Do teenage boys perform less well than teenage girls in literacy or do estimates of gender gaps depend on the test? A comparison of PISA and PIAAC

F. Borgonovi[1,2]

1 Social Research Institute, Institute of Education, University College London.

2 OECD Centre for Skills, Organisation for Economic Cooperation and Development.

Correspondence concerning this article should be addressed to:

Francesca Borgonovi, Social Research Institute, Institute of Education, University College London, 55-59 Gordon Square, London WC1H 0NU, UNITED KINGDOM

f.borgonovi@ucl.ac.uk

Keywords: literacy, gender gaps, testing, PISA, PIAAC.

Author's note: Francesca Borgonovi is a Senior Policy Analyst at the OECD, the organization that produces the PISA and PIAAC data.

Do teenage boys perform less well than teenage girls in literacy or do estimates of gender gaps depend on the test? A comparison of PISA and PIAAC

Abstract

Data from international large-scale assessments (ILSAs) of schooled populations indicate that young males have considerably poorer literacy skills than females. New evidence from a household based ILSA – the OECD Survey of Adult Skills (PIAAC) – indicates that the gender gap in literacy is negligible despite the fact that its assessment framework is very similar to that of one of the most widely used school-based assessments, the Programme for International Student Assessment (PISA). Using individual level data from 15, 16 and 17 year old males and females in countries that administered both assessments we estimate literacy gender gaps. We compare gender gaps in the two assessments after accounting for differences in target population, response rates, scoring scheme, test length, mode of delivery, the prevalence of items involving different stimuli in the two assessments (such as types of texts) and of cognitive processes test-takers need to engage in to solve assessment items (such as accessing and retrieving information or reflecting and evaluating information presented in the text). We find that these differences explain only part of the differences across the two studies in estimated literacy gender gaps: even when these factors are considered gender gaps remain large in PISA and small (though imprecisely estimated) in PIAAC. We discuss the potential role of test-taking motivation and administration conditions in explaining differences across the studies and implications for research and policy.

Keywords: literacy, achievement, gender gaps, testing, PISA, PIAAC, cross-country comparisons.

Educational Impact And Implications Statement

In this work, we compare gender gaps in the teenage years in literacy in two low-stakes international large-scale assessments: PISA and PIAAC. We find that estimates of literacy gender gaps in the two assessments are very different: young males significantly underachieve compared to young females in the PISA test but no gender gap can be identified in PIAAC. Our results suggest that before embarking on major policy reforms designed to ensure that young males acquire literacy skills based on their poor showing in the context of large-scale assessments as well as school tests, it would be important to evaluate if and how assessments reflect all of what young males know and can do, if assessments are comprehensive enough to capture dimensions of literacy that young males may be more proficient in and, crucially, if the assessments provide incentives for young males to show test administrators what they know and can do.

Do teenage boys perform less well than teenage girls in literacy or do estimates of gender gaps depend on the test? A comparison of PISA and PIAAC

International large-scale assessments (ILSAs) have been used extensively in academic research and education policy to identify gender gaps in achievement. The most widely known and used ILSAs to study gender gaps are school-based: the Programme for International Student Assessment (PISA), the Trends in Mathematics and Science Study (TIMSS), and the Programme for International Reading and Literacy Study (PIRLS). ILSAs have been used primarily to examine gender disparities in mathematics achievement (Breda, Jouini, & Napp, 2018; Else-Quest, Hyde, & Linn, 2010; Machin & Pekkarinen, 2008; Nollenberger, Rodríguez-Planas, & Sevilla, 2016; Stoet & Geary, 2018) but there is an increasing interest in the use of ILSAs to study gender gaps in literacy (Guiso, Monte, Sapienza, & Zingales, 2008; Legewie & DiPrete, 2012; Lietz, 2006; Lynn & Mikk, 2009; Reilly, 2012; Van Hek, Buchmann, & Kraaykamp, 2019).

Literacy is a fundamental pre-requisite for academic success and participation in society (OECD, 2010a; Cunningham & Stanovich, 1998; Smith, Mikulecky, Kibby, Dreher, & Dole, 2000) and numerous studies based on school-based achievement tests have indicated that male students have poorer literacy skills than girls (Buchmann, DiPrete, & McDaniel, 2008; Cole, 1997; DiPrete & Buchmann, 2013; Smith & Wilhelm, 2009). Similarly, males tend to obtain lower grades than females when their language and writing abilities are assessed at school (Voyer & Voyer, 2014).

In recent years, education policy makers have grown increasingly concerned about males' underachievement in literacy (Boys' Reading Commission, 2012; DiPrete & Buchmann, 2013; Kunnskapsdepartementet, 2019; Legewie & DiPrete, 2012; UNESCO, 2019) and ILSAs of schooled populations have been used to identify how severe the gender gap is and to examine how

institutional features and societal factors correlate with between-country differences in the severity of males' underachievement (Van Hek et al., 2019). Large-scale assessments involve large representative samples and have typically been developed to provide comparable data across languages and world regions. Therefore, results of such assessments are considered highly generalisable and with a high degree of external validity. Most research based on such data has considered generalisability to pertain to all aspects of the tests and consequently failed to critically evaluate what assessments measure, to clearly describe how broad constructs such as literacy are operationalised and tested (for example, the prevalence of different text stimuli, of items with different response formats) and the specific conditions in which the assessments were administered (for example, if assessments took place in a group setting, the nature of proctoring) and evaluated (for example, how item non-response was considered).

Evidence based on PISA, probably the most well-known school-based standardised assessment and the key source of data for most cross-country studies of achievement disparities, indicates that the gender gap in literacy[1] in favour of 15 and 16-year-old girls is large and has remained very large over the past two decades (OECD, 2001a; OECD, 2019a). Across OECD countries that participated in PISA since 2000, the standardised gender gap was $d = 0.32$ in 2000, $d = 0.39$ in 2009 and $d = 0.30$ in 2018 (OECD, 2010b; OECD, 2019a). But does the evidence from PISA and other school-based assessments indicate that teenage boys underperform compared to girls in literacy or that they underperform in the specific literacy test designed and implemented in the context of the PISA study?

Evidence from Nordic countries – Denmark, Finland, Norway and Sweden – indicates that gender gaps in literacy estimated in PISA (age 15) are considerably larger than gender gaps in

[1] Although some ILSAs refer to reading literacy, reading or literacy, in the context of this paper we will refer to literacy whenever discussing reading, reading literacy or literacy.

literacy for the same countries identified in PIRLS (age 10) and the OECD Survey of Adult Skills (PIAAC) (age 16-24) (Solheim & Lundetræ, 2018). Solheim and Lundetræ suggest that divergent results between the three studies might be due to age differences in the three surveys and/or to differences in test construction and assessment features such as the prevalence of assessment tasks in which either males or females tend to excel. Solheim and Lundetræ (2018) limited their analysis to Nordic European countries and although they identified differences across the three studies, they did not formally test the extent to which different assessment characteristics were responsible for the observed differences across the three studies in gender gaps in literacy.

