in the study. A binary variable was used to indicate the gender of each adolescent. Female was coded as 1 and male was coded as 2.

Socio-Economic Status (SES) / Economic Social-Cultural Status (ESCS). PISA developed its own instrument to measure the socio-economic characteristics of the students. The "index of economic, social, and cultural status (ESCS) variable (OECD, 2016a) was considered/identified as the SES variable, and the two labels SES and ESCS will be used interchangeably hereinafter. The ESCS has a mean of zero and a standard deviation of one. It was measured using three standardized indicators, including parental education (in years of schooling), highest parental occupation (as measured by the International Socio-Economic Index of Occupational Status), and home possessions (e.g., such as books at home), to formulate a single ESCS index. To create distinctly high and low SES groups, the ESCS variable was divided into three equal proportions. Individuals with ESCS values in the top 33% were classified as the "high ESCS" group, whereas individuals who had ESCS values in the bottom 33% were classified as the “low ESCS” group. The ESCS variable was recoded into 1 = low ESCS and 3 = high ESCS.

3.14 Analyses

The statistical results in this study were based on secondary analysis of a survey-based, nationally representative sample of 15-year-old students from Canada, Indonesia, and Singapore. Prior to analyses, several steps were taken. Student data from the three selected countries was downloaded from the OECD website and merged to create a master dataset. Factors (variables) of interest were investigated as described below. The SPSS version 27 statistical software was used in analyzing the sample in this study.

Home language background was the key defining variable in categorizing students from all three countries into two groups: students who speak the test language at home (TLH), and students who do not speak the test language at home (NTLH). To ascertain the effect of test language in each country, comparisons were made with the sample as a whole and as subsamples of TLH and NTLH students. The research focused on the differential relationships between reading proficiency and mathematics and science achievement, and comparability and consistency of score meaning for TLH and NTLH students in each country.

3.14.1 Descriptive Statistics

Descriptive statistics were obtained for the combined total scores of the PISA Reading, Math, and Science for all participants in each of the three countries separately. These statistical options included the average scores, percentages, standard deviations, and correlations. Significance testing using two-sample *t*-tests was used to see whether the difference between the language groups is significant or due to random chance. Effect size (Cohen, 1988) in terms of standardized mean difference was also calculated to evaluate the magnitude of the difference between these groups. All analyses were conducted at 0.05 significance level.

3.14.2 Analysis of Covariance (ANCOVA)

A key method for examining the degree to which a particular variable can account for variation in an outcome variable is Analysis of Covariance (ANCOVA) (Maxwell, O'Callaghan, & Delaney, 1993). ANCOVA can be thought of as an extension of Analysis of Variance (ANOVA) because they are both used to determine whether there are any significant differences between two or more independent groups on a dependent variable. However, whereas the analysis of variance (ANOVA) looks for differences in group means, the ANCOVA looks for differences in adjusted means (i.e. adjusted for the covariate). Compared to ANOVA, the ANCOVA method has the additional benefit of 'statistically controlling' for a third variable (sometimes known as a 'confounding variable') that has an influence on the dependent variable and which can affect the results. This third variable is called the *covariate.* This method allows for estimating adjusted mean scores of the outcome variable when the covariate is taken into account. They are included in the analysis to create a hierarchical regression in which the dependent variable is the outcome. In an ANCOVA analysis, the covariate is entered in the first block. In the second block, the dependent variables are entered to see what effects independent variable(s) has/have *after* the effect of the covariate is considered.

Including covariates in ANCOVA serves two purposes. The first reason is to reduce within-group error variance. The effect of the independent variable(s) can be assessed by comparing the amount of variability in the data that the study can explain against the variability that it cannot explain. If the ‘unexplained’ variance can be explained in terms of covariates, then the error variance can be reduced, thus allowing for a more accurate assessment of the effect of the independent variable(s). Another reason for including covariates is that it can help in the elimination of unmeasured variables that may confound the results. If any variables are known to influence the dependent variable(s), then ANCOVA is ideally suited to remove the bias of these variables from the comparisons. Once a possible confounding variable has been identified, it can be measured and entered as a covariate in the analysis.

ANCOVA was the main data analysis method used in this study. Reading scores served as the covariate (CV) for each group comparison of TLH and NTLH. Mathematics and science scores were the dependent variables (DV), while independent variables (IVs) included a grouping variable that identified students as TLH or NTLH, as well as gender and socio-economic status (ESCS).

ANCOVA Assumptions. ANCOVA has the same assumption as any linear model, except that there are two additional important considerations. The first assumption required for ANCOVA analyses is the independence of the covariate and treatment effect. That is, the covariate should not be different across the groups in the analysis. The second assumption is the homogeneity of regression slopes. This means that the CV must not have a differential association with the DV at different levels of the IV. If this assumption is violated, then ANCOVA results cannot be interpreted meaningfully for different levels of the IV (Henson, 1998; Shadish, Cook, & Campbell, 2002).

In the present study, all assumptions of ANCOVA were tested. (1) Linearity of regression: A linear relationship exists between the dependent variable (DV) and the covariate (CV). (2) There needs to be homogeneity of error variances. The error is a random variable with conditional zero mean and equal variances for different treatment classes and observations. (3) There needs to be homogeneity of regression slopes. The slope of the regression lines should be all equal and parallel to each other. (4) There is no interaction between the independent variable (IV) and the covariate. Residuals (error terms) should be normally distributed. (5) Independence of error terms. The errors are uncorrelated, i.e. the error covariance matrix is diagonal. If the assumptions of ANCOVA are met, then the adjusted mean scores for math and science are compared for TLH and NTLH group. If assumptions are not met, appropriate adjustments are made to the modelling of dependent, covariate and independent variables before comparing scores for TLH and NTLH.

3.15 Chapter Summary

International studies like PISA assess students on core subjects such as Reading, Mathematics and Science. However, students who do not speak the test language at home can be disadvantaged because of limitations in language proficiency. This study investigated the possible association of language proficiency and performance on Mathematics and Science assessments of students from Canada, Indonesia, and Singapore, using data from the PISA 2015. These countries were selected due to their characteristics in providing important insights about the key research questions: a large proportion of non-speakers of the test language different from English such as Indonesia, a country with language-minority immigrants such as

Canada, and bilingual speakers who were taught effectively in the test language and a second home language such as Singapore.

Most research done on language-minority students has used samples from primarily English-speaking countries, e.g., the United States, Australia, Canada, and the United Kingdom, where the focus has mostly been on English Language Learners. This research, with the methodology and approach as explained in this chapter, however, did not only include the English Language Learners, but it also included students in which English is not the major academic and assessment language as well as students taught effectively in the test language and a second home language. It is anticipated that the study may likely reveal differential impact of language ability on performance in math and science assessments of students from English and non-English backgrounds. And given the well-documented performance differences between gender and socio-economic groups, this study may also shed light as to whether the relationship between reading ability and mathematics and science performance vary with gender and socio-economic groups.

CHAPTER 4

RESULTS

4.1 Re-statement of the Research Problems

The purpose of this study is to investigate the association of reading proficiency and performance on Math and Science assessments using PISA 2015. To add to the current body of research, the following research questions were addressed in this study.

RQ1: To what degree do students' reading abilities (PISA Reading Score) account for their math performance (PISA Math Score)?

RQ2: To what degree do students' reading abilities (PISA Reading Score) account for their science performance (PISA Science Score)?

RQ3: To what degree does the relationship between reading ability and mathematics and science performance vary for gender and socio-economic groups?

The research hypothesized that speaking the test language at home or not accounts for variation on their performance in PISA 2015. For the purposes of this study, TLH students are defined as "those who speak the test language at home", whereas NTLH refers to "students whose home language is not the test language of the PISA assessment in their respective countries of residence".

4.2 Descriptive Analyses

Descriptive statistics of the study population's performance in 2015 PISA Science, Reading, and Mathematics were obtained. These statistical results included the mean scores by subject for Canada, Indonesia, and Singapore. The mean scores of the TLH and NTLH groups on individual subjects and the significance level for the three countries were also calculated (Table 4.1).

Table 4.1

Study Population's Performance in 2015 PISA Science, Reading, and Mathematics (Language Group)

Selected Countries	**Subjects**	**Test Format**	**Sample Size**	**Mean Score (subject)**	**TLH**		**NTLH**	
					Mean Score	**SE**	**Mean Score**	**SE**
Canada	Science	Computer	6348	519.66	**521.94*** **	1.20	510.37	3.19
	Reading		6348	518.94	**520.96***	1.17	511.51	3.08
	Mathematics		619	504.64	504.03	3.41	508.03	9.01
Indonesia	Science	Paper	3642	409.77	411.60	1.95	410.10	1.27
	Reading		3642	406.35	**413.81*** **	2.13	403.64	1.37
	Mathematics		745	390.26	**398.77***	4.65	385.18	3.29
Singapore	Science	Computer	2504	546.04	**570.91*** **	2.72		
	Reading		2504	525.94	**551.49*** **	2.52		
	Mathematics		239	551.68	**574.07*** **	7.42		

Note. TLH, Test Language Speakers at Home; NTLH, Non-Test Language Speakers at Home. In Canada, only students who did the English version of the PISA questionnaire were included. *p < 0.05, two-tailed. ***p < 0.001, two-tailed.

For all three countries, the TLH group has a higher mean score on all three subjects compared with the NTLH group, except for Canada where the NTLH group has a higher mean score in mathematics than the TLH group but the difference was not statistically significant. In

Canada, the mean score difference between TLH and NTLH was found to be significant for science and reading. In Indonesia, the reading and mathematics scores between TLH and NTLH were found to be significantly different, whereas in Singapore the mean score difference between TLH and NTLH was found to be significantly different for all three subjects.

4.3 Correlations

Correlation measures the strength of the linear association between two quantitative variables, in this case, the dependent variables mathematics and science scores and the covariate reading score. The reading score, mathematics score, and science score were correlated with each other using pairwise deletion, with a null hypothesis that there is no correlation among the variables. In Tables 4.2 to 4.3, Pearson's correlation matrices displaying the r values at 95% confidence level for each pair of quantitative variables from the three countries are presented.

For Canada, a fairly strong positive correlation (+/- 0.80 or higher) was found between reading and math scores ($r = 0.882$), reading and science scores ($r = 0.933$), and math and science scores ($r = 0.942$) (Table 4.2). As reading score increases, math and science scores also increase. Similarly, as math score increases, science score also tends to increase.

Table 4.2

Correlational Matrix among the Main Variables (Canada)

Canada		**Reading Score**	**Math Score**	**Science Score**
Reading Score	Pearson's *r*	1		
	Sig (2-tailed)			
	N	6348		
Math Score	Pearson's *r*	**.882*** **	1	
	Sig (2-tailed)	.000		
	N	619	619	
Science Score	Pearson's *r*	**.933*** **	**.942*** **	1
	Sig (2-tailed)	.000	.000	
	N	6348	619	6348

***p < 0.001, two-tailed.

Similar findings were also noted for Indonesia. In Table 4.3, a fairly strong positive correlation (+/- 0.80 or higher) was found between reading and math scores (r = 0.805), reading and science scores (r = 0.842), and math and science scores (r = 0.887). As reading score increases, math and science scores also increase. Similarly, as math score increases, science score also tends to increase.

Table 4.3

Correlational Matrix among the Main Variables (Indonesia)

Indonesia		**Reading Score**	**Math Score**	**Science Score**
Reading Score	Pearson's *r*	1		
	Sig (2-tailed)			
	N	3642		
Math Score	Pearson's *r*	**.805***	1	
	Sig (2-tailed)	.000		
	N	745	745	
Science Score	Pearson's *r*	**.842***	**.887***	1
	Sig (2-tailed)	.000	.000	
	N	3642	745	3642

***p < 0.001, two-tailed.

In Singapore, a fairly strong positive correlation (+/- 0.80 or higher) was found between reading and math scores (r = 0.890), reading and science scores (r = 0.945), and Math and Science scores (r = 0.941) (Table 4.4). As reading score increases, math and science scores also increase. Similarly, as math score increases, science score also tends to increase.

Table 4.4

Correlational Matrix among the Main Variables (Singapore)

Singapore		**Reading Score**	**Math Score**	**Science Score**
Reading Score	Pearson's *r*	1		
	Sig (2-tailed)			
	N	2504		
Math Score	Pearson's *r*	**.890***	1	
	Sig (2-tailed)	.000		
	N	239	239	
Science Score	Pearson's *r*	**.945***	**.941***	1
	Sig (2-tailed)	.000	.000	
	N	2504	239	2504

***p < 0.001, two-tailed.

Reading proficiency was found to be highly correlated with mathematics and science scores. The correlations for all three countries were higher between reading and science (Canada, r = 0.933; Indonesia, r = 0.842; Singapore, r = 0.945) than reading and mathematics (Canada, r = 0.882; Indonesia, r = 0.805; Singapore, r = 0.890). This suggests that reading proficiency is associated with student performance, more so in science than in mathematics.

4.4 Performance Scores and Means Comparisons

4.4.1 Language at Home

The differential relationships between reading proficiency and performance on mathematics and science assessments, and how this relationship affects the comparability of scores for TLH and NTLH students were examined. The reading, mathematics, and science scores of TLH and NTLH students were calculated, and an independent samples *t*-test was used to test for statistical differences in the sample means of the language groups within countries. A *t*-test for equality of means was obtained, using a significance level of 0.05 as the threshold. The descriptive statistics and independent samples *t*-test results are summarized in Table 4.5.

Table 4.5

Performance Scores and Language Group Comparisons

Country	Subject	TLH			NTLH			*t*-Test	
		N	Mean	SD	N	Mean	SD	*t*	*p*
Canada	Reading	5374	520.96	85.41	806	511.51	87.35	2.92	***<0.05***
	Math	528	504.03	78.42	75	508.18	77.99	-0.43	<0.67
	Science	5374	521.94	87.95	806	510.37	90.67	3.47	***<0.001***
Indonesia	Reading	1295	413.81	76.82	2217	403.64	64.65	4.01	***<0.001***
	Math	278	398.77	77.51	430	385.18	68.29	2.39	***<0.05***
	Science	1295	411.60	70.14	2217	410.10	59.93	0.64	<0.50
Singapore	Reading	1203	551.49	87.46	1292	502.64	96.59	13.26	***<0.001***
	Math	116	574.07	79.88	123	530.56	95.94	3.82	***<0.001***
	Science	1203	570.91	94.28	1292	523.15	104.83	11.98	***<0.001***

Note. TLH, Test Language Speaker at Home; NTLH, non-Test Language Speaker at Home. *t*-Test value, "equal variances not assumed" when the *t*-Test value was found to be significantly different; "equal variances assumed" when the *t*-Test value was found to be insignificant.

Findings summarized in Table 4.5 indicate significant differences between TLH and NTLH students' reading scores in all three countries (Canada, $t = 2.92$, $p < 0.05$; Indonesia, $t =$

4.01, $p < 0.001$; Singapore, $t = 13.26$, $p < 0.001$). Statistically significant differences between the two language groups' mathematics scores were also found in Indonesia and Singapore (Indonesia, $t = 2.39$, $p < 0.05$; Singapore, $t = 3.82$, $p < 0.001$). Science scores between TLH and NTLH students were also significantly different in Canada and Singapore (Canada, $t = 3.47$, $p < 0.001$; Singapore, $t = 11.98$, $p < 0.001$). Where significant differences were identified, the TLH student group outperformed the NTLH student group.

4.4.2 Gender Group Comparisons

According to research, student characteristics such as gender can affect both English language and content assessment outcomes (Cole, 1997; Coley, 2001; Freeman, 2004; Hakuta, Butler, & Witt, 2000; Klecker, 2006). This section analyzed the performance scores of females and males in each of the three countries on three assessments, reading, mathematics, and science. Gender (ST004D01T), an independent variable where 1 = female and 2 = male, was used to indicate the gender of each respondent.

The differential relationships between reading proficiency and performance on mathematics and science assessments, and how this relationship affects the comparability of scores for female and male students were examined. The reading, mathematics, and science scores of female and male students were calculated. An independent samples *t*-test was used to test for statistical differences in the sample means, using a significance level of 0.05 as the threshold. The descriptive statistics and independent samples *t*-test results are summarized in Table 4.6.

Table 4.6

Gender Group Comparisons

Country	Subject	Female			Male			*t*-Test	
		N	Mean	SD	N	Mean	SD	*t*	*p*
Canada	Reading	3109	533.21	82.37	3239	505.25	86.72	13.18	***<0.001***
	Math	295	492.62	75.65	324	515.59	78.50	-3.70	***<0.001***
	Science	3109	519.26	86.16	3239	520.05	90.66	-0.35	<0.72
Indonesia	Reading	1862	418.53	67.78	1780	393.61	69.30	10.97	***<0.001***
	Math	377	394.15	73.66	368	386.27	70.17	1.49	<0.14
	Science	1862	411.98	64.14	1780	407.47	63.28	2.14	***<0.05***
Singapore	Reading	1225	535.69	91.38	1279	516.60	98.35	5.03	***<0.001***
	Math	114	548.30	82.10	125	554.76	98.61	-0.55	<0.58
	Science	1225	542.16	97.87	1279	549.76	106.89	-1.86	<0.06

Note. *t*-Test value, "equal variances not assumed" when the *t*-Test value was found to be significantly different; "equal variances assumed" when the *t*-Test value was found to be insignificant.

Findings summarized in Table 4.6 indicate statistically significant differences between female and male students' reading scores in all three countries (Canada, t = 13.18, p < 0.001; Indonesia, t = 10.97, p < 0.001; Singapore, t = 5.03, p < 0.001). Statistically significant differences were also found between female and male students' mathematics scores in Canada (Canada, t = -3.70, p < 0.001) and science scores in Indonesia (Indonesia, t = 2.14, p < 0.05). Female students in Canada performed better in reading as compared to male students (mean score: 533.21 versus 505.25), while male students outperformed female students in math (mean score: 515.59 versus 492.62). In Indonesia, where significant differences were identified, the female group outperformed the male group in both reading (mean score: 418.53 vs. 393.61) and science (mean score: 411.98 vs. 407.47). In Singapore, females performed better in reading when compared to males (mean score: 535.69 vs. 516.60).

4.4.3 Socio-economic Status (SES) Group Comparisons

The family socio-economic status was examined using the PISA's ESCS index of economic, social, and cultural status. PISA's ESCS index is derived from three standardized indicators, including parental education, highest parental occupation, and home possessions. Its high validity and reliability have been illustrated in the PISA 2015 Technical Report (OECD, 2017b), hence this derived variable of ESCS index was used directly in the present study to measure SES.

The ESCS scores were divided into three equal proportions; individuals with ESCS values in the top 33% were classified as "high SES" (coded as High ESCS=3), whereas individuals who had the bottom 33% ESCS values were classified as "low SES" (coded as Low ESCS =1). To show distinctly the ESCS groups, only the high SES and low SES groups were included in the analyses. The performance scores of high and low ESCS group were examined using an independent samples *t*-test, with a significance threshold level of 0.05. The *t*-test results are provided in Table 4.7.

The findings showed significant differences in mean scores between the low ESCS and high ESCS students for all three subjects. Reading scores were significantly different in all three countries (Canada, $t = -22.06$, $p < 0.001$; Indonesia, $t = -21.22$, $p < 0.001$; Singapore, $t = -22.56$, $p < 0.001$). Significant differences in mathematics scores between the low ESCS and high ESCS students were also found in Canada, Indonesia, and Singapore (Canada, t = -7.39, p < 0.001; Indonesia, t = -11.19, p < 0.001; Singapore, t = -6.06, p < 0.001). The same results were also noted with the science scores, which were significantly different for low and high ESCS students in Canada, Indonesia. and Singapore (Canada, t = -21.88, p < 0.001; Indonesia, t = -21.04, p < 0.001; Singapore, t = -21.95, p < 0.001). Where significant differences were identified, the high ESCS group outperformed the low ESCS group.

Table 4.7

SES (ESCS) Group Comparisons

Country	Subject	Low ESCS			High ESCS			*t*-Test	
		N	Mean	SD	N	Mean	SD	*t*	*p*
Canada	Reading	2054	492.14	82.38	2054	548.89	82.53	-22.06	**<*0.001***
	Math	209	476.90	74.01	197	533.26	79.65	-7.39	**<*0.001***
	Science	2054	492.80	84.58	2054	550.91	85.62	-21.88	**<*0.001***
Indonesia	Reading	1211	380.31	63.63	1212	438.27	70.61	-21.22	**<*0.001***
	Math	236	359.25	61.92	261	427.47	73.93	-11.19	**<*0.001***
	Science	1211	387.04	53.99	1212	439.55	68.14	-21.04	**<*0.001***
Singapore	Reading	831	478.98	89.95	831	575.52	84.46	-22.56	**<*0.001***
	Math	74	507.29	82.79	76	591.78	87.91	-6.06	**<*0.001***
	Science	831	496.12	98.27	831	597.83	90.50	-21.95	**<*0.001***

Note. Low ESCS, 33% lowest ESCS index values; High ESCS, 33% highest ESCS index values. *t*-Test value, “equal variances not assumed” when the *t*-Test value was found to be significantly different; “equal variances assumed” when the *t*-Test value was found to be insignificant.

4.5 Analysis of Covariance (ANCOVA)

In this study, reading scores served as the covariate (CV) for each of the group performance comparisons of TLH and NTLH students, male and female groups, and low and high ESCS values. Mathematics and science scores were the dependent variables (DV), while Language at Home was the independent variable (IV) that grouped and identified students as TLH or NTLH. Other independent variables included gender and ESCS. They will be discussed in subsequent sections.

4.5.1 Assumptions

To determine if the CV (reading scores) significantly interacts with the IV (language at home), as well as whether there are interactions between the language at home group and gender, and between the language at home groups and ESCS, an ANCOVA model including the IV, CV, and the interaction term between the IV and CV was tested. All the assumptions for ANCOVA, including linearity of regression, normality of residuals, homogeneity of variances, homogeneity of regression lines, and independence of error terms were tested. For those ANCOVA analyses where the assumption of uniformity of regression lines was violated, ANCOVA was performed separately for each level of the IV (i.e. separate model for TLH and NTLH). A significance level of $p<0.05$ was applied to all ANCOVA analyses when assessing for statistical significance (*p* value).

4.6 ANCOVA Results

4.6.1 Language at Home (Language Group)

The homogeneity of variances assumption for the language group was tested using ANCOVA analysis. The Levene's test of equality of error variance was found to be statistically insignificant for the language group in both dependent variables, indicating that the variance is homogeneous for all groups. The homogeneity of regression slopes assumption was also tested, and the results of the ANCOVA analyses are presented in Table 4.8.

Table 4.8

ANCOVA Results (Language Group)

Dependent Variable (DV)	Country	Variable	F	*p*	Partial Eta Square
Mathematics	Canada	Language Group*Reading Scores	0.77	0.38	0.001
		(IV) Language Group	1.57	0.21	0.003
		(CV) Reading Scores	904.22	0.00	0.602
	Indonesia	Language Group*Reading Scores	0.20	0.65	0.000
		(IV) Language Group	0.09	0.76	0.000
		(CV) Reading Scores	1269.75	0.00	0.643
	Singapore	Language Group*Reading Scores	0.01	0.92	0.000
		(IV) Language Group	0.01	0.91	0.000
		(CV) Reading Scores	768.58	0.00	0.766
Science	Canada	Language Group*Reading Scores	0.00	0.96	0.000
		(IV) Language Group	0.09	0.76	0.000
		(CV) Reading Scores	19315.03	0.00	0.758
	Indonesia	**Language Group*Reading Scores**	***3.92***	***0.05***	***0.001***
		(IV) Language Group	***8.25***	***0.00***	***0.002***
		(CV) Reading Scores	***8661.93***	***0.00***	***0.712***
	Singapore	Language Group*Reading Scores	2.39	0.12	0.001
		(IV) Language Group	1.59	0.21	0.001
		(CV) Reading Scores	19358.87	0.00	0.886

Note. ANCOVA, analyses of covariance; DV, dependent variables; IV, independent variables, CV, covariate. ANCOVA results highlighted in bold indicate that assumption of uniformity of regression slopes was not met.

Of the six models that were used in testing the homogeneity of regression slopes assumption, only one significant interaction between the IV and CV was identified. In Indonesia, the interaction for science scores as the DV (F (1, 3508) = 3.92, $p < 0.05$) was found to be significant, which suggested violation of the homogeneity of regression slopes assumption for ANCOVA analyses. In the remaining models, no significant differences in the interaction between the IV and CV were found.

For the model where homogeneity of regression slopes assumption was violated, further analyses were conducted. Instead of conducting the ANCOVA analysis across different language groups (TLH and NTLH), the relationship between the CV and the DV was examined separately for each level of the IV (i.e. TLH and NTLH groups were analyzed separately) (Ercikan et al, 2015). For analyses that met the homogeneity of regression slopes assumption, the ANCOVA models were applied without the CV and IV interaction term.

One purpose of this analysis is to examine the degree to which mathematics and science scores of TLH and NTLH might be impacted by their reading score. In Table 4.8, it can be seen that reading scores accounted for a large proportion of variance in both mathematics (up to 77%, since partial eta squared = 0.766) and science (up to 89%, since partial eta squared = 0.886). Like the findings reported by Ercikan and colleagues (2014), in the Canadian analyses reading scores contributed to more variance in science (76%, partial eta squared = 0.758) than in mathematics scores (60%, partial eta squared = 0.602). The same findings were found for Indonesia, where reading scores contributed to more variance in science (71%, partial eta squared = 0.712) than in mathematics (64%, partial eta squared = 0.643). Reading scores also contributed to more variance in science (89%, partial eta squared = 0.886) than in mathematics (77%, partial eta squared = 0.766) for Singapore. One possible explanation is the fact that science requires higher language (reading) demand.

The ANCOVA result for the model that violated the homogeneity of regression slopes assumption is summarized in Table 4.9. In Indonesia, the interaction between the IV and CV for science scores was found to be significant. This indicated that there were different associations between reading and science scores in Indonesia for TLH and NTLH groups. In the ANCOVA analyses, reading scores exhibited a stronger association with the DV science scores for the TLH group versus the NTLH group (76% versus 68%). In Table 4.10, the group means comparison showed that the TLH performed better than the NTLH.

Table 4.9

ANCOVA Results for the Model that Violated the Homogeneity of Regression Slopes Assumption (Language Group)

Dependent Variable (DV)	Country	Variable	F	*p*	Partial Eta Square
Science	Indonesia	Language Group (TLH)*Reading Scores	**4069.79**	***0.00***	**0.759**
		Language Group (NTLH)*Reading Scores	**4624.23**	***0.00***	**0.676**

Note. ANCOVA, analyses of covariance; DV, dependent variables; TLH, test language speakers at home; NTLH, non-test language speakers at home. Since the assumption of uniformity of regression slopes was not met for the model, a separate fit statistic, significance level, and effect size was generated for TLH and NTLH groups in Indonesia (Science).

Table 4.10

ANCOVA Results, Descriptive Statistics, and Mean Differences in Science Achievement in Indonesia (Language Group and Reading Scores)

Language	Science (DV)				
Group	Mean (SE)		SD	n	
TLH	411.60 (1.95)		70.14	1295	
NTLH	410.10 (1.27)		59.93	2217	
Source	SS	df	MS	F	p
Language Group	9671.60	1	9671.60	8.25	0.004
Reading Scores	10155146.19	1	10155146.19	8661.93	0.000
Language Group*Reading Scores	4594.53	1	4594.53	3.92	***0.048****
Error	4112740.93	3508	1172.39		

Note. ANCOVA, analyses of covariance; DV, dependent variables; TLH, test language speakers at home; NTLH, non-test language speakers at home. Since the assumption of uniformity of regression slopes was not met for the model, a separate fit statistic, significance level, and effect size was generated for TLH and NTLH groups in Indonesia (Science).

*$p < 0.05$.

4.6.2 Gender

The variation of degree of association between reading and mathematics and science scores was examined. The Levene’s test of equality of error variance was found to be statistically insignificant, suggesting a homogeneous variance for the gender groups in both dependent variables. In addition, a two-way ANCOVA analysis was performed to determine whether there is an interaction effect between language group (IV) and gender (IV) on the mathematics scores (DV) and science scores (DV) of TLH and NTLH, after controlling for reading scores (CV). The same assumptions and analyses as discussed in section 4.5.1 were applied, but with language at home and gender as the IVs.

An ANCOVA model including language group and gender as IV, reading score as CV, and the interaction term between language group and gender was tested. Further ANCOVA analysis was performed separately when the assumption of uniformity of regression lines was violated. The results of the ANCOVA analyses are presented in Table 4.11.

Of the six models that were analyzed using ANCOVA, a significant interaction between language group and gender was found in one of the models, which suggested a violation of the homogeneity of regression slopes assumption for ANCOVA analyses. In the Indonesian analyses, the interaction was found to be significant for mathematics scores as the DV (F (1,703) = 5.26, $p < 0.05$). This indicates that the mathematics scores of TLH and NTLH, when reading is taken into account, vary between males and females. No significant differences in the interaction between the gender and the IV were found in the remaining models.

Table 4.11

ANCOVA Results (Language Group and Gender)

Dependent Variable (DV)	Country	Variable	F	*p*	Partial Eta Square
Mathematics	Canada	Language Group*Gender	0.08	0.78	0.000
		(IV) Language Group	3.06	0.08	0.005
		(CV) Reading Scores	2818.73	0.00	0.825
	Indonesia	**Language Group*Gender**	***5.26***	***0.02***	***0.007***
		(IV) Language Group	***0.26***	***0.61***	***0.000***
		(CV) Reading Scores	***1360.25***	***0.00***	***0.659***
	Singapore	Language Group*Gender	0.29	0.59	0.001
		(IV) Language Group	1.53	0.22	0.006
		(CV) Reading Scores	884.72	0.00	0.791
Science	Canada	Language Group*Gender	2.55	0.11	0.000
		(IV) Language Group	4.07	0.04	0.001
		(CV) Reading Scores	52724.99	0.00	0.895
	Indonesia	Language Group*Gender	0.58	0.45	0.000
		(IV) Language Group	24.63	0.00	0.007
		(CV) Reading Scores	9289.03	0.00	0.726
	Singapore	Language Group*Gender	1.89	0.17	0.001
		(IV) Language Group	2.23	0.14	0.001
		(CV) Reading Scores	24111.35	0.00	0.906

Note. ANCOVA, analyses of covariance; DV, dependent variables; IV, independent variables; CV, covariate. ANCOVA results highlighted in bold indicate that assumption of uniformity of regression slopes was not met.

Further analysis was conducted on the model that violated the homogeneity of regression slopes assumption. For analyses that met the homogeneity of regression slopes assumption, the ANCOVA models were applied without the CV and IV interaction term.

The ANCOVA results for the model that violated the homogeneity of slopes assumption is summarized in Table 4.12. In Indonesia, the interaction between the IV and CV for mathematics scores was found to be significant. This indicated that the different associations of the reading and mathematics scores varied when the interaction between the language background and gender variables were examined. In the ANCOVA analyses, reading scores exhibited a slightly stronger association for female (68.4% accounting for variation) as compared to 64.3% for male students.

Table 4.12

ANCOVA Results for the Model that Violated the Homogeneity of Regression Slopes Assumption (Language Group and Gender)

Dependent Variable	Country	Variable	F	*p*	Partial Eta Square
Mathematics	Indonesia	Language Group*Reading Scores (Female)	**383.85**	**0.00**	**0.684**
		Language Group*Reading Scores (Male)	**311.97**	**0.00**	**0.643**

Note. ANCOVA, analyses of covariance; DV, dependent variables; Since the assumption of uniformity of regression slopes was not met for the model, a separate fit statistic, significance level, and effect size was generated for female and male groups in Indonesia (Mathematics).

4.6.3 Socio-economic Status

A two-way ANCOVA analyses of the IVs, CV, and DV, including the interaction term between language group (IV) and ESCS (IV) was tested to determine if the mathematics and science scores of TLH and NTLH also vary for ESCS when reading scores is taken into account. Further ANCOVA analysis was performed separately for TLH and NTLH groups of low ESCS and high ESCS when the assumption of uniformity of regression lines was violated. The homogeneity of variances assumption for the ESCS group was also tested and found to be statistically insignificant, suggesting a homogeneous variance for the ESCS groups in both dependent variables. The results are presented in Table 4.13.

No significant interactions between language group and ESCS were found in the six models that were analyzed using ANCOVA. This indicates that the mathematics scores of TLH and NTLH, when reading is taken into account, does not vary for low ESCS and high ESCS group.

Table 4.13

ANCOVA Results (Language Group and ESCS)

Dependent Variable (DV)	Country	Variable	F	*p*	Partial Eta Square
Mathematics	Canada	Language Group*ESCS	2.44	0.09	0.008
		(IV) Language Group	5.11	0.02	0.009
		(CV) Reading Scores	1924.01	0.00	0.764
	Indonesia	Language Group*ESCS	0.15	0.86	0.000
		(IV) Language Group	7.09	0.01	0.010
		(CV) Reading Scores	1028.11	0.00	0.595
	Singapore	Language Group*ESCS	0.77	0.46	0.007
		(IV) Language Group	3.12	0.08	0.013
		(CV) Reading Scores	735.26	0.00	0.760
Science	Canada	Language Group*ESCS	0.88	0.41	0.000
		(IV) Language Group	4.30	0.04	0.001
		(CV) Reading Scores	37659.03	0.00	0.860
	Indonesia	Language Group*ESCS	0.63	0.53	0.000
		(IV) Language Group	65.09	0.00	0.018
		(CV) Reading Scores	7247.19	0.00	0.674
	Singapore	Language Group*ESCS	0.24	0.87	0.000
		(IV) Language Group	0.02	0.89	0.000
		(CV) Reading Scores	16983.59	0.00	0.872

Note. ANCOVA, analyses of covariance; DV, dependent variables; IV, independent variables; CV, covariate. ANCOVA results highlighted in bold indicate that assumption of uniformity of regression slopes was not met.

4.7 Group Differences Adjusted for Reading Proficiency

4.7.1 Language Group

The adjusted and unadjusted means of mathematics and science scores for TLH and NTLH are summarized in Table 4.14. After statistically adjusting for reading scores using the ANCOVA model for each country, the scores for the NTLH adjusted means were higher than the TLH adjusted means in both mathematics scores and science scores. This finding was similar in all three countries, except for mathematics scores in Canada.

Table 4.14

Adjusted and Unadjusted Means for Each Language Group

Country	Language Group	Mathematics		Science	
		TLH	NTLH	TLH	NTLH
Canada	Unadj Mean (SE)	504.03 (3.41)	508.18 (9.01)	521.94 (1.20)	510.37 (3.19)
	Adj Mean (SE)	**503.19*** (1.60)**	**513.80*** (4.26)**	**518.26* (0.44)**	**520.75* (1.13)**
Indonesia	Unadj Mean (SE)	398.77 (4.65)	385.18 (3.29)	411.60 (1.95)	410.10 (1.27)
	Adj Mean (SE)	388.96 (2.61)	391.68 (2.09)	**406.49*** (0.96)**	**412.96*** (0.73)**
Singapore	Unadj Mean (SE)	574.07 (7.42)	530.56 (8.65)	570.91 (2.72)	523.15 (2.92)
	Adj Mean (SE)	548.20 (4.15)	555.13 (3.90)	545.44 (1.00)	547.40 (0.96)

Note. TLH, test language speakers at home; NTLH, non-test language speakers at home.

*$p < 0.05$. ***$p < 0.001$.

The adjusted mean scores reflected a different pattern of group differences between TLH and NTLH in different countries. In mathematics, the score gap between the TLH and NTLH groups stayed significant (p<0.001) only in Canada, in favor of the NTLH adjusted scores (even though they were not significant based on the unadjusted scores). In science, the adjusted differences between the TLH and NTLH remained significant for both Canada and Indonesia, even though they were also initially insignificant based on the unadjusted mean scores. The significant differences favored both the Canada NTLH group ($p<0.05$) and the Indonesia NTLH group ($p<0.001$).

When the unadjusted and adjusted means of TLH were compared, the mathematics and science scores did not change greatly for Canada, but changed a bit for Indonesia, and changed a lot for Singapore. The changes with TLH mean scores tend to be in the opposite direction, hence leading to lower scores in the adjusted means as compared to the unadjusted means. When the unadjusted and adjusted means of the NTLH group were compared, the changes for Canada and Indonesia was also found to be small, but the change in score gap for Singapore was found to be large. The change with the mean scores was also found to be in the opposite direction but leading to higher scores in the adjusted means as compared to the unadjusted means.

4.7.2 Gender Group

Another purpose of the study was to examine to what degree the relationship between reading ability and mathematics and science performance of TLH and NTLH varies for gender groups. Since only Indonesia demonstrated statistically significant interaction between gender and the language background, as explained earlier in section 4.6.2 (see Tables 4.11 and 4.12), the adjusted and unadjusted means for each language group by gender in mathematics and science for Indonesia are presented in Table 4.15.

The results indicated somewhat differential relationship between reading and mathematics and science scores for language group and gender. Except for the science scores of males, the TLH group (unadjusted means) outperformed the NTLH group for both gender groups. But when the reading scores were taken into account, the adjusted mean scores of both gender groups indicated higher performance for NTLH compared to TLH, except for mathematics scores in males.

Table 4.15

Adjusted and Unadjusted Means for Each Language Group by Gender (Indonesia)

		INDONESIA			
Gender	Language Group	Mathematics		Science	
		TLH	NTLH	TLH	NTLH
Female	Unadj Mean (SE)	399.08 (6.32)	390.48 (4.91)	414.88 (2.66)	411.37 (1.81)
	Adj Mean (SE)	**388.44* (3.39)**	**398.42* (2.92)**	**408.53*** (1.27)**	**415.44*** (1.02)**
Male	Unadj Mean (SE)	398.40 (6.89)	380.35 (4.41)	407.70 (2.86)	408.86 (1.79)
	Adj Mean (SE)	390.89 (3.80)	384.53 (2.83)	**405.28* (1.37)**	**410.14* (0.99)**

Note. TLH, test language speakers at home; NTLH, non-test language speakers at home.

*$p < 0.05$. ***$p < 0.001$

4.8 Chapter Summary

The purpose of this study was to examine the relationship between reading proficiency and performance on mathematics and science assessments, and how this relationship affects comparability of scores and score meaning for TLH and NTLH from Canada, Indonesia, and

Singapore. The study also investigated to what degree this relationship varies for gender and socio-economic groups.

The TLH groups for all countries exhibited higher mean scores in all three subjects compared to their NTLH counterparts, except for math in Canada and Science in Indonesia where the differences were not significant. In addition, reading proficiency was found to be highly correlated with mathematics and science scores. A fairly strong positive correlational relationship (+/- 0.80 or higher) was found between reading and math scores, reading and science scores, and math and science scores for all three countries. As reading score increases, math and science scores also increase. And as math score increases, science score also tends to increase. The correlations for all three countries were higher between reading and science than between reading and mathematics.

When the performance scores were compared, the TLH and NTLH students' reading scores in all three countries revealed significant differences. The mathematics scores of TLH and NTLH were significantly different in Indonesia and Singapore, while science scores were significantly different in Canada and Singapore. Where significant differences were found, the TLH group outperformed the NTLH group.

Significant differences were also noted between female and male students' reading scores in all three countries, mathematics scores in Canada, and science scores in Indonesia. Female students outperformed the male students, except for Canada where male students performed better in mathematics. When ESCS group comparisons were made, the findings revealed significant differences in the mean scores between the low ESCS and high ESCS students in all three subjects. This result was similar in all three countries. Where significant differences were identified, the high ESCS group outperformed the low ESCS group.