The aim of our work is: 1) to extend the work of Solheim and Lundetræ by identifying differences between PISA and PIAAC in literacy gender gaps across a larger number of countries, considering only participants from the same age group who were tested in the same year – 2012, and 2) to develop analyses that allow to formally test the contribution of differences in sampling, in scoring method and differences in assessment features (item formats, text types, comprehension process) in explaining differences across the two assessments.

Although PISA and PIAAC assess mathematics/numeracy as well as literacy, and in fact the main domain tested in PISA in 2012 was mathematics, we examine literacy because the frameworks (i.e. what the tests intend to measure) in the two studies are more similar with respect to literacy than mathematics/numeracy (Gal & Tout, 2014). We wanted to minimise the role of differences in framework on estimated gender gaps. Moreover, gender gaps in literacy in the teenage years estimated in the context of PISA are generally considerably larger than gender gaps in mathematics/numeracy and appear to have been relatively stable over time when similar instruments have been used (OECD, 2015). Both PISA and PIAAC are low-stakes assessments: they do not have consequences for individuals taking part in the assessment and participants do not know their own results. Although the two studies differ in some respects (which are reviewed

and examined at length in the paper), assessment instruments and goals are aligned, so much so that both studies – PISA in 2000 and PIAAC in 2012 – administered some assessment tasks originally developed for the International Adult Literacy Survey (IALS).

The role of motivation in shaping achievement results

Recent evidence suggests that test-taking motivation plays a key role in determining achievement in low-stakes assessments (Borgonovi & Biecek, 2016; Wise & DeMars, 2005, 2010; Wolf, Smith, & Birnbaum, 1995). Motivation to take a test (which affects response rates) and test-taking motivation during the test (which influences test-taking effort and engagement) have long been recognised as important determinants of participation and performance in low-stakes standardised assessments. In recent years research has examined test-taking effort and motivation in low-stakes assessments to explain between country differences in test results (Borghans & Schils, 2012; Borgonovi & Biecek, 2016; Gneezy, List, Livingston, Qin, Sadoff, & Xu 2019; Zamarro, Hitt, & Mendez, 2019) as well as differences across key population groups (Balart & Oosterveen, 2019; Borgonovi & Biecek, 2016).

According to expectancy value theory, the motivation to participate in a test and the motivation exerted during the test depend on individuals' expected performance and task value (Eccles & Wigfield, 2002). Test-taking motivation (Baumert & Demmrich, 2001) can therefore be considered to depend on perceived success on a given test, beliefs about the amount of effort the test will consume, the perceived importance of the test, and affective reactions to individual test items and stimuli (Wise & DeMars, 2005). Cognitive skills, self-efficacy and self-concept determine students' expected task performance and shape students' convictions that they can successfully perform at designated levels (Schunk, Pintrich, Meece, & Pintrich, 2008). Task value is determined by intrinsic and instrumental rewards associated with participation, such as the

opportunity cost of time, the rewards associated with test completion as well as attitudes about working hard.

In high-stakes assessments, motivation to take the test is driven primarily by the stakes associated with participation and it is generally assumed that test-takers will invest their maximum effort for the duration of the test. By contrast, in low-stakes assessments, such as those used in our work, motivation to participate and test-taking motivation during the test are likely to be more variable and to be determined by how the assessment shapes performance expectancies and task value (Barry, Horst, Finney, Brown, & Kopp, 2010; Chan, Schmitt, DeShon, Clause, & Delbridge, 1997; Cole, Bergin, & Whittaker, 2008; Eklöf, 2010; Wise & DeMars, 2010). Moreover, when incentives are low, motivation is likely to vary across groups that differ in expected performance and task value (Braun, Kirsch, & Yamamoto, 2011; Wolf et al., 1995).

In line with findings from Solheim and Lundetræ (2018) we expect to identify differences in the literacy gender gap estimated in PISA and PIAAC. In the following paragraphs of this section, we review in detail evidence that males' level of achievement and motivation is more context dependent than females'. This evidence leads us to expect that such difference will be primarily driven by a difference in the estimated performance of males in the two assessments. In other words, our first hypothesis is that gender gaps in the two assessments will be different. We expect these differences to be driven by differences in the achievement of males in the PISA and the PIAAC assessments while we expect females' achievement to be relatively stable across different assessment conditions. Different tests promote different levels of motivation and engagement and we expect males to be more susceptible to motivational drivers.

The first mechanism through which motivation could influence estimates of gender gaps in different assessments is by shaping the participation of a different pool of males and females (selection mechanism). PISA is a school-based assessment with very high response rates among

its target population (15 and 16-year-old students) while PIAAC is a household assessment covering 16-65 year olds and has lower response rates (see the Supplementary Online Annex for detailed descriptions of the PISA and PIAAC surveys, methods used to make the two tests as comparable as possible and conduct the analyses presented in the manuscript). No systematic differences in participation by gender can be observed, neither in PISA nor in PIAAC. However, differences in motivation to participate across the two studies may occur. Opting out may be more difficult in a school setting where teachers and school principals may encourage students to participate and where the opportunity cost of participating is skipping classes, not necessarily the most favourite activity of 15-year-olds. Moreover design features (PISA covers only students) prevent some individuals from participating (although in the countries considered in this work schooling is compulsory beyond the age of 15 so school dropout is low). Consequently, our second hypothesis is that the greater sample selectivity of the PIAAC study determines a smaller gender gap in the PIAAC than in the PISA literacy assessment and that the gender gap is the same once sample selectivity is accounted for.

The second mechanism through which motivation may influence estimates of gender gaps in different assessments is by shaping effort and engagement exerted by participants during the assessments. Even assessments that are relatively well aligned like PISA and PIAAC differ on a number of characteristics that could influence engagement with the test and effort exerted while answering the test. Crucially, males and females may respond differently to assessment characteristics and, as a result, exert a different amount of effort in different tests and display a different level of engagement with different assessment tasks.

The literature has indicated that males' achievement tends to vary more between high and low stakes assessments than females' achievement, a reflection of the higher effort exerted during the assessment and engagement with assessment tasks among males in high-stakes assessments

(Braun et al., 2011; Coffman & Klinowski, 2018; Gneezy, et al., 2019). Lower conscientiousness and greater reliance on extrinsic motivational drivers to guide behaviour among males suggest that even if males and females were to possess similar levels of ability, males may be more likely than females to skip tasks and fail to provide answers in low-stakes assessment settings. Gender differences in the propensity to skip answers, a reflection of gender differences in motivation during the test, may determine differences in the PISA and PIAAC assessments because items with no response were scored differently in the two tests: they were considered as wrong in PISA but missing in PIAAC. When unattempted answers are scored as wrong, imputed levels of proficiency are lower than when unattempted answers are considered missing, thereby leading to lower estimated proficiency in PISA than in PIAAC among individuals who left many assessment tasks unanswered. Theory and prior empirical evidence suggests that males will be over-represented in this group. Therefore, our third hypothesis is that differences in scoring methods between PISA and PIAAC determine a smaller gender gap in favour of females in the PIAAC than in the PISA literacy assessment and that gender gaps across the two assessments will be similar once the same method for scoring non answered items is used.