The results from the ANCOVA analyses indicated that reading proficiency accounts for a large proportion of variance in both mathematics and science. In mathematics, across all three countries, reading proficiency accounted for as high as 77% of the variance (ranging between 60% and 77%). In science, reading proficiency accounted for approximately 89% of the variance (ranging between 71% and 89%) in Canada, Indonesia, and Singapore. Reading scores contributed to more variance in science than in mathematics, and the trend was similar across the three countries. In Canada, reading scores contributed to as high as 76% of the variance in science while in mathematics it only accounted for 60% of the variance. In Indonesia, there was more variance in science (71%) than in mathematics (64%). In Singapore, the variance in science was 89% while in mathematics it was 77%.

While the interaction effects between language groups and reading scores were insignificant in Canada and Singapore, the interaction effect between language group and reading scores was significantly different in Indonesia. This finding points to differences in the relationship between reading proficiency and performance on science assessment of TLH and NTLH in Indonesia, with the results indicating a higher association for TLH. This result supported previous research on the effects of reading proficiency on mathematics and science assessment performance (e.g., Abedi & Gandara, 2006; Hudson, Lane, & Pullen, 2005; Noble et al., 2014). The findings also provided an estimate of the size of possible language background effects on student performance in Indonesia.

The interaction effects of language group (TLH and NTLH) and ESCS in all countries were also analyzed, and no significant differences in the interaction between language group and ESCS were identified. This indicates that the differential relationship in the reading score and performance assessment between TLH and NTLH does not vary for ESCS.

There is evidence across the three countries that group differences in mathematics and science scores are different when the students' reading proficiency levels are taken into account. After statistically adjusting for reading scores, the NTLH adjusted means were higher in both mathematics and science in all three countries compared to the TLH adjusted means, even though they were initially lower than TLH based on the unadjusted mean scores except for mathematics in Canada. The varying patterns of performance difference in mathematics and science revealed that scores for NTLH students may be underestimated, a phenomenon that has already raised serious concerns (Abedi & Gandara, 2006; Korpershoek, Kuyper, & van der Werf, 2014; Solano-Flores, 2008).

Since Indonesia demonstrated statistically significant difference in the language group and gender interaction (see Tables 4.11 and 4.12), further analyses on the group means of TLH and NTLH were made by looking at the interactions between language group and gender. In Indonesia, the adjusted mean scores of females were lower in TLH than in NTLH for both mathematics and science. The significance favored the NTLH female groups for both subjects. For males, the adjusted mean was also higher for NTLH in science, but lower in mathematics. The score gap was significant only in science, favoring NTLH male students.

In the present study, when the group means were adjusted to take reading proficiency into account, the adjusted and unadjusted mean scores for both mathematics and science showed different results. Even though the relationship between reading proficiency and performance on mathematics and science assessments identified in the ANCOVA revealed significant differences in the unadjusted mean scores between TLH and NTLH in all three countries, the pattern of the differences pointed to different results when reading proficiency scores were adjusted. A similar pattern was also noted in the interaction effects between language group and gender, and between language group and ESCS in Indonesia.

The findings in the study confirmed that the reading scores reported in PISA 2015 may not truly reflect the performance of the three selected countries. More specifically, the reading scores may have been underestimated due to the presence of sizable proportions of non-speakers of test language (NTLH) who are, by default, second-language speakers where the test language is concerned since it is not the language they speak at home. The NTLH's linguistic background may distort the mean scores of their performance to varying extent. Since ANCOVA captures correlational associations but not causal relationships, the results need to be interpreted with caution. The reading performance that accounts for variation in science and mathematics scores may not necessarily indicate that reading proficiency led to the achievement gap differences identified in the study.

CHAPTER 5

DISCUSSION

5.1 Context for the Research

Today's knowledge-based economy driven by globalization, advances in information and communication technologies, and reduced trade barriers have changed the type of knowledge and skills future economy requires. There is a much higher demand for the next generation to develop a strong set of foundational skills upon which further learning can be built. Many international large-scale surveys are used worldwide to test the scholastic performance on mathematics, science, and reading of students.

The OECD has developed a common tool – PISA – to measure the extent to which youth, at age 15, have acquired some of these knowledge and skills that are essential for full participation in modern societies. PISA is designed to provide international indicators of the skills and knowledge of 15-year-old students, and to shed light on a range of factors that contribute to successful students, schools, education systems, and society (OECD, 2016a). It measures skills that are generally recognized as key outcomes of the educational process (OECD, 2016a).

PISA, administered by OECD, is a wide-scale, evaluative study used to compare the education systems of countries. It focuses on the skills of students in using their accumulated knowledge to solve problems they might face in real life rather than assessing their academic knowledge in relation to a particular school curriculum (OECD, 2016a, 2016b). The PISA survey gathers information for a thorough comparative analyses of student performance near the end of their compulsory education. It also permits exploration of the ways these student performances vary across different social and economic groups, as well as identify factors that influence achievement within and between countries. PISA has brought significant public and educational attention to international assessments and has generated data that enhanced policy makers' abilities to formulate decisions based on evidence.

5.2 Overview of Research Purpose

Since 2000, the OECD has organized PISA assessment rounds for member countries, non-member countries, and partner countries, to assess school-system performance of upper secondary students in three core subjects, mathematics, reading, and science. Due to the rigour of the background analytic reports prepared by PISA, and its large sample size, PISA has become one of the benchmarks in making international comparisons.

For large scale assessments like PISA, as a normal practice, data are collected by administering tests to students in their respective language of instruction – the language officially sanctioned as the country's medium of teaching, e.g., English or French in Canada; Bahasa Indonesia in Indonesia; and English in Singapore. However, as the language used for teaching and assessment may not be the same as the home or first language of all students, then their performances may be affected by their proficiency in the language of the test. Variations in student linguistic backgrounds may therefore distort the assessment results to varying degrees and the results may not be accurate indicators of the students' performances (Abedi, Urrutia, & Shneyderman, 2005).

The main purpose of this study was to investigate the association of reading proficiency and performance on Math and Science assessments using PISA 2015. Specifically, there were three research questions that guided the study: 1) To what degree do students' reading abilities

(PISA Reading Score) account for their math performance (PISA Math Score)? 2) To what degree do students' reading abilities (PISA Reading Score) account for their science performance (PISA Science Score)? 3) To what degree does the relationship between reading ability and mathematics and science performance vary for gender and socio-economic groups?

This chapter begins with a summary of the major findings addressing the three research questions. The implications of these findings are also discussed, as well as the limitations of the current research. The chapter concludes with future research directions.

5.3 Summary of Findings Addressing Research Questions 1 and 2

In examining the relationship between reading proficiency and performance in mathematics and science, the study hypothesized that speaking the test language at home or not influences their performance in PISA 2015. ANCOVA analyses were conducted based on language group, i.e. TLH students who speak the language of assessment at home, and NTLH students whose home language is not the language of the PISA assessment in their respective countries of residence.

ANCOVA analyses revealed that reading scores accounted for a large proportion of variance in the mathematics (up to 77%) and science scores (up to 89%) in all three countries When the interaction effects between reading scores and language groups were examined, the results revealed significant difference in the relation of reading scores with science scores of TLH and NTLH in Indonesia; the results exhibited a much stronger association for TLH group. This finding further supported previous research on the effects of reading proficiency on mathematics and science assessment performance (e.g., Abedi & Gandara, 2006; Bussiere, Knighton, & Pennock, 2001; Elosua & De Boeck, 2020; Ercikan et al., 2015; Hudson, Lane, & Pullen, 2005; Leung, 2014; Noble et al., 2014).

When the group mean differences were adjusted to take reading proficiency into account, the adjusted and unadjusted mean scores for both mathematics and science showed different results. With the exception of mathematics scores in Canada and science scores in Indonesia, even though TLH had higher unadjusted mathematics and science scores than NTLH, the findings pointed to the opposite direction when reading proficiency scores were taken into account. The adjusted mean scores reflected a different pattern of group differences; the NTLH group was found to have higher scores in both mathematics and science in each country compared to the TLH. The score differences in mathematics between TLH and NTLH stayed significant for Canada, while in science it was significant for Canada and Indonesia; the differences favored the NTLH group.

The varying patterns of performance difference in mathematics and science point to an underestimation of the scores for NTLH due to language demands of the assessment, a phenomenon that has already raised serious concerns in other research (e.g. Abedi, 2001; Ercikan et al., 2015; Smith, Frazier, Lee, & Chang, 2018; Soh, 2011, 2014; Solano-Flores, 2008).

5.4 Summary of Findings Addressing Research Question 3

Additional ANCOVA analyses were conducted, with reading scores as a covariate, to determine whether the association of reading with Mathematics and Science performance varied for gender and SES groups, when TLH and NTLH were compared.

Although gender differences have been observed in assessments (Cole, 1997; Coley, 2001; Freeman, 2004; Logan & Johnston, 2010), from the ANCOVA results in the present study, the interaction effect between the language group and gender was only found to be significant in the mathematics performance in Indonesia. That is, in Canada and Singapore, the relationship between reading proficiency and Math and Science performance was similar across the gender groups when the TLH and NTLH language groups were compared. In Indonesia, the differential relationship of reading with performance in Math and Science for gender groups revealed that the TLH and NTLH comparisons varied when reading proficiency was taken into account. For females, when reading proficiency was taken into account, NTLH outperformed the TLH group in both Mathematics and Science. Whereas for the males in Indonesia, the performance comparison for the language groups was reversed only for Science, with NTLH outperforming TLH.

On the other hand, the ANCOVA analysis on the interaction effect between TLH and NTLH and SES groups revealed no significant interactions, indicating that the relationship between reading ability and mathematics and science performance for the three countries does not vary with SES groups.

5.5 Implications of the Research Findings

The conceptual grounding of this study was informed by the literature in the research domains of ecological framework, second language acquisition theory, and componential model of reading. These theories also guided the interpretations of the findings and the implications of the research findings as discussed in subsequent sections.

5.5.1 Bronfenbrenner's Macro- and Micro-systems Ecological Framework as Validity Evidence

Bronfenbrenner (1976) argued that traditional experimental research involves largely unfamiliar, artificial, and unnecessary rigid environments for both the researcher and the research subject, which could obscure key findings. He identified the importance of using an ecological framework to understand the nested and complex relationships among the various systems that play a role in human development. Bronfrenbrenner (1977, 1995) conceptualized the ecological environment as multiple, nested levels of environmental influence, and proposed a framework that viewed the progression of human development as evolving through increasingly complex interactions between the organism and its ecological environment.

Bronfrenbrenner's ecological framework underwent two major significant changes since its first inception in the 1970s. In the initial phase, he defined ecological theory (Bronfenbrenner, 1974) as the study of human development in contextual environments, which included the upper layer of immediate settings such as home and school, and the supportive layer such as the geographical and institutional contexts (social systems). He further expanded his layer model into a more complex series of systems interdependent of each other. In this revised framework (Bronfenbrenner, 1977), the ecological environment became a nested arrangement of structures, each contained within the next. These structures were referred to as the macrosystem, exosystem, mesosystem, and the microsystem. In the 1990s, more revisions were made, namely the general ecological model or bioecological paradigm (Bronfenbrenner & Ceci, 1994), the process-person-context-time model, and inclusion of the new structure 'chronosystem' (Bronfenbrenner, 1995).

For the purpose of this study, Bronfenbrenner's model was used to understand how the students' performances in PISA 2015 were impacted by factors at both macro and micro system levels. In Bronfenbrenner's model, the *macrosystem* level is defined as the overarching level, representative of the economic, social, and political systems that is exerting influence on the students. The *microsystem* level is defined as the activities and interpersonal relations experienced by the student on a day-to-day basis, including interactions with family and peers. Both macrosystem and microsystem were included in guiding the conceptualization of this study.

Student achievements on international tests differ sharply across countries. Students in high performing countries such as Canada and Singapore score consistently higher across several international comparisons than students from low performing countries such as Indonesia, in mathematics, reading, and science (Chiu & Joh, 2015; OECD, 2011). These differences in academic achievement might stem from several factors, such as the country's economic and cultural differences. Studies using PISA data have shown that there are substantial differences in student achievement at the country macrosystem and the individual microsystem (Chiu & Khoo, 2005). Both are discussed below.

A macrosystem represents the cultural and political objectives that drive the educational priorities of a country (Bronfenbrenner, 1995). At the national (macrosystem) level, these initiatives reflect the interplay of cultural, economic, and social beliefs, and the political objectives that emerge from these beliefs. For instance, the countries with higher average test scores are generally much wealthier (e.g., as measured by gross domestic product (GDP) per capita) and have greater levels of equity (e.g., income) than the countries with lower average test scores. Students in countries with greater economic productivity also show higher academic achievement (Chiu & McBride-Chang, 2010). Their study revealed that the country's resources provide students with more learning opportunities, and governments tend to invest more in education by building more educational infrastructure and providing better educational resources. These differences suggest that both greater educational resources and more equal distribution of resources contribute to greater student learning and academic achievement.

A case in point is Singapore. Being a multi-ethnic economy, Singapore has a peculiar situation when language is concerned. Singaporean students from single or mix families normally speak another language other than English at home. This means that the assessment language of PISA, English, is not the home language for a considerable number of the Singapore sample. However, despite this, Singapore has been consistently among the top ranking countries on PISA. One possible explanation for this success is that the education system and initiatives (macrosystem) in Singapore may have provided rich learning opportunities to students, preparing them to become bilingual, self-directed learners of the 21st century (Cheung, Sit, Soh, Leong, & Mak, 2014; Deng & Gopinathan, 2016).

In Indonesia, although students who completed the PISA assessment in their home language performed better than students who were non-speakers of the test language at home, in general, Indonesia was still below the OECD average in all three subjects. One possible explanation for this phenomenon is that Indonesia's educational system (macrosystem) has been reported to be inadequate (Hidayat & Patras, 2013). Another reason is that the teaching and learning outcomes were considered misaligned with curriculum goals (Sukyadi & Mardiani, 2011). These concerns can act as a catalyst in driving policy at the macrosystem level, a policy that recognizes the importance of providing quality curricula and meaningful learning outcomes for the students in the 21st century.

On the other hand, in the microsystem, the focus is on the student and the interaction s/he has with the environment, peers, teachers, and family. The microsystem level involves all the relationships that interact directly with the student. At the school microsystem, students in richer countries often have more school resources, and have teachers with higher qualifications and better training. The teachers also use effective teaching methods to create a supportive classroom atmosphere. These advantages help students in privileged schools learn better as compared to their counterparts in poorer schools/countries. In Indonesia, learning inequality is evident between urban and rural schools. Schools in rural and remote areas have very few facilities, which may have limited their students' range of educational opportunities. Rural schools also tend to have relatively untrained teachers (or unmotivated to remain in such rural areas), hence teaching and learning is often of low quality.

At the home microsystem, many Singaporeans see family as the backbone of society. Parents maintain a culture where talent can be developed through hard work and perseverance, thus children in Singapore are expected to work hard on academics from an early age. Families have moral obligations to develop every child to their fullest potential, and to have respect for learning and scholarship. The focus is to perform well academically in order to "have a better future".

At the classroom microsystem, the schools within each country are culturally diverse, with a mix of students from different language backgrounds, which could fuel differences in learning opportunities and student achievement. For example, students in richer countries not only showed higher academic achievement, stronger links between family support and academic performance were also reported in some studies (Chiu, 2007, 2010; Chiu & Chow, 2010). Canada is one of the most diverse countries in the world, featuring a highly dispersed population, two official languages, and a large number of immigrants. Despite such diversity in language background, Canada's overall PISA scores has consistently ranked well above the OECD average on the three subjects. The students who took the English version of the PISA survey were also mostly speakers of the test language at home. Such diversity in the students' educational and skills profile has led to higher performance in the assessment.

The findings in this study highlighted contextual and socio-cultural differences at both macrosystem and microsystem levels for the three countries. This study also illustrated how the macrosystem and microsystem structures in Bronfrenbrenner's ecological framework can facilitate a better understanding of student performance in PISA 2015 (microsystem) within the greater context of the national educational priorities (macrosystem). The framework has allowed for a clear description of the relationships between a combination of contextual variables, such as home language, gender and SES, and test performance, in a systematic way. The current findings have also made it evident that multiple factors across multiple ecological contexts can directly or indirectly affect the academic achievement in students. Thus, both the macrosystem and microsystem of one's ecology should be tapped to understand how students' academic achievement develops, and study the extent to which the two levels interact, and which combination of factors best support academic achievement.

5.5.2 Language Effects Explained Using Cummins' Second Language Acquisition Theory

The second language acquisition theory by Cummins (2009) identifies two types of language proficiency: conversational (Basic Interpersonal Communicative Skills or BICS) and academic (Cognitive Academic Language Proficiency or CALP). Conversational proficiency is the ability to converse in the language used in social and personal situations, and can take two to three years to acquire (Cummins, 1980, 1983). Academic proficiency, which includes reading literacy, is the ability to understand the academic language predominantly used in schools and official settings, and can take seven to ten years of consistent effort in strong academic environments to acquire (Cummins, 1981a, 1981b, 1981c, 2009; Suarez-Orozco, Suarez-Orozco, & Todorova, 2008). Conversational language and academic language are not always acquired to the same level of proficiency (Krashen, 2000); in fact, the time to develop academic language proficiency is much longer than the conversational language proficiency (Cummins, 2009).

The relationship between reading proficiency and academic achievement has been well documented. Cummins, as early as 1981, argued that ELL students need to master the academic English skills needed to understand the content, and communicate that content knowledge on standardized tests (Cummins, 1981a, 1981b). Abedi and Gandara (2006) added that grade school students can take approximately five to seven years to acquire English language proficiency. If one does not have adequate competency in the language of the test, it becomes challenging for them to perform well on the test especially when it involves reading and writing. Brown (2005), and Rabinowitz and colleagues (2005), studied the effects of language proficiency on standardized tests, and they found that as language proficiency increases, the standardized test score also increases.

PISA administers tests in the students' medium of instruction, with a tacit assumption that the students are proficient in that language, and that their performance would not be adversely affected (Soh, 2014). However, as shown by the findings in this study, speaking the test language at home or not is strongly associated with the student performance in mathematics and science assessment. In fact, the study findings indicate that the test language of PISA may pre-condition the level of performance of students who are assessed via the test language. These language challenges can attenuate students' opportunities to demonstrate what they know. Without sufficient command of the test language, the student may not respond to the test question correctly even if they know the correct answers. PISA also involves students who speak the test language at home as well as the non-speakers. These students may therefore have an advantage over others, which may have implications for interpretation of assessment results.

The present study elucidated the problem of test language effect in international assessments, using PISA 2015 reading scores to illustrate. The proportion of NTLH varied widely among the three countries, and hence the magnitude of the effect also varied. When the proportion of NTLH in the sample is large, like in Indonesia, the effect of underestimation of mathematics and science scores at the country level may be larger than compared to the other two countries (Soh, 2014). One way to adjust for this type of underestimation is to adjust the estimation by controlling for reading scores (the 'covariate' as in this study). Another way of adjusting for the underestimation is to take into account the proportion of test language speakers versus non-speakers, and then combine the scores of TLH and NTLH duly weighted by their respective proportions in the overall sample (Soh, 2014).

Cummins' theory was relevant to the goals of the present study. His BICS/CALP distinction (Cummins, 1981a) helped explain the possibility that some second language learners or language minority students may take longer than 7 years to approach normal levels of CALP. He also showed that bilingual students were unfairly disadvantaged in literacy test results, hence assessments like PISA may have created academic deficits because they failed to distinguish between L2 BICS and L2 CALP (Cummins, 1983). With the increasing presence of language diversity in schools, Cummins' theory provided additional contextualization of the present study.

5.5.3 Home Environment as Guided by the Componential Model of Reading Framework

The *Componential Model of Reading* (CMR) (Aaron , Joshi, Gooden, & Bentum, 2008) consists of three primary domains, namely, the cognitive, psychological, and ecological domains. According to the CMR model, reading performance is influenced by the factors included in the Simple View of Reading (SVR) cognitive domain (Hoover & Gough, 1990) as well as factors from the psychological and ecological domains. The Simple View of Reading describes reading comprehension as the product of decoding and linguistic comprehension component abilities (Gough & Tunmer, 1986). The psychological domain includes teacher and student variables such as locus of control, learning styles, motivation and interest, and gender differences (Aaron et al, 2008). The ecological factors include characteristics of students' home environment, such as families' socio-cultural status, number of books in the home, and home literacy environment and dialects (Chiu, McBride-Chang, & Lin, 2012).

In a longitudinal study on the emergence of reading comprehension problems in L1 and L2, Kieffer (2010) found L2 students and students from low SES homes to be overrepresented in the group of struggling readers. In most OECD countries, students with immigrant background and indigenous languages and cultures have a more disadvantaged socio-economic background than their native peers (OECD, 2010b), hence partially explaining for their low performance level in reading and other domains. Jerrim (2012) investigated the association of SES (represented by parental occupation) and reading achievement for different performance levels across Australia, Canada, Finland, Germany, the UK, and the USA. He found that the association was stronger in the USA and the UK than in most other countries, and that it is particularly strong at high performance levels.

Cross-country research by Martins and Veiga (2010) used a multi-dimensional SES measure including, for example, parents' education, and books and home possessions. They, too, showed that SES gaps in math achievement were particularly large in the UK, Germany, Belgium, Greece, and Portugal but much lower in Sweden and Finland. They added that inequalities decreased in Germany and Spain but increased in France and Italy. Chiu et al (2012) also tested the CMR model using PIRLS 2006 data, and found that the ecological domain comprising variables such as the school climate, attitude about reading, number of books at home, socio-economic status, and resources explained the greatest variance in reading comprehension at the country level (61%).

The psychological and ecological domains of the CMR framework were used to help understand and explain the indirect effects and contributions, if any, of home literacy environments, gender differences, and socio-economic status to the reading abilities of the TLH and NTLH groups.

5.5.4 Measurement Comparability

The final implication of the research findings indicates the importance of the type of analyses that should be conducted at the test level for examining comparability of performance results across countries. Previous research has demonstrated differences in measurement incomparability identified at item and test levels, such as those identified by confirmatory factor analyses and DIF analysis (Byrne, 2008; Byrne, & van de Vijver, 2010, 2014; Ercikan & Koh, 2005; Oliveri, Ercikan, Zumbo & Lawless, 2014). Ercikan and colleagues (2015) demonstrated the limitations in validity of comparisons, when the *population heterogeneity* of language groups is not properly taken into account.

The results in this study point to the possibility of using ANCOVA analyses as an alternative approach to examining measurement comparability at the test level (Rheinheimer, & Penfield, 2011; Shieh, 2020). Evidence of differential score meaning and comparability for TLH and NTLH mathematics and science assessments highlighted the importance of minimizing the effects of reading on assessments, as an inaccurate estimation of student performance in mathematics and science assessment, especially for the NTLH group, can impact the accuracy of group comparisons (e.g. Ercikan & Solano-Flores, 2014; Hu & Bentler, 1999). Furthermore, the findings indicate that comparability of findings need to be examined and interpreted with more caution. If the comparability of scores is not demonstrated, then the intention to compare TLH and NTLH language groups may not yield accurate results for meaningful inferences. For example, decisions concerning educational policies, comparisons of student achievement across language groups, and program / evaluation development may not be done appropriately.

The results of this study highlight the necessity to look at the possible role of contextual factors at both macro and micro levels, such as educational policies and socio-cultural background, language effect on performance in math and science assessments, home environments such as gender and socio-economic status, and measurement comparability when assessment results of TLH and NTLH groups are compared. The current study integrated theories from the ecological framework, second language acquisition, reading model, and measurement concepts in order to gain a more holistic understanding of the association between reading proficiency and performance in mathematics and science assessments of TLH and NTLH students from Canada, Indonesia, and Singapore. These theories were also utilized in guiding the interpretations of the findings and future recommendations.

5.6 Contributions of the Dissertation to PISA Research Literature

The purpose of this study is to examine the comparability of mathematics and science scores among students who speak the test language at home and those who are non-speakers of the test language at home. The study also investigates whether the association of reading language proficiency and performance on mathematics and science assessments in Canada, Indonesia, and Singapore varies with gender and socio-economic groups. The results of this dissertation have implications for educational policy and practice relating to using ecological framework as validity evidence, test language effect, and measurement comparability.

The results of this dissertation made several contributions to PISA research literature. First, although linguistic abilities of students seemed to affect their performance on assessments, the use of ecological framework suggested that the process may be far more complex than just the language competencies of the students alone. A possibility exists for academic performance to be interrelated with other factors, including educational policies and initiatives, academic literacy level, gender, and SES, which could produce confounding effects.

Next, the dissertation findings also provided some clarity around test language effect on student performance in math and science, for native speakers as well as ELLs and language minority students, who are either test language speakers or non-test language speakers at home. In previous research on examining the effects of test language, researchers have compared PISA results of ELLs and language minority students using samples drawn primarily from English-speaking countries. The present study not only looked at an English-speaking country, such as Canada, but also at a country in which English is no the major academic and assessment language, such as Indonesia. Moreover, the study also investigated a country where students are using both English and a Mother Tongue Language as the lingua franca, such as Singapore.

In addition, whereas other studies looked at the effect of test language *between* and *within* countries and systems, using country as the unit of analysis and country means as the 'country score' for comparisons, the current study only looked at the effect country by country, and found that the use of the 'country score' may not be entirely applicable at the student level. The findings suggested that there is a systematic relationship between being able to speak the test language at home and performance on the assessment across the three selected countries. The varying patterns of performance difference in mathematics and science also pointed to an under-reporting of the scores for NTLH due to language demands of the assessment.

Therefore, the country scores (and the rankings based on these) are likely to be confounded by the presence of NTLH in the national samples. Ranking countries on their ‘country score’, with the presence of NTLH, could lead to questions regarding the validity of the reading means as truly reflecting students’ reading competency. Conducting a multi-level analysis that separates country effects and student effects for a deeper understanding of test language effect may be an area worth investigating in future.

Lastly, the results of the dissertation are likely to be useful in informing the research on measurement comparability in PISA assessment. Specifically, the results in the present study indicated the importance of examining measurement comparability to better understand the score differences of student groups in the respect subject domains. The fact that the group means were different / reversed when reading scores (as covariate) were controlled underscores the importance of investigating for measurement comparability before making any claims about the score differences. Given that the PISA reading literacy test is often used as a basis for numerous comparisons of youth across the globe, this study added more evidence to the importance of determining consistency in the findings to arrive at more definitive and accurate results.

5.7 Limitations

The first limitation is due to the use of reading proficiency as the only variable for examining potential language effects. Writing proficiency in English, measured using constructed response items, may be another subskill that can be used to determine performance assessments. The inclusion of writing proficiency as an additional covariate (CV) may account for greater degrees of variation in mathematics and science assessments.

Second, using only one variable ‘international language at home’ to determine language grouping has low measurement accuracy. In PISA 2015, students were asked to report what language they spoke at home, using a single item in the questionnaire [#22, International Language at Home, *“which language do you speak at home?”*]. Based on the answers, the samples for this study were divided into TLH and NTLH. The self-reporting of data by 15-years-olds of their most-commonly spoken language at home may have included some inaccuracies, as their responses can easily be influenced by how they understand the question, by their perceptions, or by social desirability bias. Thus, by relying on a single self-report indicator used in PISA, it is reasonable to expect some level of discrepancy in the language groupings. Such low reliability in reported language used at home may have added unintended variability to the reading proficiency levels of TLH and NTLH. This can lead to underestimation of correlational relationships between reading ability and mathematics and science scores.

5.8 Future Directions

Educational policymaking is becoming increasingly influenced by standardized global comparisons, assessments, and policy borrowing (Sivesind, Afsar & Bachmann, 2016). Due to PISA, national curricula are now beginning to resemble one another with regards to curricular subjects, content, and skills (Zilliacus, Holm & Sahlstrom, 2017). The results of PISA have stirred discussions in many countries and have been of great concerns to researchers about how well they are preparing students for life in an increasingly technological world in which increasing levels of literacy are needed (e.g. Zhao, 2020).

There are several different areas of research that deserve further study.

First, the effectiveness of efforts on minimizing test language effects on mathematics and science performance may vary for TLH and NTLH students, therefore limitations of score meaning and how scores are reported need to be addressed. One suggested approach is to include measurement error due to language group, i.e. language can be considered as a source of such measurement error. Including error as part of reporting is likely to lead to more cautious interpretation of scores. For example, small performance differences identified as statistically insignificant between TLH and NTLH may change to a statistically significant difference when measurement error due to language (e.g. reading scores) is taken into consideration. As a result of the inclusion of measurement error due to language, changes in scores will undoubtedly affect the outcome, either as a gain or drop in performance.

Another area that needs to be addressed is how the students are grouped using the variable 'International Language at Home'. The groupings of the students into TLH and NTLH based on one question in PISA assessment can impact the accuracy of group comparison as well as overall results. Asking students to self-report whether the most-commonly used language at home was the test language or not may also lead to measurement inaccuracies; the possibility of a discrepancy between what is reported and what is the actual case can exist. One suggestion involving NTLH groupings is to measure and report the language proficiency of these students along with their mathematics and science scores. In OECD's framework (OECD, 2016b), Level III is considered a key measure of success in PISA and is often used as a benchmark for national standards in several countries (Fleckenstein et al., 2016; OECD, 2011). Reading proficiency at this level is characterized by Bussiere and colleagues (2001) as having the ability to "compare, contrast, and categorize competing information according to a range of criteria" (p. 24). The language proficiency at Level III needs to be established using empirical evidence before findings can be meaningfully interpreted. For example, students with scores below the identified 'meaningful' level threshold will be considered inappropriate in measuring mathematics and science knowledge and skills and will therefore be excluded when comparing TLH and NTLH performance.

What constitutes 'International Language at Home' is also another area of future research. Central in this notion is the fact that language is not fixed and clear-cut, but rather it is a fluid construct which incorporates multiple levels of complexity. One suggestion is to provide a clear definition of the variable 'international language at home' to address the increasing linguistic diversity brought about by globalization. There needs to be a clear definition of language, e.g. 'home language', 'mother tongue', 'immigrant language', and 'indigenous language'. As students with different languages and identities interact in a culturally transforming society, a move from a restricted definition of language towards a pluralistic approach is needed now more than ever. Policies and practices that respect the language of linguistic minority students, while finding common ground between the home and mainstream culture, remains important.

The findings from the study underscore the importance of identifying the elements and relations within multiple ecological levels to describe and establish the condition of the learning ecology of the student. One suggestion is the use of the ecological model of Bronfenbrenner to systematically analyze the horizontal and vertical interactions of factors affecting student achievement. Such approach can lead to a comprehensive understanding of the dynamics of the interactions between students and their multi-level interacting environments. Future comparative work on student achievement should consider country-level predictors in a more extensive way to help explain further the patterns of educational achievement in each country.

Although PISA test items go through multiple phases of piloting and field testing, OECD does not report any efforts on minimizing language burden on the examinees. Another direction for future research can be the inclusion of language and cultural perspectives, or reviews for language complexity, in the test development. ANCOVA or other measurement comparability statistical analyses needs to be complemented by other sources of validity evidence such as those based on student response processes. For example, cognitive interviews that focus on examinee response processes may help identify problems with assessment tasks or determine how overall language of the assessment may be affecting NTLH performance (see Noble et al., 2014). Such approaches are useful in determining linguistic and cultural aspects of items that may contribute to measurement error for NTLH students.

Lastly, the trend towards multiculturalism and multilinguality may have implications in determining what should be regarded as appropriate patterns of school organizations in different societies, indicating the need for cross-cultural aspects in educational assessment and policy. One suggestion is to take into consideration the influence of the socio-cultural dimensions at the national level to make the educational system in each country more effective and socially responsive. As 21st century students live in an interconnected, diverse, and rapidly changing world, developing an assessment with a global and intercultural outlook is both an opportunity and a challenge. Different approaches to teaching, learning, and assessment can only be truly understood when placed in their respective cultural context.

5.9 Conclusion

There is growing evidence that limited language proficiency has significant implications on students' success in mathematics and science assessment. The present study investigated the association between reading ability and mathematics and science performance of TLH and NTLH students (research questions 1 and 2). It also investigated whether such association varies for gender and SES (research question 3). The present study allowed for a deeper understanding of reading proficiency effect on mathematics and science performance of 15-year-old students from Canada, Indonesia, and Singapore. Findings from the study have shown the importance of reading ability in mathematics and science assessments. The results revealed reading proficiency to be crucial for both mathematics and science performance; students' mathematics and science scores were influenced by their level of reading ability. Although gender and SES differences have been observed in a number of cross-sectional studies (Cole, 1997; Coley, 2001; Freeman, 2004), these differences did not result in differential relationship between reading proficiency and Math and Science performance in the present study. The association of reading ability with mathematics and science performance of TLH and NTLH was different only in males and females' mathematics performance in Indonesia. The association of reading ability and mathematics and science performance did not vary for SES in all three countries.

The results in the present study also offered explanations and contributed information to the comparability and consistency of score meaning, which is central to the validity of score interpretations. Based on the results of this study, it can be concluded that caution must be exercised when comparing language groups within countries, and therefore in making comparisons across countries. The findings in this study highlight the point that inferences made from the ANCOVA analyses which include reading as a covariate are likely to result in more appropriate score interpretations. Given that the PISA reading literacy test is often used as a basis for numerous comparisons of youth across the globe, this study added more evidence to the importance of comparability of assessment results within and across countries.

REFERENCES

Aaron, P.G., Joshi, R.M., Gooden, R., & Bentum, K.E. (2008). Diagnosis and treatment of reading disabilities based on the component model of reading. *Journal of Learning Disabilities, 41*, 67-84.

Abedi, J. (2001). Validity consideration in the assessment of LEP students using standardized achievement tests. Paper presented at *The 2001 Annual Meeting of the American Educational Research Association, Seattle, Washington.* Retrieved from: https://www.researchgate.net/publication/234675898_Validity_Considerations_in_the_Assessment_of_LEP_Students_Using_Standardized_Achievement_Tests

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and ability issues. *Educational Researcher, 33*(1), 4-14.

Abedi, J. (2008). Measuring students' level of English proficiency: Educational significance and assessment requirements. *Educational Assessment*, *13*(2-3), 193-214.

Abedi, J., & Gandara, P. (2006). Performance of English language learners as a subgroup in large-scale assessment: Interaction of research and policy. *Educational Measurement: Issues and Practice*, *25*(4), 36-46.

Abedi, J., & Herman, J.L. (2010). Assessing English language learners' opportunity to learn Mathematics: Issues and limitations. *Teachers College Record*, *112*(3), March 2010, 723-746.

Abedi, J., Herman, J.L., Courtney, M., Leon, S., & Kao, J. (2004). *English language learners and math achievement: A study on classroom-level opportunity to learn.* Los Angeles:

University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (2001). *NAEP mathematics performance and test accommodations: Interactions with student language background* (CSE Technical Report 536). Retrieved from: http://cresst.org/wp-content/uploads/newTR536.pdf

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, *74*(1), 1-28.

Abedi, J., Leon, S., & Mirocha, J. (2001). Validity of standardized achievement tests for English language learners. Paper presented at the *American Educational Research Association Conference,* Seattle, WA, April 10-14, 2001. Retrieved from: https://files.eric.ed.gov/fulltext/ED455292.pdf

Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student's language background on content-based assessment: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing. Retrieved from: https://pdfs.semanticscholar.org/6e2d/b6b2ef14393500a0ddb50b93151e1ba7ec79.pdf

Abedi, J., & Linquanti, R. (2012). Issues and opportunity in improving the quality of large-scale assessment systems for English language learners. In K. Hakuta & M. Santos (Eds.), *Commissioned papers on language and literacy issues in the Common Core State Standards and Next Generation Science Standards, April 5, 2012* (pp. 75-85). Palo Alto: Stanford University Press. Retrieved from:

https://mes.sccoe.org/resources/ALI%202012/11_KenjiUL%20Stanford%20Final%205-9-12%20w%20cover.pdf

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, *14*(3), 219-234.

Abella, R., Urrutia, J., & Shneyderman, A. (2005). An examination of the validity of English language achievement test scores in an English Language Learner population. *Bilingual Research Journal*, *29(1),* 127-144.

Alt, M., Arizmendi, G.D., & Beal, C.R. (2014). The relationship between mathematics and language: Academic implications for children with specific language impairment and English language learners. *Language, Speech, and Hearing Services in Schools*, *45,* 220-233.

American Education Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Ammermueller, A. (2013). Institutional features of schooling systems and educational inequality: Cross-country evidence from PIRLS and PISA. *German Economic Review*, *14*(2), 190-213.

Ammermueller, A., & Pischke, J.S. (2009). Peer effects in European primary schools: Evidence from the progress in international reading literacy study. *Journal of Labor Economics*, *27*(3), 315-348.

Anderson, J.O., Lin, H.S., Tregust, D.F., Ross, S.P., & Yore, L.D. (2007). Using large-scale assessment datasets for research in science and mathematics education: Programme for

International Student Assessment (PISA). *International Journal of Science and Mathematics Education*, *5*(4), 591-614.

Asil, M., & Brown, G.T. (2016). Comparing OECD PISA reading in English to other languages: Identifying potential sources of non-invariance. *International Journal of Testing*, *16*(1), 71-93.

August, D., & Hakuta, K. (1997). *Improving schooling for language-minority children: A research agenda.* Washington, DC: National Academy Press.

Aydin, A., Erdag, C., & Tas, N. (2011). A comparative evaluation of PISA 2003-2006 results in reading literacy skills: An example of top five OECD countries and Turkey. *Educational Sciences: Theory and Practice*, *11*(2), 664-673.

Bailey, A., & Butler, F.A. (2007). A conceptual framework of academic English language for broad application to education. In A. Bailey (Ed.), *The language demands of school: Putting academic English to the test.* New Haven, CT: Yale University Press.

Bailey, A., Maher, C., & Wilkinson, L.C. (2018). *Language, literacy, and learning in the STEM disciplines: Language counts for English learners.* New York, NY & Oxford, UK: Routledge Taylor Francis.

Baker, D.P., Goesling, B., & Letendre, G.K. (2002). Socioeconomic status, school quality, and national economic development: A cross-national analysis of the "Heyneman-Loxley effect" on mathematics and science achievement. *Comparative Education Review*, *46*, 291-312.Baumgartner, H., & Steenkamp, J.B.E.M. (1998). Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. *Marketing Letters*, *9*, 21-35. Retrieved from: doi:10.1023/A:1007911903032.

Benoliel, P., & Berkovich, I. (2018). A cross-natural examination of the effect of the Schwartz cultural dimensions on PISA performance assessments. *Soc Indic Res,* 139, 825-845.

Bourdieu, P. (2011). The form of capital. *Cultural Theory: An anthology, 1*, 81-93.

Broadfoot, P. (2007). *An introduction to assessment.* London: Continuum International Publishing Group.

Bronfenbrenner, U. (1974). Developmental research, public policy, and the ecology of childhood. *Child Development*, *45,* 1-5.

Bronfenbrenner, U. (1976). The experimental ecology of education. *Educational Researcher*, *5*(9), 5-15.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*, 513-531.

Bronfenbrenner, U. (1995). Developmental ecology through space and time: A future perspective. In P. Moen, G.H. Elder. Jr., & K. Luscher (Eds.), *Examining lives in context: Perspectives on the ecology of human development* (pp. 617-649). Washington, DC: American Psychological Association.

Bronfenbrenner, U., & Ceci, S.J. (1994). Nature-nurture reconceptualized in developmental perspective: A bio-ecological model *Psychological Review, 101*, 568-586.