Maintaining high levels of accuracy and alertness during a long test requires motivation and the exercise of self-control (Baumeister, Heatherton & Tice, 1994) and the literature indicates that males tend to display lower levels of conscientiousness (Duckworth & Seligman, 2006; Matthews, Ponitz, & Morrison, 2009). Prior empirical research has demonstrated that males' engagement with assessment tasks has been shown to decline rapidly during long assessments (Balart & Oosterveen, 2019; Borgonovi & Biecek, 2016). As a result of differential effort expended over the course of long assessments by males and females, males' achievement at the start of the assessment tends to be considerably higher than their achivement towards the end of the assessment, while females' level of accuracy tends to be more stable (Borgonovi & Biecek,

2016). Because PISA is a considerably longer assessment than PIAAC, our fourth hypothesis is that the shorter nature of the PIAAC test determines a smaller gender gap in favour of females in the PIAAC than in the PISA literacy assessment and that gender gaps across the two assessments will be similar once comparisons are conducted using tests of similar length.

A combination of motivation and skill profile is likely to underlie the variation in gender differences in literacy depending on mode of administration across assessments (i.e. digital literacy vs. printed literacy assessments). Young males tend to have greater familiarity with technology and to be more attracted to technology than young females: among recent cohorts of adolescents, males generally reported having started to use computers and the internet at a younger age and to spend more time on digital devices than females (OECD, 2015). Young males also report greater self-efficacy using digital tools than young females (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014). These factors may induce greater motivation among males when completing assessment tasks on a computer. There is also evidence that females may perform less well than males in solving problems on digital technologies, may have poorer navigation skills and may be less interested in ICT related skills than males (Zhou, 2014). Prior literature has observed that literacy gender gaps tend to be smaller especially in tasks that require digital reading skills (such as clicking on hyperlinks, clicking on tabs to find information, scrolling etc…) (Borgonovi, 2016; Støle, Mangena, & Schwippert 2020). Because the PIAAC assessment was delivered on computer and the main PISA assessment in 2012 was delivered using paper-based instruments, our fifth hypothesis is that differences in mode of administration will determine a smaller gender gap in favour of females in the PIAAC than in the PISA literacy assessment and that gender gaps in the two assessments will be similar when considering assessments administered using the same mode.

Males' literacy achievement has been shown to vary to a larger extent than females' across contexts, a reflection of males' more variable level of engagement with texts that cater (or not) to their interests (Taube & Munck, 1996). Our sixth and final hypothesis is that the larger gender gap in favour of females observed in the PIAAC than in the PISA literacy assessment will be due to the different prevalence of assessment tasks involving different types of texts in the stimulus or requiring different cognitive processes to be solved.

In order to consider more directly the role of motivation in shaping gender gaps, we exploit information from the PISA background questionnaire in which students were asked to report how much they agreed/disagreed that school is a waste of time. We compare the gender gap estimated in PISA among individuals who reported that they consider school as a waste of time and those who do not. Furthermore, we provide individual level evidence on the size of the gender gap in literacy in Canada when the test was delivered to 15 year olds and when the same test was delivered to the same individuals in their homes nine years later. We use external evidence on 24 year olds from PIAAC in Canada to indirectly account for ageing effects.

Method

Data sources

The Programme for International Student Assessment (PISA)

PISA is a triennial large-scale low-stakes standardised assessment conducted since 2000 and targeting the schooled population of children between the ages of 15 years and three months and 16 years and two months at the time of administration. Each PISA cycle assesses three core domains (reading, mathematics and science) although in each cycle one domain is considered the main domain and, as a result, is examined in greater depth in the assessmenr. Students take the test in a class with other students under supervision and after they finish the test

they are asked to complete a questionnaire. Our main analyses rely on data from the PISA 2012 study although we use data from PISA 2000 when examining longitudinal evidence from Canada. Data are publicly available from http://www.oecd.org/pisa/data/.

The OECD Survey of Adult Skills (PIAAC)

PIAAC is a low-stakes assessment that was primarily administered in 2012 (additional administration rounds were organised in 2015 and 2017). The PIAAC instruments were designed to be comparable with IALS and the Adult Literacy and Lifeskills Survey (ALL). The PIAAC target population includes all non-institutionalised adults between age 16 and 65 (inclusive) whose usual place of residence is in the country of assessment at the time of data collection. Key assessment domains in PIAAC are literacy, numeracy and problem solving in technology rich environments. PIAAC is a household-based study. Trained interviewers first administered the background questionnaire which was conducted using Computer Assisted Personal Interviewing (CAPI). The interviewers then handed the direct assessment to respondents.

Materials

The PISA assessment is timed and is designed to take around 2 hours to complete. Testing material is organised around subject specific clusters (domains tested in the main 2012 administration were literacy, mathematics and science) that take around 30 minutes each to complete, and each testing booklet that students receive contains four clusters of test items, and different booklets contain a different selection of clusters. After the end of the assessment session, students take a short break and then complete a background questionnaire designed to take around 30 minutes to complete (questionnaire completion is untimed). In 2012 the main assessment domain in PISA was mathematics. This means that in 2012 the mathematics item pool is greater than either the literacy or science item pool. Because PISA uses a random matrix design, all

students are administered the same overall amount of testing material, i.e. each booklet contains four clusters, but different students receive different clusters, meaning that they receive different combinations of test items in different subjects. In 2012 around 46% of students were administered a booklet containing one cluster of literacy items with the rest being a combination of math and science items; 23% of students were administered a testing booklet containing two literacy clusters while the remaining 31% were not administered any literacy item. By contrast, no student was not administered any mathematics item, 46% were administered three mathematics clusters, 23% were administered two mathematics clusters and the remaining 30% were administered one cluster of mathematics items. Among those being administered at least one cluster of literacy test items, around a quarter were administered literacy items at the start of the assessment, another quarter in the early middle part, another in the late middle part and the remaining quarter at the end of the assessment.

The core test and questionnaires in PISA 2012 were delivered through paper booklets. However, in a number of countries an optional computer-based assessment of literacy, mathematics and problem solving was administered. The computer-based study always post-dated the administration of the paper-based study. In most countries it was administered in the afternoon of the same day of the core test while it was administered the day after in the Slovak Republic and within a week in Italy. The computer-based test was timed, with a total length of 40 minutes. The test was organised around two clusters designed to take twenty minutes to complete each.

In PIAAC the questionnaire was administered first and took around 40 minutes to complete on average while the assessment took slightly less than an hour. However, no time limit was imposed and respondents could use as much time as needed to complete the test. Testing material in PIAAC was organised around two subject specific clusters of equal expected length. Delivery was conducted on a computer, although individuals who lacked familiarity with a

computer (or a willingness to sit a test with a computer) were offered a paper-based version of the test. A mode effect study was conducted and findings indicate that mode of delivery did not affect response rates or accuracy (OECD, 2013a, Chapters 18 and 19). Response rates varied greatly across countries, but they ranged between 45% in Sweden to 75% in Korea. Response bias analyses conducted to validate the quality of the PIAAC data indicate that non-respondents share common background characteristics to respondents (OECD, 2019b, Chapter 16). Furthermore, hard to reach individuals (defined as those for whom several contact attempts by the interviewers had to be made to achieve participation) did not have different levels of literacy and numeracy from those individuals whose participation did not require additional effort. These may be due to compensating effects: low skilled individuals may be less willing to sit a test because they might fear test-like situations. However, the opportunity cost of time is higher among the highly skilled, who therefore may be less willing to participate in a survey like PIAAC. The majority of participants in PIAAC were administered a computer-based assessment.