Brosseau, L., & Dewing, M. (2009). Canadian multiculturalism: Background paper. Canadian Government: Parliament of Canada. Retrieved from: https://lop.parl.ca/sites/PublicWebsite/default/en_CA/ResearchPublications/200920E

Brown, C.L. (2005). Equity of literacy-based math performance assessments for English language learners. *Bilingual Research Journal*, *29*(2), 337-364.

Brown, J.D. (2014). The future of world Englishes in language testing. *Language Assessment Quarterly*, *11*(1), 5-26.

Bussière, P., Knighton, T., & Pennock, D. (2001). *Measuring up: The performance of Canada's youth in reading, mathematics, and science (p. 24)*. Ottawa: Statistics Canada.

Byrne, B.M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*, 872-882.

Byrne, B.M., & van de Vijver, F.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of non-equivalence. *International Journal of Testing*, *10*, 107-132.

Byrne, B.M., & van de Vijver, F.R. (2014). Factorial structure of the family values scale from a multilevel-multicultural perspective. *International Journal of Testing*, *14*(2), 168-192.

Cavallaro, F., & Ng, B.C. (2014). Language in Singapore: From multilingualism to English plus. In J. Hajek & Y. Slaughter (Eds.), *Challenging the Monolingual Mindset,* Chapter 2, pp. 33-48. Bristol: Multilingual Matters. https://www.degruyter.com/view/book/9781783092529/10.21832/9781783092529-005.xml

Chall, J.S. (1996). *Stages of reading development* (2nd Ed.). Fort Worth, TX: Harcourt-Brace.

Chee, M.T., Chin, T.Y., Loh, M.Y., Ng, H.L., Chew, L.P., Sin, K.H., Tang, S.Y., Tay, W.B., & Yen, Y.P. (2016). Singapore. In V.S. Mullis, M.O. Martin, S. Goh, & K. Cotter (Eds.), *TIMSS 2015 Encyclopedia: Education policy and curriculum in mathematics and science.* Retrieved from: http://timssandpirls.bc.edu/timss2015/encyclopedia/.

Chen, Y.F., & Jiao, H. (2014). Exploring the utility of background and cognitive variables in explaining latent differential item functioning: An example of the PISA 2009 Reading assessment. *Educational Assessment*, *19*(2), 77-96. Retrieved from: https://www.researchgate.net/publication/262336014_Exploring_the_Utility_of_Background_and_Cognitive_Variables_in_Explaining_Latent_Differential_Item_Functioning_An_Example_of_the_PISA_2009_Reading_Assessment/download

Cheung, K.C., Sit, P.S., Soh, K.C., Leong, M.K., & 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-Pacific Education Researcher*, *23*(4), 895-909.

Chiu, M.M. (2007). Families, economies, cultures, and science achievement in 41 countries: Country, school, and student level analyses. *Journal of Family Psychology*, *21*, 510-519.

Chiu, M.M. (2009). Inequalities' harmful effects on both disadvantaged and privileged students: Sources, mechanisms, and strategies. *Journal of Education Research*, *3*, 109-128.

Chiu, M.M. (2010). Inequality, family, school, and mathematics achievement. *Social Forces*, *88*(4), 1645-1676.

Chiu, M.M., & Chow, B.W.Y. (2010). Culture, motivation, and reading achievement: High school students in 41 countries. *Learning and Individual Differences*, *20*, 579-592.

Chiu, M.M., & Joh, S.W. (2015). Cross-national comparisons in Education: Findings from PISA. *International Encyclopedia of the Social and Behavioral Sciences,* 2nd Ed., Vol. 5, 342-348. Retrieved from: http://dx.doi.org/10.1016/B978-0-08-097086-8.92144-5

Chiu, M.M., & Khoo, L. (2005). Effects of resources, inequality, and privilege bias on achievement: country, school, and student level analyses. *American Educational Research Journal*, *42*, 575-603.

Chiu, M.M., & McBride-Chang, C. (2010). Family and reading in 41 countries: Differences across cultures and students. *Scientific Studies of Reading*, *14*, 514-543.

Chiu, MM., McBride-Chang, C., & Lin, D. (2012). Ecological, psychological, and cognitive components of reading difficulties: Testing the component model of reading in fourth graders across 38 countries. *Journal of Learning Disabilities*, *45*, 391-405.

Clements, D.H., & Sarama, J. (2008). Experimental evaluation of the effects of a research-based preschool mathematics curriculum. *American Educational Research Journal*, *45*(2), 443-494. Retrieved from: https://psycnet.apa.org/record/2008-07769-007.

Cocking, R.R., & Chipman, S. (1988). Conceptual issues related to mathematics achievement of language minority children. In R.R. Cocking & J.P. Mestre (Eds.), *Linguistic and cultural influence on learning mathematics* (pp. 17-46). Hillsdale: Erlbaum.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd Ed.)*. Hillsdale, NJ: Erlbaum.

Cole, N. (1997). *The ETS gender study: How females and males perform in educational settings.* Princeton, NJ: Educational Testing Service.

Coley, R.J. (2001). *Differences in the gender gap: Comparisons across racial / ethnic groups in education and work.* Princeton, NJ: Educational Testing Service.

Collier, V.P. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly*, *21*(4), 617.

Cope, M., & Kalantzis, M. (2012). *Multiliteracies: Literacy learning and the design of social futures (2nd Ed).* New York: Routledge.

Council of Ministers of Education Canada (CMEC) (2016). *Protocol for agreements for minority-language education and second language instruction: 2013-2014 to 2017-2018.* Canada: Council of Ministers of Education, Canada. Retrieved from: https://www.cmec.ca/docs/programsInitiatives/olp/protocol/Protocol-2013-2018-EN.pdf

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed)., *Intelligence: Measurement theory and public policy* (pp. 147-171). Urbana, IL: University of Illinois Press.

Cronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52*(4), 281-302.

Cummins, J. (1979). Linguistic interdependence and educational development of bilingual children. *Review of Educational Research*, *49*, 222-251.

Cummins, J. (1980). The construct of proficiency in bilingual education. In J.E. Alatis (Ed.), *Schooling and language minority students: A theoretical framework* (pp. 81-103). Washington, DC: Georgetown University Press.

Cummins, J. (1981a). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics*, *2*, 132-149.

Cummins, J. (1981b). *Bilingualism and minority language children.* Ontario, Canada: Ontario Institute for Studies in Education.

Cummins, J. (1981c). Empirical and theoretical underpinnings of bilingual education. *Journal of Education*, *163*, 16-29.

Cummins, J. (1983). Language proficiency and academic achievement. In J. Oiler (Ed.), *Issues in language testing research* (pp. 97-132). Rowley, RA: Newbury House.

Cummins, J. (2009). Putting language proficiency in its place: Responding to critiques of the conversational/academic language distinction. Retrieved from: https://esltaggart.files.wordpress.com/2013/04/research-article_-j-cummins.pdf

Darling-Hammond, L. (2010). *The flat world of education: How America's commitment to equity will determine our future.* New York: Teachers College Press.

Deng, Z., & Gopinathan, S. (2016). PISA and high performing education systems: Explaining Singapore's education success. *Comparative Education*, *52*(4), 449-472. Retrieved from: https://doi.org/10.1080/03050068.2016.1219535

Dharma, S. (2014). Tregedi nol buku: Tragedi di dunia pendidikan Indonesia. Retrieved from: https://satriadharma.com/2014/01/28/tragedi-nol-buku-tragedi-di-dunia-pendidikan-indonesia/ (abstract in English).

Dilas, D.B., Mackie, C., Huang, Y., & Trines, S. (2019). Education in Indonesia. World Education News and Reviews. Retrieved from: https://wenr.wes.org/2019/03/education-in-indonesia-2

Dixon, Q., Zhao, J., Shin, J., Wu, S., Su, J., Burgess-Brigham, R., Gezer, M., & Snow, C. (2012). What we know about second language acquisition: A synthesis from four perspectives. *Review of Educational Research*, *82*(1), 5-60. Retrieved from: https://journals.sagepub.com/doi/abs/10.3102/0034654311433587

Droop, M., & Verhoeven, L. (2003). Language proficiency and reading ability in first- and second-language learners. *Reading Research Quarterly*, *38*, 78-103. Retrieved from: https://ila.onlinelibrary.wiley.com/doi/abs/10.1598/RRQ.38.1.4

Duran, R.P. (1989). Testing of linguistic minorities. In R. Linn (Ed.), *Educational measurement.* 3rd ed. (pp.573-587). New York, NY: Macmillan.

Durrant, G.B. (2009). Imputation methods for handling item non-response in practice: Methodological issues and recent debates. *International Journal of Social Research Methods*, *12*(4), 293-304.

Ellis, E. (2004). The invisible multilingual teacher: The contribution of language background to Australian ESL teachers' professional knowledge and beliefs. *The International Journal of Multilingualism*, *1*(2), 90-108.

Elosua, P., & De Boeck, P. (2020). Educational assessment issues in linguistically diverse contexts: A case study using a generalised linear mixed model. *Language, Culture, and Curriculum*, *33*(3), 305 – 318.

Ercikan, K., & Koh, K. (2005). Examining the construct comparability of the English and French versions of TIMSS. *International Journal of Testing*, *5*(1), 23-35.

Ercikan, K., & Solano-Flores, G. (2014). Introduction to the special issue: Levels of analysis in the assessment of linguistic minority students. *Applied Measurement in Education*, *27*(4), 233-235. Retrieved from: https://www.tandfonline.com/doi/abs/10.1080/08957347.2014.944462

Ercikan, K., Chen, M.Y., Lyons-Thomas, J., Goodrich, S., Sandilands, D., Roth, W.M., & Simon, M. (2015). Reading proficiency and comparability of mathematics and science scores for students from English and non-English backgrounds: An international perspective. *International Journal of Testing, 15(2),* 153-175. Retrieved from: https://doi.org/10.1080/15305058.2014.957382

Ercikan, K., McCreith, T., & Lapointe, V. (2005a). Factors associated with mathematics achievement and participation in advanced mathematics courses: An examination of gender differences from an international perspective. *School Science & Mathematics*, *105*(1), 5-14. Retrieved from: https://doi.org/10.1111/j.1949-8594.2005.tb18031.x

Ercikan, K., McCreith, T., & Lapointe, V. (2005b). How are non-school related factors associated with participation and achievement in science? An examination of gender differences in Canada, the USA, and Norway. In S.J. Howie & T. Plomp (Eds.*), Contexts of learning mathematics and science: Lessons learned from TIMSS* (pp. 211-225). Lisse, Netherlands: Swets & Zeitlinger.

Ercikan, K., Oliveri, M., & Sandilands, D. (2013). Large-scale assessments of achievement in Canada. In J. Hattie & E.M Anderman (Eds.), *International guide to student achievement* (pp. 456-459). Ohio, USA: Routledge.

European Commission (EC) (2012). *Foreign language learning statistics.* EU: Eurostat. Retrieved from: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Foreign_language_learning_statistics

Faulkner-Bond, M., & Sireci, S.G. (2015). Validity issues in assessing linguistic minorities. *International Journal of Testing*, *15*(2), 114-135.

Feniger, Y., & Lefstein, A. (2014). How not to reason with PISA data: An ironic investigation. *Journal of Education Policy, 29*(6), 845-855.

Fischman, G.E., & Topper, A.M. (2017). *An examination of the influence of international large-scale assessments and global learning metrics on national school reform policies.* CASGE Working Paper #2. Tempe, Arizona: Arizona State University, The Center for Advances Studies in Global Education. Retrieved from: https://education.asu.edu/sites/default/files/casge_working_papers_2_updated.pdf

Fleckenstein, J., Leucht, M., Pant, H.A., & Koller, O. (2016). Proficient beyond borders: Assessing non-native speakers in a native speakers' framework. *Large-scale Assessments in Education*, *4*(1), 1-19.

Freeman, C.E. (2004). *Trends in educational equity of girls and women.* Washington, DC: National Center for Education Statistics, Institute of Education Studies, US Department of Education.

Frenette, M. (2007). Why are youth from lower-income families less likely to attend university? Evidence from academic abilities, parental influences, and financial constraints. *Analytical Studies Branch Research Paper Series*, *295.* Statistics Canada Catalogue No. 11F0019M. Ottawa: Statistics Canada.

Gandara, P., Rumberger, R., Maxwell-Jolly, J., & Callahan, R. (2003). English learners in California schools: Unequal resources, unequal outcomes. *Education Policy Analysis Archives*, *11(36).* Retrieved from: http://epaa.asu.edu/epaa/v11n36/

Geva, E. (2006). Second language oral proficiency and second language literacy. In D. August and T. Shanahan (Eds.), *Developing literacy in second language learners: Report of the national literacy panel on language minority children and youth (*pp. 123-139). Mahwah, NJ: Lawrence Erlbaum Associates.

Goh, C.B., & Gopinathan, S. (2008). Education in Singapore: Development since 1965. In B. Fredriksen & J.P. Tan (Eds.), *An African exploration of the East Asian education* (pp. 80-108). Washington, DC: The World Bank.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioural Statistics*, *32*(3), 252-286.

Gottardo, A., & Mueller, J. (2009). Are first- and second-language factors related in predicting second-language reading comprehension? A study of Spanish speaking children acquiring English as a second language from first to second grade. *Journal of Educational Psychology*, *101*, 330 – 344.

Gough, P.B., & Tunmer, W.E. (1986). Decoding, reading, and reading disability. *Remedial & Special Education*, *7*, 6-10. Retrieved from: https://journals.sagepub.com/doi/10.1177/074193258600700104

Grabe, W., & Stoller, F. (2013). Teaching reading for academic purposes. In M. Celcel-Murcia, D., Brinton, & M.A. Snow (Eds.), *Teaching English as a second or foreign language* (4th Ed). Boston: Heinle Cengage.

Graddol, D. (2006). *English next.* London: The British Council.

Grisay, A., de Jong, H.A.L., Gebhardt, E., Breezier, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement*, *8*(3), 249-266. Retrieved from: https://www.acer.org/files/grisay.pdf

Gupta, A.F. (1998). The situation of English in Singapore. In J.A. Foley, T. Kandiah, L. Wee, B. Zhiming, and A. Fraster-Gupta (Eds.), *English in New Cultural Contexts: Reflections from*

*Singapore (*pp. 106-126). Oxford: Oxford University Press. https://www.degruyter.com/view/title/535789?tab_body=toc

Haertel, E.H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, *18*(4), 5-9.

Hakuta, K., Butler, Y., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Policy Report 2000-1. University of California: Linguistic Minority Research Institute. Retrieved from: https://files.eric.ed.gov/fulltext/ED443275.pdf

Haladyna, T.M., & Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-27.

Hambleton, R.K., Merenda, P., & Spielberger, C. (2005). *Adapting educational and psychological tests for cross-cultural assessment.* Hillsdale, NJ: Lawrence Erlbaum.

Henson, R. K. (1998). ANCOVA with intact groups: Don't do it! Paper presented at *the annual meeting of the Mid-South Educational Research Association*, November, New Orleans, LA.

Hidayat, R., & Patras, Y.E. (2013). Evaluasi system Pendidikan nasional Indonesia. *2nd International Seminar on Quality and Affordable Education* (ISQAE 2013). Indonesia, January 2012.

Hill, C., & Parry, K. (2014). *From testing to assessment: English as an international language.* (2nd Edition). New York: Routledge.

Hoover, W.A., & Gough, P.B. (1990). The simple view of reading. *Reading and Writing: An Interdisciplinary Journal*, *2*, 127-160.

Hopfenbeck, T.N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the programme for international student assessment. *Scandinavian J of Educational Research*, *62*(3), 333-353.

Hopmann, S.T. (2008). No child, no school, no state left behind: Schooling in the age of accountability. *Journal of Curriculum Studies*, *40*(4), 417-456.

Horton, N.J., & Kleinman, K.P. (2007). Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *The American Statistician*, *61*(1), 79-90.

Hu, L.T., & Bentler, P.M. (1999). Cut-off criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. Retrieved from: https://www.tandfonline.com/doi/abs/10.1080/10705519909540118

Huang, X., Wilson, M., & Wang, L. (2016). Exploring plausible causes of differential item functioning in the PISA science assessment: Language, curriculum, or culture. *Educational Psychology*, *36*(2), 378-390.

Hudson, R.F., Lane, H.B., & Pullen, P.C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher*, *58*(8), 702–714.

Ilma, Z., & Pratama, R.K. (2015). Transformation in Indonesian language curriculum: Pros and cons between KTSP 2006 and Curriculum 2013 in Indonesia. Presented at *International Conference on Trends in Economics, Humanities and Management, March 27-28, 2015, Singapore*. Retrieved from: https://icehm.org/upload/5097ED0315007.pdf

International Test Commission (ITC) (2017). The ITC guidelines for translating and adapting tests. Second Edition. Retrieved from: https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf.

International Test Commission (ITC) (2018). ITC guidelines for the large-scale assessment of linguistically and culturally diverse populations. Retrieved from: https://www.intestcom.org/files/guideline_diverse_populations.pdf

Jang, E.E.H., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *J of Educational Measurement, 44(1),* 1-21. Retrieved from: https://doi.org/10.1111/j.1745-3984.2007.00024.x

Jerrim, J. (2012). The socio-economic gradient in teenagers' reading skills: How does England compare with other countries? *Fiscal Studies*, *33*(2), 159-184.

Karoly, L., & Cannon, J. (2007). *Achieving success for all students; A statewide initiative on closing the achievement gap.* USA: The RAND (Labour and Population) Corporation.

Kieffer, M.J. (2010). Socioeconomic status, English proficiency, and late-emerging reading difficulties. *Educational Researcher*, *39*, 484-486. Retrieved from: https://www.researchgate.net/publication/242700060_Socioeconomic_Status_English_Proficiency_and_Late-Emerging_Reading_Difficulties/download

Klecker, B.M. (2006). The gender gap in NAEP fourth-, eighth-, and twelfth-grade reading scores across years. *Reading Improvement*, *43*(1), 50-56.

Klinger, D.A., De Luca, C., & Miller, T. (2008). The evolving culture of large-scale assessments in Canadian education. *Canadian Journal of Educational Administration, and Policy*, *76*(3), 1-34.

Klinger, D.A., Maggi, S., & D'Angiulli, A. (2011). School accountability and assessment: Should we put the roof up first? *The Educational Forum*, *75*(2), 114-128.

Kocay, V.G. (2019). Evaluating best practices in large scale assessment: Defining the purpose of province-wide assessment programs in Canada. *Edcan Network,* March 12, 2019. Retrieved from: https://www.edcan.ca/articles/large-scale-assessment/

Koh, K., & Paris, S.G. (2011). Developing new reading assessments to promote beginning reading in Singapore. *Asia Pacific Education Review*, *12*, 23-33.

Komatsu, H., & Rappleye, J. (2017). A new global policy regime founded on invalid statistics? Hanushek, Woessmann, PISA, and economic growth. *Comparative Education, 53*(2), 166-191.

Kopriva, R.J. (2000). *Ensuring accuracy in testing for English language learners.* Washington, DC: Council of Chief State School Officers. Retrieved from: https://eric.ed.gov/?id=ED454703

Korpershoek, H., Kuyper, H., & Van Der Werf, G. (2014). The relation between students' math and reading ability and their mathematics, physics, and chemistry examination grades in secondary education. *International Journal of Science and Mathematics Education, 13,* 1013-1037.

Krashen, S. (2000). Bilingual education, the acquisition of English and the retention and loss of Spanish. In A. Roca (Ed.), *Research on Spanish in the US: Linguistic issues and challenges.* Somerville, NA: Cascadilla Press.

Kuo, E. (1980). The sociolinguistic situation in Singapore. Unity in diversity. In A.E. Afendras & E. Kuo (Eds.), *Language and society in Singapore (*pp. 39-62). Singapore: NUS Press.

Kurniawati, S., Suryadarma, D., Bima, L., & Yusrina, A. (2018). Education in Indonesia: A white elephant? *Journal of Southeast Asian Economies*, *35*(2), 185-199. Retrieved from: https://muse.jhu.edu/article/702081

Lauder, A. (2008). The status and function of English in Indonesia: A review of key factors. *Makara, Sosial Humaniora*, *12*(1), Jul, 9-20.

Lee, S.K., Goh, C.B., & Birger, F. (2008). *Toward a better future: Education and training for economic development in Singapore since 1965*. Washington, DC: The World Bank. Retrieved from: http://siteresources.worldbank.org/INTAFRREGTOPEDUCATION/Resources/444659-1204656846740/4734984-1212686310562/Toward_a_better_future_Singapore.pdf

Leung, F.K.S. (2014). What can and should we learn from international studies of mathematics achievement? *Mathematics Education Research Journal*, *26*, 579-605. Retrieved from: https://link-springer-com.ezproxy.library.ubc.ca/content/pdf/10.1007%2Fs13394-013-0109-0.pdf

Lingard, B. (2010). Policy borrowing, policy learning: Testing times in Australian schooling. *Critical Studies in Education, 51*(2), 129-147.

Logan, S., & Johnston, R. (2010). Investigating gender differences in reading. *Educational Review*, *62,* 175-187. Retrieved from: https://doi.org/10.1080/00131911003637006

Makel, M., & Plucker, J. (2014). Facts are more important than novelty: Replication in the education sciences. *Educational Researcher, 43,* 304-316.

Martiniello, M. (2008). Language and the performance of English language learners in math word problems. *Harvard Educational Review*, *78*(2), 333-368.

Martiniello, M. (2009). Linguistic complexity, schematic representations, and differential item functioning for English language learners in math tests. *Educational Assessment*, *14*(3-4), 160-189.

Martins, L., & Veiga, P. (2010). Do inequalities in parents' education play an important role in PISA student mathematics achievement test score disparities? *Economics of Education Review*, *29*(6), 1016-1033.

Marx, A., Stanat, P., Roick, T., Segerer, R., Marx, P., & Schneider, W. (2015). Components of reading comprehension in adolescent first-language and second-language students from low-track schools. *Read Write, 28*, 891 – 914. Retrieved from: DOI 10.1007/s11145-015-9554-3

Maxwell, S.E., O'Callaghan, M.F., & Delaney, H.D. (1993). Analysis of covariance. In L.K. Edwards (Ed.), *Applied analysis of variance in behavioural science (pp. 63-104).* New York, NY: Marcel Dekker.

Menken, K. (2008). English learners left behind: Standardized testing as language policy. Clevedon, UK: Multilingual matters. Retrieved from: http://journals.sagepub.com/doi/10.1177/0265532211404200

Menken, K. (2009). No Child Left Behind and its effects on language policy. *Annual Review of Applied Linguistics, 29,* 103-117. Retrieved from: https://doi.org/10.1017/S0267190509090096

Messick, S. (1989a). Validity. In R.L. Linn (Ed.), *Educational measurement.* 3rd edition (pp. 13-104). New York: Macmillan.

Messick, S. (1989b). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18(2),* 5-11.

Ministry of Education (MOE Singapore) (2014). *Information sheet of 21st century competencies.* Retrieved from: https://www.moe.gov.sg/news/press-releases/information-sheet-on-21st-century-competencies.

Ministry of Education (MOE Singapore) (2016a). *Education Statistics Digest 2016.* Retrieved from: https://www.moe.gov.sg/docs/default-source/document/publications/education-statistics-digest/esd-2016.pdf

Ministry of Education (MOE Singapore) (2016b). *Secondary school education: Shaping the next phase of your child's learning journey.* Retrieved from: https://www.moe.gov.sg/docs/default-source/document/education/secondary/files/secondary-school-education-booklet.pdf

Mislevy, R.J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56, 177-196.

Monseur, C., & Adamas, R. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement, 10*(3), 320-334.

Moore, K., Glei, D., Driscoll, A., Zaxlow, M., & Redd, Z. (2002). Poverty and welfare patterns: Implications for children. *Journal of Social Policy*, *31*(2), 207-227.

Nardi, E. (2008). Cultural biases: A non-Anglophone perspective. *Assessment in Education: Principles, Policy & Practice*, *15*(3), 259-266.

National Mathematics Advisory Panel (NMAP) (2008). *Foundations for success: The final report of the National Mathematics Advisory Panel.* Washington, DC: US Department of

Education. Retrieved from:
https://www2.ed.gov/about/bdscomm/list/mathpanel/report/final-report.pdf

Newkirk-Turner, B.L., & Johnson, V.E. (2018). Curriculum-based language assessment with culturally and linguistically diverse students in the context of mathematics. *Language, Speech, & Hearing Services in Schools*, *49*(2), 189. Retrieved from: https://doi.org/10.1044/2017_LSHSS-17-0050

Ng, E.H. (2008). Educational reform in Singapore: From quantity to quality. *Educational Research for Policy and Practice, 7*(1), 5-15.

Ni Riordain, M., & O'Donoghue, J. (2009). The relationship between performance on mathematical word problems and language proficiency for students learning through the medium of Irish. *Educational Studies in Mathematics*, *71*(10), 43-64.

No Child Left Behind (NCLB) (2002). *No Child Left Behind* Act of 2001. Conference Report to Accompany H.R., 1, Rep. No. 107-334, House of Representatives, 107th Congress, 1st Session, December 13, Pub. L. No. 107-110, 115 Stat. 1425.

Noble, T., Risebery, A., Suarez, C., Warren, B., & O'Connor, C. (2014). Science assessments and English language learners: Validity evidence based on response processes. *Applied Measurement in Education*, *27*, 248-260.

Oliveri, M.E., & Ercikan, K. (2011). Do different approaches to examining construct comparability in multilanguage assessments lead to similar conclusions? *Applied Measurement in Education*, *24*, 349-366. Retrieved from: https://www.tandfonline.com/doi/abs/10.1080/08957347.2011.607063

Oliveri, M.E., Ercikan, K., & Simon, M. (2015). A framework for developing comparable multilingual assessments for minority populations: Why context matters. *International Journal of Testing, 15(2),* 1-20.

Oliveri, M.E., Ercikan, K., Zumbo, B.D., & Lawless, R. (2014). Uncovering substantive patterns in student responses in international large-scale assessments: Comparing a latent class to a manifest DIF approach. *International Journal of Testing*, *14*(3), 265-287. Retrieved from: https://www.researchgate.net/publication/264461819_Uncovering_Substantive_Patterns_in_Student_Responses_in_International_Large-Scale_Assessments-Comparing_a_Latent_Class_to_a_Manifest_DIF_Approach/download

Oliveri, M.E., Olson, B.F., Ercikan, K., & Zumbo, B.D. (2012). Methodologies for investigating item- and test-level measurement equivalence in international large-scale assessments. *International Journal of Testing*, *12*(3), 203-223.

Organization for Economic Cooperation and Development (OECD) (2010a). *PISA 2009 results: What students know and can do – Student performance in reading, mathematics, and science (Vol. I).* Paris: PISA, OECD Publishing. Retrieved from: https://www.oecd-ilibrary.org/education/pisa-2009-results-what-students-know-and-can-do_9789264091450-en

Organization for Economic Cooperation and Development (OECD)(2010b). *PISA 2009 results: Overcoming social background: Equity in learning opportunities and outcomes (Vol. II).* Paris: PISA, OECD Publishing. Retrieved from: https://www.oecd-ilibrary.org/education/pisa-2009-results-overcoming-social-background_9789264091504-en

Organization for Economic Cooperation and Development (OECD) (2011). Singapore: Rapid improvement followed by strong performance. In *Lessons from PISA for the United States*. Paris: PISA, OECD Publishing. Retrieved from: https://www.oecd-ilibrary.org/education/lessons-from-pisa-for-the-united-states/singapore-rapid-improvement-followed-by-strong-performance_9789264096660-8-en

Organization for Economic Cooperation and Development (OECD) (2012a). *Learning beyond fifteen: Ten years after PISA.* Paris: OECD Publishing.

Organization for Economic Cooperation and Development (OECD) (2012b). *Lesson from PISA for Japan, strong performers, and successful reformers in education.* Paris: PISA, OECD Publishing.

Organization for Economic Cooperation and Development (OECD) (2014a). *PISA – Try the test: Sample questions.* Paris: PISA, OECD Publishing. Retrieved from: https://www.oecd.org/pisa/test-2012/

Organization for Economic Cooperation and Development (OECD) (2014b). *PISA 2012 technical report.* Paris: OECD Publishing.

Organization for Economic Cooperation and Development (OECD) (2016a). *PISA 2015 Results (Volume 1): Excellence and equity in education*. Paris: PISA, OECD Publishing. Retrieved from: https://www.oecd-ilibrary.org/education/pisa-2015-results-volume-i_9789264266490-en

Organisation for Economic Cooperation and Development (OECD) (2016b). *PISA 2015 Results (Volume II): Policies and Practices for Successful Schools (p. 252)*, OECD Publishing, Paris, http://dx.doi.org/10.1787/9789264267510-en

Organization for Economic Cooperation and Development (OECD) (2017a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving. Revised edition.* Paris: PISA, OECD Publishing. Retrieved from: https://www.oecd-ilibrary.org/education/pisa-2015-assessment-and-analytical-framework_9789264281820-en

Organization for Economic Cooperation and Development (OECD) (2017b). *PISA 2015 technical report.* Paris: OECD Publishing.

Organization for Economic Cooperation and Development (OECD) (2020). Canada: Overview of the education system. OECD: Education GPS. Retrieved from: http://gpseducation.oecd.org

Paris, S.F. (2005). Reinterpreting the development of reading skills. *Reading Research Quarterly*, *40*(2), 184-202.

Pellegrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment.* National Research Council's Committee on the Foundations of Assessment. Washington, DC: National Academy Press.

Poon, C.L., Lam, K.W.L., Chan, M., Chang, M., Kwek, D., & Tan, S. (2016). Preparing students for the twenty-first century: A snapshot of Singapore's approach. In S. Choo et al. (Eds.), *Education for the 21st century perspectives: Policies and practices from around the world* (pp. 225-241). Singapore: Springer.

Prais, S.J. (2004). Cautions on OECD's recent educational survey (PISA): Rejoinder to OECD's response. *Oxford Review of Education*, *30*(4), 33-46.

Qualifications and Curriculum Authority. (QCA) (2010). *National Curriculum Assessments Code of Practice.* UK: Office of Qualifications and Examinations Regulation. Retrieved from: https://dera.ioe.ac.uk/461/1/2009-02-nc-assessments-code-of-practice.pdf

Rabinowitz, S., Ananda, S., & Bell, A. (2005). Strategies to assess the core academic knowledge of English language learners. *Journal of Applied Testing Technology*, *7*(1), 1-12.

Rheinheimer, D.C., & Penfield, D.A. (2011). The effects of Type I error rate and power of the ANCOVA F test and selected alternatives under non-normality and variance heterogeneity. *Journal of Experimental Education, 69,* 373 – 391. Retrieved from: https://doi.org/10.1080/00220970109599493

Rios, J.A., & Sireci, S.G. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing*, *14*(4), 289-312.

Rosser, A. (2018). *Beyond access: Making Indonesia's education system work.* Victoria, Australia: Lowly Institute.

Roth, W.M., Ercikan, K., Simon, M., & Fola, R. (2015). The assessment of mathematical literacy of linguistic minority students: Results of a multi-method investigation. *Journal of Mathematical Behavior, 40*, 88-105.

Rubin, D.B. (1987). *Multiple imputations for non-response in surveys.* New York: Wiley.

Schleppegrell, M.J. (2004). *The language of schooling: A functional linguistics perspective.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Schleppegrell, M.J. (2007). The linguistic challenges of mathematics teaching and learning: A research review. *Reading & Writing Quarterly*, *23*(2), 139-159.

Schuelka, M.J. (2013). Excluding students with disabilities from the culture of achievement: The case of the TIMSS, PIRLS, and PISA. *Journal of Education Policy*, *28*(2), 216-230.

Schwippert, K. (2007). *Progress in reading literacy: The impact of PIRLS 2001 in 13 countries.* New York: Waxmann.

Schwippert, K., & Lenkeit, J. (2012). *Progress in reading literacy in national and international context: The impact of PIRLS 2006 in 12 countries*. New York: Waxmann Verlag.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. New York, NY: Houghton Mifflin.

Shaftel, J., Belton-Kocher, E., Glasnapp, D., & Poggio, J. (2006). The impact of language characteristics in mathematics test items on the performance of English language learners and students with disabilities. *Educational Assessment*, *11*(2), 105-106.

Shaw, S., Crisp, V., & Johnson, N. (2012). A framework for evidencing assessment validity in large-scale, high-stakes international examinations. *Assessment in Education: Principles, Policy & Practice*, *19*(2), 159-176.

Shieh, G. (2020). Power analysis and sample size planning in ANCOVA designs. *Psychometrika*, *85*, 101-120. Retrieved from: https://doi.org/10.1007/s11336-019-09692-3

Shohamy, E. (2013). The discourse of language testing as a tool for shaping national, global, and transnational identities. *Language and Intercultural Communication*, *13*(2), 225-236.

Sivesind, K., Afsar, A., & Bachmann, K. (2016). Transnational policy transfer over three curriculum reforms in Finland: The constructions of conditional and purposive programs (1994-2016). *European Educational Research Journal, 15*, 345-365.

Sjoberg, S. (2015). PISA and global educational governance: A critique of the project, its uses and implications. *Eurasia Journal of Mathematics, Science and Technology Education, 11*(1), 111-127.

Smith, P., Cheema, J., Kumi-Yeboah, A., Warrican, S.J., & Alleyne, M. (2018). Language-based literacy differences in the literacy performance of bidialectal youth. *Teachers College Record*, *120*(1), 010304. Retrieved from: https://www.researchgate.net/publication/317007732_Language-based_Differences_in_the_Literacy_Performance_of_Bidialectal_Youth_Teachers_College_Record

Smith, P., Frazier, P., Lee, J., & Chang, R. (2018). Incongruence between native and test administration languages: Towards equal opportunity in international literacy assessment. *International Journal of Testing, 18(3),* 276-296. Retrieved from: https://doi.org/10.1080/15305058.2017.1407767

Soh, K. (2011). Statistically speaking, correctly. *North Star*, *2*(2), 108-127.

Soh, K. (2012). Fifteen-years-old students of seven east Asian cities in PISA 2009: A secondary analysis. *New Horizons in Education*, *60*(1), 83-91.

Soh, K. (2013). Finland and Singapore in PISA 2009: Similarities and differences in achievement and school management. *Compare: A Journal of Comparative and International Education*, *44*, 455-471.

Soh, K. (2014). Test language effect in international achievement comparisons: An example from PISA 2009. *Cogent Education*, *1*, 955247. Retrieved from: https://www.cogentoa.com/article/10.1080/2331186X.2014.955247

Solano-Flores, G. (2006). Language, dialect, and register: Sociolinguistics and the estimation of measurement error in the testing of English language learners. *The Teachers College Record*, *108*(1), 2354-2379.

Solano-Flores, G. (2008). Who is given tests in what language and by whom, when, and where? The need for probabilistic views of language in the testing of English language learners. *Educational Researcher, 37*(4), 189-199.

Solano-Flores, G., Backhoff, E., Contreras-Nino, L.A., & Vazquez-Munoz, M. (2015). Language shift and the inclusion of indigenous populations in large-scale assessment programs. *International Journal of Testing*, *15*(2), 136-152.

Solano-Flores, G., & Trumbull, E. (2003). Examining language in context: The need for new research and practice paradigms in the testing of English language learners. *Educational Researcher*, *32*(2), 3-13.

Solheim, O.K., & Lundetrae, K. (2018). Can test construction account for varying gender differences in international reading achievement tests of children, adolescents and young adults? A study based on Nordic results in PIRLS, PISA and PIAAC. *Assessment in Education: Principles, Policy and Practice*, *25*(1), 107-126.

Statistics Canada (2011). *Linguistic characteristics of Canadians. Census 2011.* Statistics Canada. Retrieved from: http://www12.statcan.ca/census-recensement/2011/as-sa/98-314-x/98-314-x2011001- eng.cfm

Statistics Canada (2018). *Canada at a glance. 2018*. Canada: Statistics Canada. Retrieved from: https://www150.statcan.gc.ca/n1/en/pub/12-581-x/12-581-x2018000-eng.pdf?st=YnYztyq9

Suarez-Orozco, C., Suarez-Orozco, M.M., & Todorova, I. (2008). *Learning in a new land.* Cambridge, MA: Harvard University Press.

Sukyadi, D., & Mardiani, R. (2011). The Washback Effect of the English National Examination (ENE) on English Teachers' Classroom Teaching and Students' Learning. *Kata*, *13*(1), 96-111. Retrieved from: https://doi.org/10.9744/kata.13.1.96-111

Suprapto, N. (2016). What should educational reform in Indonesia look like? Learning from the PISA science scores of East-Asian countries and Singapore. *Asia-Pacific Forum on Science Learning and Teaching*, *17*(2), Article 8, 1-21.

Tan, C., Koh, K., & Choy, W. (2016). The education system in Singapore. In S. Juszczyk (Ed.), *Asian education systems* (pp. 129-148). Torun: Adam Marszalek Publishing House. Retrieved from: https://www.academia.edu/30689089/The_education_system_in_Singapore

Tan, T.H. (2015). *Is it time for a new approach to education in Singapore? Towards education for a flourishing life. THF Lecture Series.* Singapore: National University of Singapore.

Tilak, J.B. (2003). *Vocational education and training in Asia: International handbook of educational research in the Asia-Pacific Region* (pp. 673-686). Dordrecht:Springer.

UNESCO (2006). *EFA global monitoring report 2006: Education for all.* Paris: UNESCO.

UNESCO (2015). *A decade of promoting multilingualism in cyberspace*. Paris: UNESCO.

United Nations (UN) (2016). *Youth and migration.* UN: Department of Economic and Social Affairs. Retrieved from: https://www.un.org/esa/socdev/documents/youth/fact-sheets/youth-migration.pdf

United Nations (UN) (2017). *Sustainable development goals.* Retrieved from: https://www.un.org/sustainabledevelopment/sustainable-development-goals/

United Nations Development Program (UNDP) (2014). *Human development report 2014: Indonesia.* Retrieved from: http://hdr.undp.org/sites/all/themes/hdr_theme/country-notes/IDN.pdf.

US Government of Education (2017). *Transitioning to the Every Student Succeeds Act (ESSA).* USA: Department of Education. Retrieved from: https://www2.ed.gov/policy/elsec/leg/essa/essatransitionfaqs11817.pdf

Von Davier, M., Gonzalez, E., & Mislevy, R.J. (2009). What are plausible values and why are they useful? *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments*,2, 9-36. Retrieved from: http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI _Monograph_Volume_02_Chapter_01.pdf

Wee, L. (2003). Linguistic instrumentalism in Singapore. *Journal of Multilingual class and multicultural*, *24*(3), 211-224.

Wei, R., & Su, J. (2012). The statistics of English in China. *English Today*, *28*(3), 10-14.

Wilson, M.D., & Lueck, K. (2014). Working with missing data: Imputation of non-response items in categorical survey data with a non-monotone missing pattern. *Journal of Applied Mathematics, 368791*, 1-9.

Wyse, D., & Torgerson, C. (2017). Experimental trials and 'what works?' in education: The case of grammar for writing. *British Educational Research Journal*, *43*(6), 1019-1047.

Xie, Y., & Wilson, M. (2008). Investigating DIF and extensions using an LLTM approach and also an individual differences approach: An international testing context. *Psychology Science Quarterly*, *50*(3), 403-416.

Yusny, R. (2013). ELT in Indonesian context: Issues and challenges. *Englisia Journal*, *Nov 1*(1), 81-99. https://jurnal.ar-raniry.ac.id/index.php/englisia/article/view/140

Zhao, Y. (2020). Two decades of havoc: A synthesis of criticism against PISA. *Journal of Educational Change*, *21*, 245-266.