The PIAAC assessment design is more balanced than PISA: two thirds of the overall sample were administered one literacy module and the rest was only administered problem-solving test items. Out of participants administered literacy items, half were administered literacy items at the start of the test and half were administered literacy items after they had completed either a numeracy or a problem solving module. The PIAAC test was partially adaptive, meaning that individuals were assigned test items of different expected difficulty with a different probability determined by background characteristics (such as educational attainment and immigrant status) as well as achievement in previous parts of the test. For example, individuals with tertiary level qualifications were more likely than individuals with upper secondary qualifications to be assigned a module consisting of a larger pool of difficult assessment items. Yet, some individuals with tertiary level qualifications were assigned modules with a larger pool

of easier test items and individuals with upper secondary qualifications were assigned modules with difficult items. Detailed information on the PIAAC adaptivity routing scheme is available in OECD (2019b, Chapter 1). In the section Analytic Strategy we detail how we account for item difficulty to ensure that the adaptive design does not influence results.

The PIAAC and the PISA literacy frameworks are very similar. Both share the same (action-oriented or functional) definition of skills. They share a common approach to the specification of constructs, a comparable definition of measured abilities, similar content definitions and contexts in which tasks are embedded [for more details see (OECD, 2013b)]. In fact, many of the international experts involved in the development of PISA were also involved in the development of PIAAC.

Although the assessment frameworks are very similar in the two studies, the prevalence of texts used in the stimulus and of tasks requiring different response formats or designed to identify specific cognitive processes differ across the two studies. For example, non-continuous texts comprise over 50% of stimulus material for test items in PISA but only 2% of texts in PIAAC. By contrast, multiple and mixed texts make up the majority of stimuli used in test items in PIAAC (77% in combined terms) but only around a quarter of texts in PISA. Tasks requiring individuals to access and retrieve information comprise only 23% of test items in PISA, but over 55% in PIAAC. Similarly, in PISA 55% of the literacy items require constructed responses while in PIAAC there were no constructed responses as such: almost 90% of items required individuals to click on the correct answer, highlight a piece of text to give an answer, or respond to multiple choice questions. Only in 12% of test questions individuals had to enter text or a number to provide an answer and since answers were computer coded, no extensive writing was involved. Tables S5 and S6 in the Supplementary Online Annex identify the prevalence of different assessment items or items relying on different stimuli in PISA and PIAAC.

Both PISA and PIAAC describe the items used in the assessments according to stimuli used (i.e. if the text used in the stimulus was a continuous, non-continuous, or mixed text), type of response required (i.e. constructed, multiple choice) and cognitive process involved (access and retrieve information, integrate and interpret information, and evaluate and reflect upon information). Most assessment tasks used in the PISA and PIAAC administration are confidential and therefore are not publicly available. They are part of an item bank used in successive administrations and/or other assessments. However, a number of sample tasks are disclosed to illustrate the range of tasks involved and their difficulty. Sample tasks from the PISA and PIAAC literacy assessments cannot be reproduced but are available from OECD publications (OECD, 2013c, pp. 203-213 for PISA and OECD, 2013d, pp. 22-25 for PIAAC). Crucially, a study designed to establish concordance scores between the PISA and PIAAC literacy assessments indicates that the overall level of difficulty of the two tests is comparable; that both PISA and PIAAC include a range of items at easy, medium and high level of difficulty for a target population of youngsters similar to the target population in this study; and that a similar distribution of literacy abilities can be retrieved in a population of youngsters whether the PISA or the PIAAC literacy assessment is used (Pokropek & Borgonovi, 2019).

Participants

We consider test takers from the 26 countries that took part in PISA in 2012 and that took part in PIAAC in either the 2012 round or in the 2015 round (2015 countries are denoted with an * next to the country name). We decided to include PIAAC 2015 countries to increase country coverage and the sample size of the PIAAC samples. Countries considered are: Australia, Austria, Belgium, Canada, Chile*, the Czech Republic, Denmark, Estonia, Finland, France, Germany, Ireland, Italy, Japan, Korea, Lithuania*, the Netherlands, New Zealand*, Norway, Poland, the Slovak

Republic*, Slovenia*, Spain, Sweden, Turkey*, Singapore*, England and Northern Ireland and the United States.

We only consider 16 and 17 year-olds in PIAAC and 15 and 16-year-olds in PISA (estimates are very similar when we restrict the PISA sample to 16 year olds, but standard errors are larger because the sample is reduced by around 75%). We do not consider students who sat the PISA *une heure* booklet (administered in some countries to individuals with special education needs or who just arrived in their country of residency and therefore have limited language abilities). The two samples are highly unbalanced: our working sample is 225933 individuals for PISA and 7495 for PIAAC. We report 95% confidence intervals for all our estimates and test for differences in estimates across the two studies for key specifications (Schenker & Gentleman, 2001; Wolfe & Hanley, 2002).

Around 3% of PIAAC participants failed a core module of literacy and numeracy tasks designed to identify if they had at least the basic reading and mathematical skills needed to take part in the full assessment. These test-takers were administered a reading components module rather than the full PIAAC assessment. Reading components are the basic set of decoding skills that are essential for extracting meaning from written texts: knowledge of vocabulary (word recognition), the ability to process meaning at the level of the sentence, and fluency in reading passages of text. The exclusion of individuals with very low levels of literacy from the main PIAAC administration could potentially bias results of a PISA-PIAAC comparison since these individuals would be excluded from PIAAC but not PISA. However, there are no gender differences in the prevalence of individuals with very low literacy skills: 50.1% of individuals failing the core literacy test are men and 49.9% are women and virtually no participant among the youngest cohort - the focus of this work - was part of the group classified as having difficulties with basic text decoding skills (Grotlüschen, Mallows, Reder, & Sabatini, 2016).

A second difference between PISA and PIAAC is that PISA is limited to students while the PIAAC target population is not limited to individuals who are in education. However, because compulsory schooling legislation mandate participation in education until the age of 17 in the vast majority of the countries considered in our analysis and the fact that many 16 and 17 year olds are still in school even in the absence of compulsory schooling legislation, the vast majority of PIAAC respondents in our sample reported being in school. We provide results for the full PIAAC sample and the subsample of PIAAC test takers who report being in school in Table 2. As an additional test, because those who abandon school at the age of 16 or 17 are generally among low achievers (the difference in the PIAAC score in reading/literacy between 16 and 17 year olds who were not in education and those who were was 94% of a SD), in Table 2 we present gender gaps in PISA on the subsample of respondents which excludes the lowest achieving 5% of males and lowest achieving 5% of females (to match the PIAAC distribution).

Response rates are also very different across the two studies: response rates are higher in PISA than in PIAAC. In PISA 2012 the student level response rate was above 90% for the core assessment. PISA uses a two-stage sampling design: schools are selected first, with the probability of selection being associated with the number of 15-year-old students attending the school. Within selected school a sample of 15-year-olds is randomly drawn. Participating countries aim to achieve a minimum of around 150 participating schools and a total of around 4500 participating students (30 students per school). However, in a number of countries samples were larger, in order, for example, to examine regional variation (the student sample was over 30 000 students in Italy and over 20 000 students in Canada and Spain). Achieved samples in PIAAC varied from 3 761 in Northern Ireland to 27 285 in Canada where the sample was designed to provide reliable estimates at provincial level as well as for a range of subgroups of the population such as the indigenous population and linguistic minorities.

The literature suggests that males' performance is more variable than females' (Lindberg, Hyde, Petersen, & Linn, 2010; Machin & Pekkarinen, 2008). Although non response bias analyses indicate that there is no difference in response rates between males and females and individuals with different levels of educational attainment in PIAAC (OECD, 2019b, Chapter 16), it is possible that the PIAAC sample may exclude lowest achievers. Even if the same proportion of lowest achieving males and females did not take part in PIAAC but took part in PISA, excluded males could be expected to have a considerably lower performance than excluded females, thereby leading to a relatively large gender gap in favour of females in PISA and a narrower gap in PIAAC.

Analytic Strategy

The standard proficiency scales in PISA and PIAAC are estimated using Item Response Theory (IRT) models. Individual responses on the assessment are combined with background information to estimate, for each respondent, a distribution of proficiency, from which a set of plausible values is drawn (Jacob & Rothstein, 2016). Scaling depends on the set of countries that take part in the assessment, as well as on the specific IRT models used (a one-parameter Rasch model for PISA, a two-parameters model for PIAAC). In PISA 2012 a set of five plausible values was drawn while in PIAAC a set of ten plausible values was drawn. IRT models are a form of multiple imputation and allow to derive precise estimates of ability considering the likely responses test-takers would have provided had they been administered the entire pool of assessment items by considering observed relations between test items and the distribution of responses across students with different background characteristics.

Descriptive statistics

In order to test the first hypothesis, we report descriptive statistics extending findings from Solheim and Lundetræ (2018) to the set of countries that participated in both PISA and

PIAAC. Given the country specific focus in Figure 1 we extend the age range of the PIAAC sample to include 16 to 20 year olds. We then map the the distribution of gender differences among low-achieving, high-achieving and middle achieving individuals. PISA and PIAAC identify levels of competencies that individuals at different levels of proficiency can be expected to achieve (OECD, 2013c; OECD, 2016). Although there is a high degree of overlap in the way in which PISA and PIAAC identify expected proficiency levels and the types of tasks individuals at different levels can be expected to perform, PISA and PIAAC proficiency scales are not directly comparable. We exploit recent evidence on linking conversion scores to derive the distribution of proficiency of males and females participating in PIAAC using the PISA proficiency levels (Pokropek & Borgonovi, 2019). Results are reported in Figure 2.

Next, we turn to identifying if discrepancy in results can be explained by sample selection, test length, coding scheme, mode of delivery, and the prevalence of items relying on different texts in the stimuli or involving different cognitive processes.

Selection mechanisms

In order to test second hypothesis, we analyse the effect of selection using the PISA and PIAAC plausible values and combine them using Rubin's rule (Rubin, Wiley, York, Brisbane, & Singapore, 1987). Because of differences in scaling, in order to compare gender gaps across the two studies, in Figure 1 and Table 2 in the main text we present Cohen's *d* statistics. Results are virtually identical when effect size are estimated using Hedges' g. All models were estimated controlling for country of administration fixed effects. In Figure 3 and Table 2 we present several analyses comparing the gender gap in PIAAC and the gender gap in the PISA. First we remove individuals in PIAAC who were not in education at the time of the assessment, to make the PIAAC sample comparable to the PISA sample. Second, we remove from the PISA sample the 5% of the lowest achieving students according to national specific distributions of achievement as

these are those most likely to drop out between the age of 15 and the age of 16-17 (the PIAAC sample age). Next, we compare gender gaps in the PISA subsample that excludes the lowest achieving 25% of males and the lowest achieving 25% of females to match response rates in PISA with response rates in PIAAC. This should represent an upper bound estimate of the response rate effect on estimated gender gaps since it implies that all individuals who do not take part in PIAAC achieve at the lowest levels of proficiency.

Test-taking mechanisms

To account for differences between PISA and PIAAC in assessment characteristics or stimuli used in the item we employ item level data from the PISA and PIAAC tests and consider the responses test takers give to specific sets of questions that are similar across the two surveys. We do so in order to estimate gender gaps considering only similar questions in PIAAC and PISA. Because in these sets of analyses the outcome variable is dichotomous (it represents if a respondent gave a correct answer to a particular test question), we fit logistic regression models to estimate the probability of success for females compared males, controlling for country of administration fixed effects. Our primary aim is to compare estimates across samples (PISA and PIAAC, different item pools) and across models. It has been suggested that odds ratios cannot be compared meaningfully across samples and models (Mood, 2009), therefore we present both odds ratios as well as average marginal effects.

Although items in PISA are randomly allocated to respondents, the adaptive nature of the PIAAC test means that different individuals can be assigned test items of different level of difficulty depending on their background characteristics and response patterns during the test. Therefore, for PIAAC, we present estimated unadjusted results as well as results adjusted for item difficulty. Estimates that account for item difficulty net out potential gender differences in the motivational response to being administered assessment items that are more likely to fall close to

one's proficiency level (like in PIAAC) or that can be in line, above or below one's level of proficiency (like in PISA). In 2018 PISA adopted an adaptive design similar to the one adopted in PIAAC and results from a validation study indicate that such design had no effect on estimates of proficiency by gender (i.e. estimates of gender gaps derived when adopting an adaptive test were similar to those estimated using a non adaptive design) while improving efficiency of the assessment instruments (more information could be acquired keeping the same burden for test-takers) (OECD, 2020).

All models take into account the complex survey design of PISA and PIAAC: PISA results were estimated using final student weights as well as balanced repeated replicate weights while for PIAAC, all estimates were derived applying jackknife replicate weights. We consider the following features to select pools of items to derived standardised estimates of gender gaps: coding scheme; test length; delivery mode; type of text used in the stimulus and cognitive process required to solve the item.