Zhao, Y., & Gearin, B. (2016). Squeezed out. In D. Ambrose & R.J. Sternberg (Eds.), *Creative intelligence in the 21st century* (pp. 121-138). Berlin: Springer.

Zilliacus, H., Holm, G., & Sahlstrom, F. (2017). Taking steps towards institutionalizing multicultural education – The national curriculum of Finland. *Multicultural Education Review, 9*(4), 231-248.

APPENDICES

Appendix A

Student Questionnaire (Background)

STUDENT QUESTIONNAIRE

Computer-based version/main survey version

Students complete the student questionnaire after the literacy assessment. The questionnaire takes about 35 minutes to complete.

The core questions on home background are similar to those used in previous PISA assessments. The questionnaire covers:

- Student, student's family and student's home
- Student's view about his/her life
- Student's school
- Student's school schedule and learning time
- Science learning in school
- Student's views about science

Technical terms are given in <brackets> and are adapted to the national context by the national data collection centre of the participating country or economy. In this annex, an explanation of the technical terms is given below the questionnaire item.

Student, student's family and student's home

ST001
ST001Q01TA

What <grade> are you in?
(Please select from the drop-down menu to answer the question.)

Please choose ▼

Option A
Option B
Option C
Option …

Drop down menu, including all possible grades attended by 15-year-olds, according to your study programme table as agreed on in the Demographic Tasks.

ST002
ST002Q01TA

Which one of the following <programmes> are you in?
(Please select one response.)

<Programme 1>	☐ 1
<Programme 2>	☐ 2
<Programme 3>	☐ 3
<Programme 4>	☐ 4
<Programme 5>	☐ 5
<Programme 6>	☐ 6

ST003	**On what date were you born?** *(Please select the day, month and year from the drop-down menus to answer the question.)*				
ST003Q01TA	Day	*Please choose* ▼ *Option A* *Option B* *Option C* *Option …*	1 2 3 4 5 6 7 8 10 …		
ST003Q02TA	Month	*Please choose* ▼ *Option A* *Option B* *Option C* *Option …*		January February March April May June July August September October November December	
ST003Q03TA	Year	*Please choose* ▼ *Option A* *Option B* *Option C* *Option …*			1998 1999 2000 2001

Consistency check/soft reminder if day, month or year is missing: "Please enter your complete birth date".

ST004 ST004Q01TA	**Are you female or male?** *(Please select one response.)*		
		Female	**Male**
		☐1	☐2

ST005 ST005Q01TA	**What is the <highest level of schooling> completed by your mother?** *If you are not sure which box to choose, please ask the <test administrator> for help.* *(Please select one response.)*	
	<ISCED level 3A>	☐1
	<ISCED level 3B, 3C>	☐2
	<ISCED level 2>	☐3
	<ISCED level 1>	☐4
	She did not complete <ISCED level 1>	☐5

ST006	**Does your mother have any of the following qualifications?** *If you are not sure how to answer this question, please ask the <test administrator> for help. (Please select one response in each row.)*		
		Yes	No
ST006Q01TA	<ISCED level 6>	☐1	☐2
ST006Q02TA	<ISCED level 5A>	☐1	☐2
ST006Q03TA	<ISCED level 5B>	☐1	☐2
ST006Q04TA	<ISCED level 4>	☐1	☐2

ST007 ST007Q01TA	**What is the <highest level of schooling> completed by your father?** *If you are not sure which box to choose, please ask the <test administrator> for help. (Please select one response.)*	
	<ISCED level 3A>	☐1
	<ISCED level 3B, 3C>	☐2
	<ISCED level 2>	☐3
	<ISCED level 1>	☐4
	He did not complete <ISCED level 1>	☐5

ST008	**Does your father have any of the following qualifications?** *If you are not sure how to answer this question, please ask the <test administrator> for help. (Please select one response in each row.)*		
		Yes	No
ST008Q01TA	<ISCED level 6>	☐1	☐2
ST008Q02TA	<ISCED level 5A>	☐1	☐2
ST008Q03TA	<ISCED level 5B>	☐1	☐2
ST008Q04TA	<ISCED level 4>	☐1	☐2

ST011	**Which of the following are in your home?** *(Please select one response in each row.)*		
		Yes	No
ST011Q01TA	A desk to study at	☐1	☐2
ST011Q02TA	A room of your own	☐1	☐2
ST011Q03TA	A quiet place to study	☐1	☐2
ST011Q04TA	A computer you can use for school work	☐1	☐2
ST011Q05TA	Educational software	☐1	☐2
ST011Q06TA	A link to the Internet	☐1	☐2
ST011Q07TA	Classic literature (e.g. <Shakespeare>)	☐1	☐2
ST011Q08TA	Books of poetry	☐1	☐2
ST011Q09TA	Works of art (e.g. paintings)	☐1	☐2
ST011Q10TA	Books to help with your school work	☐1	☐2
ST011Q11TA	<Technical reference books>	☐1	☐2
ST011Q12TA	A dictionary	☐1	☐2
ST011Q16NA	Books on art, music or design	☐1	☐2
ST011Q17TA	<Country-specific wealth item 1>	☐1	☐2
ST011Q18TA	<Country-specific wealth item 2>	☐1	☐2
ST011Q19TA	<Country-specific wealth item 3>	☐1	☐2

ST012	**How many of these are there at your home?** *(Please select one response in each row.)*	None	One	Two	Three or more
ST012Q01TA	Televisions	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q02TA	Cars	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q03TA	Rooms with a bath or shower	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q05NA	<Cell phones> with Internet access (e.g. smartphones)	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q06NA	Computers (desktop computer, portable laptop or notebook)	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q07NA	<Tablet computers> (e.g. <iPad®>, <BlackBerry® PlayBook™>)	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q08NA	E-book readers (e.g. <Kindle™>, <Kobo>, <Bookeen>)	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST012Q09NA	Musical instruments (e.g. guitar, piano)	$\square_1$	$\square_2$	$\square_3$	$\square_4$

ST013 ST013Q01TA

How many books are there in your home?

There are usually about 40 books per metre of shelving. Do not include magazines, newspapers or your schoolbooks.

(Please select one response.)

0-10 books	$\square_1$
11-25 books	$\square_2$
26-100 books	$\square_3$
101-200 books	$\square_4$
201-500 books	$\square_5$
More than 500 books	$\square_6$

ST014

The following two questions concern your mother's job:

(If she is not working now, please tell us her last main job.)

ST014Q01TA What is your mother's main job?
(e.g. school teacher, kitchen-hand, sales manager)

Please type in the job title. ______________________________

ST014Q02TA What does your mother do in her main job?
(e.g. teaches high school students, helps the cook prepare meals in a restaurant, manages a sales team)

Please use a sentence to describe the kind of work she does or did in that job. ______________________________

ST015

The following two questions concern your father's job:

(If he is not working now, please tell us his last main job.)

ST015Q01TA What is your father's main job?
(e.g. school teacher, kitchen-hand, sales manager)

Please type in the job title. ______________________________

ST015Q02TA What does your father do in his main job?
(e.g. teaches high school students, helps the cook prepare meals in a restaurant, manages a sales team)

Please use a sentence to describe the kind of work he does or did in that job. ______________________________

ST123	**Thinking about the <this academic year>: to what extent do you agree or disagree with the following statements?** *(Please select one response in each row.)*				
		Strongly disagree	Disagree	Agree	Strongly agree
ST123Q01NA	My parents are interested in my school activities.	☐1	☐2	☐3	☐4
ST123Q02NA	My parents support my educational efforts and achievements.	☐1	☐2	☐3	☐4
ST123Q03NA	My parents support me when I am facing difficulties at school.	☐1	☐2	☐3	☐4
ST123Q04NA	My parents encourage me to be confident.	☐1	☐2	☐3	☐4

ST019	**In what country were you and your parents born?** *(Please select one response in each column.)*			
		You ST019Q01TA	Mother ST019Q01TB	Father ST019Q01TC
	<Country A>	☐1	☐1	☐1
	<Country B>	☐2	☐2	☐2
	<Country C>	☐3	☐3	☐3
	<Country D>	☐4	☐4	☐4
	<…etc.>	☐5	☐5	☐5
	Other country	☐6	☐6	☐6

This is a filter question. If the student was born in <country of test> skip ST021. If he or she was NOT born in <country of test> go to ST021. ELSE go to ST022.

ST021
ST021Q01TA

→ Only if answer in ST019 "you" was NOT "<country of test>". ELSE skip and proceed to ST022.

How old were you when you arrived in <country of test>?

(Please select from the drop-down menu to answer the question.
If you were less than 12 months old, please select "age 0-1" [age zero to one])

Please choose ▼

Option A
Option B
Option C
Option …

age 0-1	1
age 1	2
age 2	3
age 3	4
age 4	5
age 5	6
age 6	7
age 7	8
age 8	9
age 9	10
age 10	11
age 11	12
age 12	13
age 13	14
age 14	15
age 15	16
age 16	17

ST022	**What language do you speak at home most of the time?** *(Please select one response.)*	
ST022Q01TA	<Language 1>	☐1
ST022Q02TA	<Language 2>	☐2
ST022Q03TA	<Language 3>	☐3
ST022Q04TA	< …etc. >	☐4
ST022Q05TA	Other language	☐5

ST125 ST125Q01NA	**How old were you when you started <ISCED 0>?** *(Please choose from the drop-down menu to answer the question.)*	
	Years:	*Please choose* ▼ *Option A* *Option B* *Option C* *Option …*

Drop-down menu, offering answers "1 year or younger", 2 years, 3 years, 4 years, 5 years, "6 years or older", "I did not attend <ISCED 0>", "I do not remember".

ST126 ST126Q01TA	**How old were you when you started <ISCED 1>?** *(Please choose from the drop-down menu to answer the question.)*	
	Years:	*Please choose* ▼ *Option A* *Option B* *Option C* *Option …*

Drop-down menu, offering answers "3 years or younger", 4 years, 5 years, 6 years, 7 years, 8 years, "9 years or older".

ST127 **Have you ever repeated a <grade>?** *(Please select one response in each row.)*

		No, never	Yes, once	Yes, twice or more
ST127Q01TA	At <ISCED 1>	☐1	☐2	☐3
ST127Q02TA	At <ISCED 2>	☐1	☐2	☐3
ST127Q03TA	At <ISCED 3>	☐1	☐2	☐3

Student's view about his/her life

ST016 ST016Q01NA	*The following question asks how satisfied you feel about your life, on a scale from "0" to "10". Zero means you feel "not at all satisfied" and "10" means "completely satisfied".* **Overall, how satisfied are you with your life as a whole these days?** *(Please move the slider to the appropriate number.)*

Slider bar: parking position, range 0-10 (not at all satisfied, completely satisfied), step = 1.

ST111 ST111Q01TA	**Which of the following do you expect to complete?** *(Please select one response.)*	
	<ISCED level 2>	☐1
	<ISCED level 3B or C>	☐2
	<ISCED level 3A>	☐3
	<ISCED level 4>	☐4
	<ISCED level 5B>	☐5
	<ISCED level 5A or 6>	☐6

ST114 ST114Q01TA	**What kind of job do you expect to have when you are about 30 years old?**
	Please type in the job title. ______________________________

ST118 **To what extent do you disagree or agree with the following statements about yourself?**
(Please select one response in each row.)

		Strongly disagree	Disagree	Agree	Strongly agree
ST118Q01NA	I often worry that it will be difficult for me taking a test.	☐1	☐2	☐3	☐4
ST118Q02NA	I worry that I will get poor <grades> at school.	☐1	☐2	☐3	☐4
ST118Q03NA	Even if I am well prepared for a test I feel very anxious.	☐1	☐2	☐3	☐4
ST118Q04NA	I get very tense when I study for a test.	☐1	☐2	☐3	☐4
ST118Q05NA	I get nervous when I don't know how to solve a task at school.	☐1	☐2	☐3	☐4

ST119 **To what extent do you disagree or agree with the following statements about yourself?**
(Please select one response in each row.)

		Strongly disagree	Disagree	Agree	Strongly agree
ST119Q01NA	I want top grades in most or all of my courses.	☐1	☐2	☐3	☐4
ST119Q02NA	I want to be able to select from among the best opportunities available when I graduate.	☐1	☐2	☐3	☐4
ST119Q03NA	I want to be the best, whatever I do.	☐1	☐2	☐3	☐4
ST119Q04NA	I see myself as an ambitious person.	☐1	☐2	☐3	☐4
ST119Q05NA	I want to be one of the best students in my class.	☐1	☐2	☐3	☐4

ST121	**Please read the descriptions about the following three students. Based on the information provided here, how much would you disagree or agree with the statement that this student is <u>motivated</u>?** *(Please select one response in each row.)*				
		Strongly disagree	Disagree	Agree	Strongly agree
ST121Q01NA	<NAME 1> gives up easily when confronted with a problem and is often not prepared for his classes. **<Name 1> is motivated.**	☐1	☐2	☐3	☐4
ST121Q02NA	<NAME 2> mostly remains interested in the tasks she starts and sometimes does more than what is expected from her. **<Name 2> is motivated.**	☐1	☐2	☐3	☐4
ST121Q03NA	<NAME 3> wants to get top grades at school and continues working on tasks until everything is perfect. **<Name 3> is motivated.**	☐1	☐2	☐3	☐4

Student's school

ST082	**To what extent do you disagree or agree with the following statements about yourself?** *(Please select one response in each row.)*				
		Strongly disagree	Disagree	Agree	Strongly agree
ST082Q01NA	I prefer working as part of a team to working alone.	☐1	☐2	☐3	☐4
ST082Q02NA	I am a good listener.	☐1	☐2	☐3	☐4
ST082Q03NA	I enjoy seeing my classmates be successful.	☐1	☐2	☐3	☐4
ST082Q08NA	I take into account what others are interested in.	☐1	☐2	☐3	☐4
ST082Q09NA	I find that teams make better decisions than individuals.	☐1	☐2	☐3	☐4
ST082Q12NA	I enjoy considering different perspectives.	☐1	☐2	☐3	☐4
ST082Q13NA	I find that teamwork raises my own efficiency.	☐1	☐2	☐3	☐4
ST082Q14NA	I enjoy cooperating with peers.	☐1	☐2	☐3	☐4

ST034	**Thinking about your school: to what extent do you agree with the following statements?** *(Please select one response in each row.)*				
		Strongly agree	Agree	Disagree	Strongly disagree
ST034Q01TA	I feel like an outsider (or left out of things) at school.	☐1	☐2	☐3	☐4
ST034Q02TA	I make friends easily at school.	☐1	☐2	☐3	☐4
ST034Q03TA	I feel like I belong at school.	☐1	☐2	☐3	☐4
ST034Q04TA	I feel awkward and out of place in my school.	☐1	☐2	☐3	☐4
ST034Q05TA	Other students seem to like me.	☐1	☐2	☐3	☐4
ST034Q06TA	I feel lonely at school.	☐1	☐2	☐3	☐4

ST039	During the past 12 months, how often did you have the following experiences at school? *(Please select one response in each row.)*				
		Never or almost never	A few times a year	A few times a month	Once a week or more
ST039Q01NA	Teachers called on me less often than they called on other students.	☐1	☐2	☐3	☐4
ST039Q02NA	Teachers graded me harder than they graded other students.	☐1	☐2	☐3	☐4
ST039Q03NA	Teachers gave me the impression that they think I am less smart than I really am.	☐1	☐2	☐3	☐4
ST039Q04NA	Teachers disciplined me more harshly than other students.	☐1	☐2	☐3	☐4
ST039Q05NA	Teachers ridiculed me in front of others.	☐1	☐2	☐3	☐4
ST039Q06NA	Teachers said something insulting to me in front of others.	☐1	☐2	☐3	☐4

ST038	During the past 12 months, how often have you had the following experiences in school? *(Please select one response in each row.)*				
		Never or almost never	A few times a year	A few times a month	Once a week or more
ST038Q01NA	I got called names by other students.	☐1	☐2	☐3	☐4
ST038Q02NA	I got picked on by other students.	☐1	☐2	☐3	☐4
ST038Q03NA	Other students left me out of things on purpose.	☐1	☐2	☐3	☐4
ST038Q04NA	Other students made fun of me.	☐1	☐2	☐3	☐4
ST038Q05NA	I was threatened by other students.	☐1	☐2	☐3	☐4
ST038Q06NA	Other students took away or destroyed things that belonged to me.	☐1	☐2	☐3	☐4
ST038Q07NA	I got hit or pushed around by other students.	☐1	☐2	☐3	☐4
ST038Q08NA	Other students spread nasty rumours about me.	☐1	☐2	☐3	☐4

Student's school schedule and learning time

ST059	How many <class periods> per week are you typically <u>required to attend</u> for the following subjects? *(Please enter a number in each row. Enter "0" [zero] if you have none.)*	
ST059Q01TA	Number of <class periods> per week in <test language>	
ST059Q02TA	Number of <class periods> per week in mathematics	
ST059Q03TA	Number of <class periods> per week in <science>	

Open text entry full numbers only. Consistency check, if entries are greater than 15.

ST060 ST060Q01NA	In a normal, full week at school, how many <class periods> are you required to attend in total? *(Please move the slider to the number of <class periods> per week.)*	
	Number of ALL <class periods>	

Slider bar: parking position, range 0–"80 or more", step 1; consistency check/soft reminder for values smaller than 10 and greater than 60.

ST061 ST061Q01NA	How many minutes, on average, are there in a <class period>? *(Please move the slider to the number of minutes per <class period>.)*	
	Average minutes in a <class period>	

Slider range 0–"120 or more"; consistency check/soft reminder for values smaller than 10 and greater than 80; step 5.

ST062	**In the last two full weeks of school, how often did the following things occur?** *(Please select one response in each row.)*				
		Never	One or two times	Three or four times	Five or more times
ST062Q01TA	I <skipped> a whole school day	☐1	☐2	☐3	☐4
ST062Q02TA	I <skipped> some classes	☐1	☐2	☐3	☐4
ST062Q03TA	I arrived late for school	☐1	☐2	☐3	☐4

ST071	**This school year, approximately how many hours per week do you spend learning in <u>addition to your required school schedule</u> in the following subjects?** *(Please include the total hours for homework, additional instruction and private study.)* *(Please move the slider to the number of total hours. Select "0" [zero] if you do not do homework, study or practice for a subject.)*	
ST071Q01NA	<School science>	slider
ST071Q02NA	Mathematics	slider
ST071Q03NA	<Test language>	slider
ST071Q04NA	<Foreign language>	slider
ST071Q05NA	Other	slider

Slider bar: parking position, range 0–"30 hours per week or more", step = 1; consistency check/soft reminder for values > 20.

ST031
ST031Q01NA

This school year, on average, on how many days do you attend physical education classes each week?
(Please select from the drop-down menu to answer the question.)

Please choose ▼
Option A
Option B
Option C
Option …

Drop down menu, 0"<number of instructional days per calendar week> days".

ST032	**Outside of school, during the past 7 days, on how many days did you engage in the following?** *(Please select one response from the drop-down menus to answer the questions.)*	
ST032Q01NA	<u>Moderate physical activities</u> for a <u>total of at least 60 minutes per day</u> (e.g. walking, climbing stairs, riding a bike to school, <country-specific>)	*Please choose* ▼ *Option A* *Option B* *Option C* *Option …*
ST032Q02NA	<u>Vigorous physical activities</u> for <u>at least 20 minutes per day</u> that made you sweat and breathe hard (e.g. running, cycling, aerobics, soccer, skating, <country-specific>)	*Please choose* ▼ *Option A* *Option B* *Option C* *Option …*

Drop down menus 0-7 days.

Science learning in school

ST063 **Which of the following <school science> courses did you attend this school year or last school year?**
(Please select all that apply in each row.)