Coding scheme: treatment of non-reached or not answered items

We test the third hypothesis by examining differences in gender gaps estimated using PISA and PIAAC when non answered answers are scored similarly in the two tests. In PISA when test takers do not provide an answer to a question, an incorrect answer code is assigned and non-reached items (i.e. unanswered items at the end of test booklets) are considered as wrong when estimating student proficiency (i.e. in the "scoring" step) but as not administered when estimating item parameters (in the "scaling" step). By contrast, in PIAAC when test takers do not provide an answer to a question, a missing information code is assigned and non-reached items are treated as not administered both when estimating student proficiency and item parameters. Because in PISA and PIAAC no penalties are assigned for wrong answers, the PISA "coding scheme" means that

achievement is typically higher when individuals attempt to provide an answer because there is a non-zero probability that this will be correct. Differences in coding scheme have two, potentially contrasting effects on estimates of gender gaps estimated in the two studies. Disengaged respondents, i.e. those who leave many questions unanswered, are particularly penalised in terms of estimated proficiency in PISA compared to PIAAC because, other things being equal, the same behaviour – leaving a task unattempted – is associated with lower proficiency in PISA than in PIAAC. Among engaged respondents, the PISA approach benefits test takers who engage in guessing behaviour. The PIAAC approach eliminates potential advantages for test takers who randomly guess answers but also reduces the influence of test engagement on the final score: only information provided by engaged respondents is considered. If males have lower level of test engagement and conscientiousness, they may have lower than expected results in the PISA coding scheme condition. If males are more likely to engage in guessing behaviour (Baldiga, 2014), they may be advantaged by the PISA coding scheme. All analyses present two sets of results: the first set considers non reached and not answered items as wrong (the PISA coding scheme) while the second set considers non reached and not answered items as missing.

Test length

We test the fourth hypothesis by examining differences in gender gaps estimated using PISA and PISA in hypothetical tests truncated to have the same length. As illustrated in the section Materials, PISA is a timed two-hours assessment while PIAAC is an untimed assessment designed to take around 40 minutes to complete. In PISA the background questionnaire follows test administration while in PIAAC the assessment takes place after questionnaire administration. The differential length of the test may give rise to test engagement effects and fatigue effects that differ across genders. The shorter length of the PIAAC assessment may therefore lead to a smaller

estimated gender gap in favour of females in PIAAC than in PISA. However, because the assessment in PIAAC follows the administration of the background questionnaire, constructing a comparison between PIAAC and PISA that accounts for length is challenging.

When considering test length effects we consider only literacy items administered in PISA clusters one and two (first 30 minutes of testing time and testing time starting after around 30 minutes of other PISA test material) and PIAAC literacy module one (first part of the PIAAC test) or module two (second part). Several comparison groups can be identified: if the administration of the background questionnaire is ignored, it is possible to compare gender differences in the probability (or the odds) of giving correct responses when a reading cluster in PISA and a literacy module in PIAAC was administered in the first position, effectively comparing only literacy materials administered at the start of the test. It is also possible to examine gender gaps after a longer elapsed time and compare gender gaps in the probability of giving a correct response of participants in PISA who were administered a literacy cluster in the second position and PIAAC respondents who were administered a literacy module in the second position. If the administration of the background questionnaire in PIAAC is considered, the more appropriate comparison would be between participants in PISA who were administered a literacy cluster in the second position (after around half an hour of assessment) and PIAAC respondents who were administered a literacy module in the first position (since these participants had already taken part in the questionnaire before the first part of the test).

Delivery mode: paper-based vs. computer-based

We test the fifth hypothesis by examining differences in gender gaps estimated using PISA and PIAAC in hypothetical tests delivered using similar mode. In PISA the core assessment instruments were paper-based while in PIAAC they were computer-based. It is therefore possible

that difference in gender gaps across the two assessments may reflect gender specific preferences for and ability to complete tests delivered on computer vs. paper. In 2012, countries participating in PISA where offered the opportunity of administering, on top of the paper-based assessment, a computer-based assessment of digital literacy (OECD, 2013b). We provide results for computer-based administered tests by reporting the gender gap in the optional PISA 2012 computer-based literacy assessment and the gender gap estimated in PIAAC. We complement these analyses by exploiting the fact that PIAAC was designed to be comparable with previous adult skills assessments (the International Adult Literacy Survey -IALS- and the Adult Literacy and Lifeskills Survey -ALL- studies), which were paper-based. As many as 62% of the literacy items in PIAAC were items originally designed to be delivered in the context of paper-based administration but that were adapted to computer delivery. A mode effect study was conducted to identify if the switch to computer rendered the PIAAC assessment not comparable with prior paper-based assessments (OECD, 2016) and results supported comparability (in fact, a paper-based version of the assessment was delivered to test-takers who were either unable or unwilling to take the assessment on a computer). We present estimates of gender gaps that comparing the PISA 2012 main paper-based assessment with the subset of PIAAC instruments that were originally developed for paper administration but were delivered via computer.

Text type and cognitive demands

We test the sixth hypothesis by examining differences in gender gaps estimated using PISA and PISA in hypothetical tests containing only items involving similar text types in the stimulus or only items requiring test-takers to engage in similar cognitive processes. In the Supplementary Annex Tables S5 and S6 we report study-specific gender differences for each of the characteristics of the stimulus material defined by the PISA reading and the PIAAC literacy

experts groups. The prevalence of items involving different stimuli or other characteristics (text type, response format, cognitive process involved in solving a task) was hypothesised by Solheim and Lundetræ (2018) as being responsible for the differences they observed in the literacy gender gaps estimated in PISA and PIAAC in Nordic countries.

Because the distribution of test items across the two studies is different (see a description in the Materials section) and the sample size in PIAAC is small, estimates comparing gender gaps for items involving similar stimuli or that share similar characteristics across the two studies are generally very imprecise (very large confidence intervals) and therefore we decided not to report all results on differences across text types and cognitive demands in the main text (results are available in the Supplementary Online Annex). However, it is possible to develop relatively precise comparisons across the two studies for certain items that are well represented in both studies. More specifically, we report comparisons of gender gaps in PISA and PIAAC when considering literacy items involving stimuli material comprising only mixed text types (a type of literacy text) and when considering only items involving tasks that require individuals to access and retrieve information (a cognitive process required to solve literacy items in the two assessments).

Mixed texts are texts that contain elements of both continuous and non-continuous texts. Continuous texts consist of sentences formed into paragraphs. Examples for this format are novels, newspaper reports, or e-mails. Non-continuous texts rather use typographic feature to organise information. Tables, graphs or forms are classified as non-continuous texts.

The PISA and PIAAC literacy assessment tasks cover three cognitive processes that readers use in order to approach written texts: accessing and retrieving information, integrating and interpreting information and reflecting and evaluating information (PIAAC Literacy Expert Group, 2009). The “access and retrieve” aspect in PISA and PIAAC refers to selecting, locating

and retrieving one or more information from a text. By contrast, the "integrate and interpret" aspect involves processing what has been read and assigning a meaning to it and the "reflect and evaluate" aspect involves using knowledge and attitudes outside the text, relating these to text content and making judgements based on the overall information acquired.