		This year	Last year
ST063Q01N	Physics	☐1	☐1
ST063Q02N	Chemistry	☐1	☐1
ST063Q03N	Biology	☐1	☐1
ST063Q04N	<Earth and space>	☐1	☐1
ST063Q05N	Applied sciences and technology (e. g. <country-specific example>)	☐1	☐1
ST063Q06N	<General, integrated, or comprehensive science> course (e. g. <country-specific example>)	☐1	☐1

This is a filter question. ST064-ST107 only apply if sum of clicks in category "this year" is greater than 0 (at least one science course this year). Else skip ST064-ST107 and proceed to section on "Student's view on science".

ST064 → Only applies if the student answered to attend at least one <school science> course in this school year in ST063.
To what extent can you choose the following for your <school science> courses?
(Please select one response in each row.)

		No, not at all	Yes, to a certain degree	Yes, I can choose freely
ST064Q01NA	I can choose the <school science> course(s) I study.	☐1	☐2	☐3
ST064Q02NA	I can choose the level of difficulty.	☐1	☐2	☐3
ST064Q03NA	I can choose the number of <school science> courses or <class periods>.	☐1	☐2	☐3

ST097 → Only applies if the student answered to attend at least one <school science> course in this school year in ST063.
How often do these things happen in your <school science> lessons?
(Please select one response in each row.)

		Every lesson	Most lessons	Some lessons	Never or hardly ever
ST097Q01TA	Students don't listen to what the teacher says.	☐1	☐2	☐3	☐4
ST097Q02TA	There is noise and disorder.	☐1	☐2	☐3	☐4
ST097Q03TA	The teacher has to wait a long time for students to quiet down.	☐1	☐2	☐3	☐4
ST097Q04TA	Students cannot work well.	☐1	☐2	☐3	☐4
ST097Q05TA	Students don't start working for a long time after the lesson begins.	☐1	☐2	☐3	☐4

ST098

→ Only applies if the student answered to attend at least one <school science> course in this school year in ST063.

When learning <school science> topics at school, how often do the following activities occur?

(Please select one response in each row.)

		In all lessons	In most lessons	In some lessons	Never or hardly ever
ST098Q01TA	Students are given opportunities to explain their ideas.	☐1	☐2	☐3	☐4
ST098Q02TA	Students spend time in the laboratory doing practical experiments.	☐1	☐2	☐3	☐4
ST098Q03NA	Students are required to argue about science questions.	☐1	☐2	☐3	☐4
ST098Q05TA	Students are asked to draw conclusions from an experiment they have conducted.	☐1	☐2	☐3	☐4
ST098Q06TA	The teacher explains how a <school science> idea can be applied to a number of different phenomena (e.g. the movement of objects, substances with similar properties).	☐1	☐2	☐3	☐4
ST098Q07TA	Students are allowed to design their own experiments.	☐1	☐2	☐3	☐4
ST098Q08NA	There is a class debate about investigations.	☐1	☐2	☐3	☐4
ST098Q09TA	The teacher clearly explains the relevance of <broad science> concepts to our lives.	☐1	☐2	☐3	☐4
ST098Q10NA	Students are asked to do an investigation to test ideas.	☐1	☐2	☐3	☐4

When answering the following questions, please keep one of your current <school science> courses in mind all the time. You are free to choose which course this should be.

ST065
ST065Q01NA

→ Only applies if the student answered to attend at least one <school science> course in this school year in ST063.

What is the name of this <school science> course?

(Please type the name of the course.)

ST100

→ Only applies if the student answered to attend at least one <school science> course in this school year in ST063.

How often do these things happen in your <school science> lessons?

(Please select one response in each row.)

		Every lesson	Most lessons	Some lessons	Never or hardly ever
ST100Q01TA	The teacher shows an interest in every student's learning.	☐1	☐2	☐3	☐4
ST100Q02TA	The teacher gives extra help when students need it.	☐1	☐2	☐3	☐4
ST100Q03TA	The teacher helps students with their learning.	☐1	☐2	☐3	☐4
ST100Q04TA	The teacher continues teaching until the students understand.	☐1	☐2	☐3	☐4
ST100Q05TA	The teacher gives students an opportunity to express opinions.	☐1	☐2	☐3	☐4

ST103	→ Only applies if the student answered to attend at least one <school science> course in this school year in ST063. **How often do these things happen in your lessons for this <school science> course?** *(Remember to answer this question in reference to the <school science> course you indicated earlier.) (Please select one response in each row.)*				
		Never or almost never	Some lessons	Many lessons	Every lesson or almost every lesson
ST103Q01NA	The teacher explains scientific ideas.	☐1	☐2	☐3	☐4
ST103Q03NA	A whole class discussion takes place with the teacher.	☐1	☐2	☐3	☐4
ST103Q08NA	The teacher discusses our questions.	☐1	☐2	☐3	☐4
ST103Q11NA	The teacher demonstrates an idea.	☐1	☐2	☐3	☐4

ST104	→ Only applies if the student answered to attend at least one <school science> course in this school year in ST063. **How often do these things happen in your lessons for this <school science> course?** *(Remember to answer this question in reference to the <school science> course you indicated earlier.) (Please select one response in each row.)*				
		Never or almost never	Some lessons	Many lessons	Every lesson or almost every lesson
ST104Q01NA	The teacher tells me how I am performing in this course.	☐1	☐2	☐3	☐4
ST104Q02NA	The teacher gives me feedback on my strengths in this <school science> subject.	☐1	☐2	☐3	☐4
ST104Q03NA	The teacher tells me in which areas I can still improve.	☐1	☐2	☐3	☐4
ST104Q04NA	The teacher tells me how I can improve my performance.	☐1	☐2	☐3	☐4
ST104Q05NA	The teacher advises me on how to reach my learning goals.	☐1	☐2	☐3	☐4

ST107	→ Only applies if the student answered to attend at least one <school science> course in this school year in ST063. **How often do these things happen in your lessons for this <school science> course?** *(Remember to answer this question in reference to the <school science> course you indicated earlier.) (Please select one response in each row.)*				
		Never or almost never	Some lessons	Many lessons	Every lesson or almost every lesson
ST107Q01NA	The teacher adapts the lesson to my class's needs and knowledge.	☐1	☐2	☐3	☐4
ST107Q02NA	The teacher provides individual help when a student has difficulties understanding a topic or task.	☐1	☐2	☐3	☐4
ST107Q03NA	The teacher changes the structure of the lesson on a topic that most students find difficult to understand.	☐1	☐2	☐3	☐4

Student's view on science

ST092 **How informed are you about the following environmental issues?**
(Please select one response in each row.)

		I have never heard of this	I have heard about this but I would not be able to explain what it is really about	I know something about this and could explain the general issue	I am familiar with this and I would be able to explain this well
ST092Q01TA	The increase of greenhouse gases in the atmosphere	□1	□2	□3	□4
ST092Q02TA	The use of genetically modified organisms (<GMO>)	□1	□2	□3	□4
ST092Q04TA	Nuclear waste	□1	□2	□3	□4
ST092Q05TA	The consequences of clearing forests for other land use	□1	□2	□3	□4
ST092Q06NA	Air pollution	□1	□2	□3	□4
ST092Q08NA	Extinction of plants and animals	□1	□2	□3	□4
ST092Q09NA	Water shortage	□1	□2	□3	□4

ST093 **Do you think problems associated with the environmental issues below will improve or get worse over the next 20 years?**
(Please select one response in each row.)

		Improve	Stay about the same	Get worse
ST093Q01TA	Air pollution	□1	□2	□3
ST093Q03TA	Extinction of plants and animals	□1	□2	□3
ST093Q04TA	Clearing of forests for other land use	□1	□2	□3
ST093Q05TA	Water shortages	□1	□2	□3
ST093Q06TA	Nuclear waste	□1	□2	□3
ST093Q07NA	The increase of greenhouse gases in the atmosphere	□1	□2	□3
ST093Q08NA	The use of genetically modified organisms (<GMO>)	□1	□2	□3

ST094 **How much do you disagree or agree with the statements about yourself below?**
(Please select one response in each row.)

		Strongly disagree	Disagree	Agree	Strongly agree
ST094Q01NA	I generally have fun when I am learning <broad science> topics.	□1	□2	□3	□4
ST094Q02NA	I like reading about <broad science>.	□1	□2	□3	□4
ST094Q03NA	I am happy working on <broad science> topics.	□1	□2	□3	□4
ST094Q04NA	I enjoy acquiring new knowledge in <broad science>.	□1	□2	□3	□4
ST094Q05NA	I am interested in learning about <broad science>.	□1	□2	□3	□4

ST095	To what extent are you interested in the following <broad science> topics? *(Please select one response in each row.)*					
		Not interested	Hardly interested	Interested	Highly interested	I don't know what this is
ST095Q04NA	Biosphere (e.g. ecosystem services, sustainability)	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$
ST095Q07NA	Motion and forces (e.g. velocity, friction, magnetic and gravitational forces)	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$
ST095Q08NA	Energy and its transformation (e.g. conservation, chemical reactions)	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$
ST095Q13NA	The Universe and its history	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$
ST095Q15NA	How science can help us prevent disease	$\square_1$	$\square_2$	$\square_3$	$\square_4$	$\square_5$

ST113	How much do you agree with the statements below? *(Please select one response in each row.)*				
		Strongly agree	Agree	Disagree	Strongly disagree
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.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
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.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST113Q03TA	Studying my <school science> subject(s) is worthwhile for me because what I learn will improve my career prospects.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST113Q04TA	Many things I learn in my <school science> subject(s) will help me to get a job.	$\square_1$	$\square_2$	$\square_3$	$\square_4$

ST129	How easy do you think it would be for you to perform the following tasks on your own? *(Please select one response in each row.)*				
		I could do this easily	I could do this with a bit of effort	I would struggle to do this on my own	I couldn't do this
ST129Q01TA	Recognise the science question that underlies a newspaper report on a health issue.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST129Q02TA	Explain why earthquakes occur more frequently in some areas than in others.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST129Q03TA	Describe the role of antibiotics in the treatment of disease.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST129Q04TA	Identify the science question associated with the disposal of garbage.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST129Q05TA	Predict how changes to an environment will affect the survival of certain species.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST129Q06TA	Interpret the scientific information provided on the labelling of food items.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST129Q07TA	Discuss how new evidence can lead you to change your understanding about the possibility of life on Mars.	$\square_1$	$\square_2$	$\square_3$	$\square_4$
ST129Q08TA	Identify the better of two explanations for the formation of acid rain.	$\square_1$	$\square_2$	$\square_3$	$\square_4$

ST131	**How much do you disagree or agree with the statements below?** *(Please select one response in each row.)*				
		Strongly disagree	Disagree	Agree	Strongly agree
ST131Q01NA	A good way to know if something is true is to do an experiment.	☐1	☐2	☐3	☐4
ST131Q03NA	Ideas in <broad science> sometimes change.	☐1	☐2	☐3	☐4
ST131Q04NA	Good answers are based on evidence from many different experiments.	☐1	☐2	☐3	☐4
ST131Q06NA	It is good to try experiments more than once to make sure of your findings.	☐1	☐2	☐3	☐4
ST131Q08NA	Sometimes <broad science> scientists change their minds about what is true in science.	☐1	☐2	☐3	☐4
ST131Q11NA	The ideas in <broad science> science books sometimes change.	☐1	☐2	☐3	☐4

ST146	**How often do you do these things?** *(Please select one response in each row.)*				
		Very often	Regularly	Sometimes	Never or hardly ever
ST146Q01TA	Watch TV programmes about <broad science>	☐1	☐2	☐3	☐4
ST146Q02TA	Borrow or buy books on <broad science> topics	☐1	☐2	☐3	☐4
ST146Q03TA	Visit web sites about <broad science> topics	☐1	☐2	☐3	☐4
ST146Q04TA	Read <broad science> magazines or science articles in newspapers	☐1	☐2	☐3	☐4
ST146Q05TA	Attend a <science club>	☐1	☐2	☐3	☐4
ST146Q06NA	Simulate natural phenomena in computer programs/ virtual labs	☐1	☐2	☐3	☐4
ST146Q07NA	Simulate technical processes in computer programs/ virtual labs	☐1	☐2	☐3	☐4
ST146Q08NA	Visit web sites of ecology organisations	☐1	☐2	☐3	☐4
ST146Q09NA	Follow news of science, environmental, or ecology organizations via blogs and microblogging	☐1	☐2	☐3	☐4

ST076	**On the most recent day *you attended school*, did you do any of the following before going to school?** *(Please select one response in each row.)*		
		Yes	No
ST076Q01NA	Eat breakfast	☐1	☐2
ST076Q02NA	Study for school or homework	☐1	☐2
ST076Q03NA	Watch TV/<DVD>/Video	☐1	☐2
ST076Q04NA	Read a book/newspaper/magazine	☐1	☐2
ST076Q05NA	Internet/Chat/Social networks (e.g. <Facebook>, <country-specific social network>)	☐1	☐2
ST076Q06NA	Play video games	☐1	☐2
ST076Q07NA	Meet friends or talk to friends on the phone	☐1	☐2
ST076Q08NA	Talk to your parents	☐1	☐2
ST076Q09NA	Work in the household or take care of other family members	☐1	☐2
ST076Q10NA	Work for pay	☐1	☐2
ST076Q11NA	Exercise or practice a sport	☐1	☐2

ST078	On the most recent day *you attended school*, did you do any of the following after leaving school? *(Please select one response in each row.)*		
		Yes	No
ST078Q01NA	Eat dinner	□1	□2
ST078Q02NA	Study for school or homework	□1	□2
ST078Q03NA	Watch TV/<DVD>/Video	□1	□2
ST078Q04NA	Read a book/newspaper/magazine	□1	□2
ST078Q05NA	Internet/Chat/Social networks (e.g. <Facebook>, <country-specific social network>)	□1	□2
ST078Q06NA	Play video games	□1	□2
ST078Q07NA	Meet friends or talk to friends on the phone	□1	□2
ST078Q08NA	Talk to your parents	□1	□2
ST078Q09NA	Work in the household or take care of other family members	□1	□2
ST078Q10NA	Work for pay	□1	□2
ST078Q11NA	Exercise or practice a sport	□1	□2

Appendix B

PISA 2015 Reading Proficiency Levels

Level	Lower score limit	Task characteristics
6	698	Tasks at this level typically require the reader to make multiple inferences, comparisons and contrasts that are both detailed and precise. They require demonstration of a full and detailed understanding of one or more texts and may involve integrating information from more than one text. Tasks may require the reader to deal with unfamiliar ideas, in the presence of prominent competing information, and to generate abstract categories for interpretations. *Reflect and evaluate* tasks may require the reader to hypothesise about or critically evaluate a complex text on an unfamiliar topic, taking into account multiple criteria or perspectives, and applying sophisticated understandings from beyond the text. A salient condition for access and retrieve tasks at this level is precision of analysis and fine attention to detail that is inconspicuous in the texts.
5	626	Tasks at this level that involve retrieving information require the reader to locate and organise several pieces of deeply embedded information, inferring which information in the text is relevant. Reflective tasks require critical evaluation or hypothesis, drawing on specialised knowledge. Both interpretative and reflective tasks require a full and detailed understanding of a text whose content or form is unfamiliar. For all aspects of reading, tasks at this level typically involve dealing with concepts that are contrary to expectations.
4	553	Tasks at this level that involve retrieving information require the reader to locate and organise several pieces of embedded information. Some tasks at this level require interpreting the meaning of nuances of language in a section of text by taking into account the text as a whole. Other interpretative tasks require understanding and applying categories in an unfamiliar context. Reflective tasks at this level require readers to use formal or public knowledge to hypothesise about or critically evaluate a text. Readers must demonstrate an accurate understanding of long or complex texts whose content or form may be unfamiliar.
3	480	Tasks at this level require the reader to locate, and in some cases recognise the relationship between, several pieces of information that must meet multiple conditions. Interpretative tasks at this level require the reader to integrate several parts of a text in order to identify a main idea, understand a relationship or construe the meaning of a word or phrase. They need to take into account many features in comparing, contrasting, or categorising. Often the required information is not prominent or there is much competing information; or there are other text obstacles, such as ideas that are contrary to expectation or negatively worded. Reflective tasks at this level may require connections, comparisons, and explanations, or they may require the reader to evaluate a feature of the text. Some reflective tasks require readers to demonstrate a fine

		understanding of the text in relation to familiar, everyday knowledge. Other tasks do not require detailed text comprehension but require the reader to draw on less common knowledge.
2	407	Some tasks at this level require the reader to locate one or more pieces of information, which may need to be inferred and may need to meet several conditions. Others require recognising the main idea in a text, understanding relationships, or construing meaning within a limited part of the text when the information is not prominent, and the reader must make low level inferences. Tasks at this level may involve comparisons or contrasts based on a single feature in the text. Typical reflective tasks at this level require readers to make a comparison or several connections between the text and outside knowledge, by drawing on personal experience and attitudes.
1a	335	Tasks at this level require the reader to locate one or more independent pieces of explicitly stated information; to recognise the main theme or author's purpose in a text about a familiar topic, or to make a simple connection between information in the text and common, everyday knowledge. Typically, the required information in the text is prominent and there is little, if any, competing information. The reader is explicitly directed to consider relevant factors in the task and in the text.
1b	262	Tasks at this level require the reader to locate a single piece of explicitly stated information in a prominent position in a short, syntactically simple text with a familiar context and text type, such as a narrative or a simple list. The text typically provides support to the reader, such as repetition of information, pictures, or familiar symbols. There is minimal competing information. In tasks requiring interpretation, the reader may need to make simple connections between adjacent pieces of information.

Note. Source: PISA 2015 assessment and analytical framework (OECD, 2017a, p.61)

Appendix C

PISA 2015 Mathematics Proficiency Levels

Level	Lower score limit	Task characteristics
6	669	At Level 6, students can conceptualise, generalise, and utilise information based on their investigations and modelling of complex problem situations, and can use their knowledge in relatively non-standard contexts. They can link different information sources and representations and flexibly translate among them. Students at this level are capable of advanced mathematical thinking and reasoning. These students can apply this insight and understanding, along with a mastery of symbolic and formal mathematical operations and relationships, to develop new approaches and strategies for attacking novel situations. Students at this level can reflect on their actions and can formulate and precisely communicate their actions and reflections regarding their findings, interpretations, arguments, and the appropriateness of these to the original situation.
5	607	At Level 5, students can develop and work with models for complex situations, identifying constraints and specifying assumptions. They can select, compare, and evaluate appropriate problem-solving strategies for dealing with complex problems related to these models. Students at this level can work strategically using broad, well developed thinking and reasoning skills, appropriate linked representations, symbolic and formal characterisations, and insight pertaining to these situations. They begin to reflect on their work and can formulate and communicate their interpretations and reasoning.
4	545	At Level 4, students can work effectively with explicit models for complex, concrete situations that may involve constraints or call for making assumptions. They can select and integrate different representations, including symbolic, linking them directly to aspects of real-world situations. Students at this level can utilise their limited range of skills and can reason with some insight, in straightforward contexts. They can construct and communicate explanations and arguments based on their interpretations, arguments, and actions.
3	482	At Level 3, students can execute clearly described procedures, including those that require sequential decisions. Their interpretations are sufficiently sound to be a base for building a simple model or for selecting and applying simple problem-solving strategies. Students at this level can interpret and use representations based on different information sources and reason directly from them. They typically show some ability to handle percentages, fractions, and decimal numbers, and to work with proportional relationships. Their solutions reflect that they have engaged in basic interpretation and reasoning.
2	420	At Level 2, students can interpret and recognise situations in contexts that require no more than direct inference. They can extract relevant information from a single source

		and make use of a single representational mode. Students at this level can employ basic algorithms, formulae, procedures, or conventions to solve problems involving whole numbers. They are capable of making literal interpretations of the results.
1	358	At Level 1, students can answer questions involving familiar contexts where all relevant information is present, and the questions are clearly defined. They are able to identify information and to carry out routine procedures according to direct instructions in explicit situations. They can perform actions that are almost always obvious and follow immediately from the given stimuli.

Note. Source: PISA 2015 assessment and analytical framework (OECD, 2017a)