Remaining assessment differences: the Canadian PISA retest and the role of attitudes towards school in PISA

As a robustness check to the main analyses, we use published results from Canadian data from the Youth in Transition Survey (YITS), which followed the students who participated in the first PISA assessment in 2000 through to their young adulthood (OECD, 2012). At two-year intervals, the original PISA respondents were contacted and asked to provide information on their activities related to education and employment, their life choices, and their attitudes. In 2009, on top of standard questionnaire-based instruments, they were (re)administered a PISA like test. The 2009 sample of YITS was representative of the population of 15-year-old Canadian students in 2000 and involved approximately 2 000 individuals. Of these 1 297 respondents took the assessment, which was conducted during May-June 2009 and consisted of a follow-up assessment of readings skills (but not mathematics skills) and a background questionnaire. The assessment was conducted in people's homes, much like PIAAC but the assessment was very much like PISA. The PISA-YITS longitudinal assessment used a selection of assessment questions known as the PISA link items. This selection of test items was also used for testing reading as a minor domain in PISA 2003 and PISA 2006 and allowed for trend analyses. This ensured that the two tests not only describe the same underlying construct but are, in fact fully comparable. The PISA-YITS assessment results were scored in conjunction with the main PISA assessment in 2009. Since the PISA-YITS items were also included in the PISA 2009 assessment, qualified coders

who scored the PISA 2009 test booklets also scored the PISA-YITS test items. Adjusted weights were included in the final data, ensuring that the sample remained representative (OECD, 2012).

The Canadian study allows to identify for the same individuals performance in the PISA test as well as performance in a PISA test conducted under PIAAC-like conditions (namely one to one proctoring in people's homes). However, the disadvantage of the Canadian study is that results could be due to an ageing effect (respondents were 9 years older when they sat the PISA-YITS study) and a cohort effect (the cohorts observed in PISA 2000 and PISA 2012 are not the same and the PISA-YITS cohort is not the same as the PIAAC 2012 cohort of 16-17 year olds). While these problems cannot be adequately dealt with, we present gender gaps in PISA 2000, PISA 2012, PISA-YITS, PIAAC 2012 (16-17 year olds sample) and PIAAC 2012 (23-24-25 year olds sample) to illustrate how cohort and ageing effects are unlikely to be driving differences across the two studies.

As a final analysis to identify the role of motivation and engagement, we use questionnaire data from students participating in the 2012 PISA study, when students were asked to report if they strongly agreed, agreed, disagreed or strongly disagreed that school is a waste of time, an indicator of lack of motivation at school. We consider gender differences in the share of males and females with positive attitudes towards school, i.e. who report disagreeing or strongly disagreeing that school is a waste of time, consider the literacy proficiency of engaged and disengaged students, and if the literacy gender gap among engaged individuals is smaller than the gender gap among disengaged individuals.

Results

Descriptive evidence: gender gaps in PISA and PIAAC

Table 1 presents gender specific descriptive statistics for the PISA and PIAAC studies as well as sample size of the two studies by gender. Figure 1 presents literacy gender gaps (females-

males) in PISA and PIAAC. Contrary to what one would expect to find and in line with the study by Solheim and Lundetræ (2018), estimates obtained using the two surveys are very different: if one were to consider only evidence from PISA, findings reported in Figure 1 would indicate that young males importantly underachieve compared to young females in all countries. However, if one were to consider only evidence from PIAAC, the same figure would suggest that there is no gender gap in literacy (although lack of precision because of the small sample size in PIAAC means that estimates could range from small negative to small positive).

TABLE 1

FIGURE 1

In Figure 2 we investigate the distribution of literacy achievement among males and females in PISA and PIAAC to identify the nature of males' underachievement and the extent to which any differences in average achievement can be traced to a differential representation of males and females among low, high and middle achievers. Results indicate some differences in the prevalence of males and females at different levels of achievement in the two tests. In particular, while in line with our first hypothesis the distribution of females across levels of proficiency in the two tests is broadly similar, males appear to be more over-represented at the lowest levels of proficiency in PISA relative to PIAAC and, by contrast, to be under-represented at the highest levels of proficiency in PISA relative to PIAAC.

FIGURE 2

Sampling and gender gaps in PISA and PIAAC

Figure 3 illustrates the gender gap in the PISA and PIAAC samples when comparing 15 and 16 year-olds in PISA and 16 and 17 year-olds in PIAAC while accounting for differences across the two studies in sampling and participation. Full results are available in Table S1 in the Supplementary Online Annex. The gender gap in favour of females in the full PISA sample of 15 and 16 year olds is d = .347 while in PIAAC it is less than a third: d = .104. Adjustments to account for out-of-school populations reduce further the gender gap in PIAAC (d = .085 when individuals who indicate they dropped out of education are removed from the PIAAC sample) and PISA (d = .336 when individuals in the bottom 5% of the PISA gender specific and country specific distribution of literacy achievement are excluded). The adjustment for response rate differences across the two studies reduces the gender gap in PISA (d =.298 when the bottom 25% of the gender specific and country specific distribution of PISA literacy achievement is removed from the sample). These results are in line with prior research suggesting that males underperform particularly at the bottom end of the literacy distribution (Hedges & Nowell, 1995; Machin & Pekkarinen, 2008) but also that neither response rates nor differences in target population can explain the large difference between gender gaps in literacy identified in PISA and PIAAC.

FIGURE 3

Table 2 reports t tests for the difference in estimated gender gaps across relevant specifications: the main PISA sample against a hypothetical PIAAC sample that excludes out-of-school individuals; the PISA sample adjusted for individuals who would be most likely to leave school if anyone left school between the age of 16 and 17; the PISA sample adjusted for response rate differences across PISA and the main PIAAC sample, and the PIAAC sample net of the out-

of-school population. In all specifications the estimated gender gap in PISA is larger than the estimated gap in PIAAC, the difference is statistically significant and quantitatively meaningful (i.e. the difference is between 19% and 26% of a standard deviation depending on the specification). These results do not provide support for the second hypothesis: differences across gender gaps in the two tests cannot be explained by selection effects.

TABLE 2

Scoring method and gender differences in PISA and PIAAC

In Table S2 in the Supplementary Online Annex we report detailed results on the percentage of correct answers in PISA and PIAAC by gender, odds ratios and the average marginal effects for females relative to males when we apply the PISA coding scheme (treating non response as a wrong answer) and when we apply the PIAAC coding scheme (treating non response as missing information)[2]. Main findings are illustrated in Figure 4 and formal test of differences across estimates are presented in Table 2.

Figure 4 indicates that in PISA the gender gap is wider when unattempted answers are treated as wrong than when they are treated as missing (because males tend to leave more questions unanswered even though they may be more likely to engage in guessing behaviour for particular set of items). Males' greater propensity to leave questions unattempted could be a reflection of lower overall performance but also of lower test engagement. The effect of coding

[2] The computer-based administration of PIAAC means that the actual scoring scheme made use of response time to identify engaged respondents, rapid guesses and identify if missing information should be considered as not attempted or wrong. In PIAAC a missing response is treated as wrong when test-takers spend more than 5 seconds with the item stimulus and perform a minimum number of actions. In order to develop meaningful comparisons with PISA we discard this information and estimate all models twice, once treating all missing information as wrong and once as not attempted.

scheme on estimates of the gender gap is more pronounced in PISA than in PIAAC. Table 2 indicates that differences in the gender gap estimated across the two studies after considering differences in coding schemes remains statistically significant and corresponds to around 6-7% of a standard deviation. These results do not support the third hypothesis: gender gaps estimated in PISA and PIAAC remain different even when the same scoring method is used to treat item non response in the two assessments.

Test length and gender differences in PISA and PIAAC

In Table S3 in the Supplementary Online Annex we report the size of the gender gap for test material that occurred at different elapsed time during the test session. The PISA test is composed of four clusters designed to take 30 minutes each to complete (total testing time is fixed at 2 hours) while the PIAAC test is composed of two modules designed to take around 20-25 minutes each to complete (total testing time is not fixed but is around 45 minutes to 1 hour for most participants.

Figure 4 indicates that the gender gap is smaller at the start of the test and grows larger the longer the test session: in PISA, when unanswered items are considered as wrong, the gender gap is very large, not just when considering test questions in the second part of the test (third cluster position), but also at the start. Males are considerably more likely than females to leave questions unanswered not just because of fatigue effects after one hour of testing, but also at the very start of the test session. In PIAAC, estimates are not precisely estimated because the sample size is small for specific age groups, a problem that is accentuated by the fact that isolating effects by adaptively administered modules reduces the sample even further. However, estimated effects work in the same direction: males' relative performance is better at the start of the test than at the

end, so much so that males may even perform better than females at the start of the test but just as well as females in the second part of the test.

Table 2 presents results of formal tests of differences of estimated gender gaps across the two assessments. Results indicate that the difference between the gender gap estimated only considering the first cluster in PISA and the first module in PIAAC is large and statistically significant. Estimates based on the second module in PIAAC (both compared to the first cluster in PISA and to the second cluster in PISA) are quantitatively large and in the hypothesised direction but are not statistically significant at conventional levels. These findings do not support hypothesis number four: estimated gender gaps in the two tests remain different even when the two tests are truncated to comprise the same length.

FIGURE 4

Mode of delivery and gender differences in PISA and PIAAC

Figure 5 illustrates the gender gap in the main PISA paper-based assessment instruments, the PISA computer-based assessment of literacy as well as the gender gap in the subset of PIAAC items that were developed for paper-based administration and those that were developed for computer-based administration. Estimates for the computer-based items are very imprecisely estimated due to the small number of observations (only a small subset of items were developed for computer delivery and the sample size for specific age groups in PIAAC is small). Table S4 in the Supplementary Online Annex indicates that irrespective of scoring method, the gender gap in the PISA computer-based literacy assessment is large and in favour of females. In PIAAC, estimates for the items that were originally developed for paper-based administration are in line with the main results (the estimated gender gap is in favour of males). Although it is not possible

to reject the hypothesis of no gender differences in PIAAC, results presented in Table 2 indicates that it is possible to establish that estimated gender gaps in PISA and PIAAC are different, whether comparing the paper-based estimates in PISA with the PIAAC instruments originally developed for paper-based administration or whether comparing the computer-based PISA assessment with the PIAAC test. These results do not support the fifth hypothesis that differences in gender gaps estimated in PISA and PIAAC reflected differences in mode of administration.

Text type, cognitive processes and gender differences in PISA and PIAAC

Figure 5 illustrate the gender gap in PISA and PIAAC on "access and retrieve" type of items and items involving stimulus material containing "mixed" texts. Results indicate that females perform considerably better than males in access and retrieve items in PISA but not in PIAAC. Table 2 indicates that the difference in the estimated gender gap in these items is statistically significant and quantitatively meaningful. Differences across the two studies are also apparent when comparing items involving mixed text stimuli: the gender gap is in favour of girls in PISA while this is not the case in PIAAC.

FIGURE 5

Although limitations with the item pool prevent the development of extensive analyses on the influence of text type and item format on differences in estimated gender gaps in the two assessments, these preliminary findings indicated that the sixth hypothesis was not supported in this initial analysis.

Longitudinal evidence and the role of attitudes towards school

Table 3 illustrates the gender gap in Canada in PISA 2000, 2009, 2012, in the subset of PISA participants in 2000 who were followed in the Youth in Transition Study (YITS), and in YITS participants in 2009. The comparison between PISA 2000, 2009 and 2012 reveals that the gender gap in reading in PISA has remained stable over the period in Canada and therefore it is possible to compare the gender gap across different cohorts. The gender gap in the subset of PISA 2000 participants who were followed in YITS was the same, indicating that the YITS sample is representative of the Canadian PISA participating population.

TABLE 3

Table 3 indicates that among YITS participants in 2009 females outperformed males by a considerably smaller margin than in 2000: the average gap was 0.18 while it was 0.32 in 2000. Interestingly, the narrower gap in YITS 2009 compared to YITS 2000 was due primarily to the fact that males achieved a considerably higher score in 2009 compared to 2000: the difference in the performance of males was 0.63 while it was 0.50 among females. The literature considered these differences to reflect a differential growth in achievement among males, ignoring differences in testing conditions between the original PISA 2000 administration and the YITS 2009 study which can, effectively, be considered as a PISA test administered under PIAAC conditions. These results are in line with findings reported in previous sections on the larger literacy gender gaps observed in PISA than in PIAAC.

A final indirect test of the role of motivational factors in shaping gender gaps in literacy among youngsters is presented in Table 4, where we report the gender gap in the percentage of males and females in PISA 2012 who reported that they strongly agree that school is a waste of time as well as the gender gap in literacy among individuals with similar attitudes towards school.

Results indicate that males are more likely than females to believe is that school is a waste of time (only 85% of males report disagreeing or strongly disagreeing that school is a waste of time compared to 93% of females). Among those who have negative attitudes towards school the gender gap is $d = 0.36$. By contrast, the gender gap is smaller among students who have positive attitudes towards school ($d = 0.27$). The difference in the gender gap across groups of students who report that school is a waste of time and those who do not is $d = 0.9$ and is statistically significant only at the 10% level. The closing of the gap appears to be due to a particularly steep decline in performance among male students who perceive school to be a waste of time.

TABLE 4

Discussion, Conclusion and Implications

The aim of this work was to identify if literacy gender gaps evaluated in different large-scale assessments are aligned or not and, should differences be identified, to assess the influence of test content and administration conditions in shaping how large gender gaps are. We began by observing that gender gaps in the literacy domain in PISA and PIAAC are remarkably different: the gender gap in favour of females in PISA is large while no gap can be observed on average among young people participating in PIAAC. In fact, in some countries, males may outperform young females. Such difference appears to be driven primarily by the fact that males' achievement in the two tests differs, while females' achievement is relatively stable across the two assessments.

We developed analyses to test if such differences can be ascribed to differences across the two studies in sample selectivity (which may reflect gender differences in motivation or ability to sit the test) or to differences across the two studies in scoring method, test length, mode of administration, type of texts used in the assessment or cognitive processes involved in solving

assessment tasks (factors that may have a bearing on or reflect differences in the motivation to engage with the test of males and females, conditional on participation).

We failed to find support for the five hypotheses identifying causes for observed differences in estimated gender gaps across the two tests. Differences in sample selectivity, test length, scoring method, mode of delivery, text type and cognitive process involved in solving assessment tasks do not appear to explain differences in gender gaps between PISA and PIAAC. Because in our study we rely on ex-post adjustments, we can only speculate on what differences across the two studies that we could not account for could be responsible for remaining differences in estimated gender gaps. Chief among these are the fact that PISA is administered in a school setting with one-to-many proctoring, while PIAAC is administered in people's homes under one-to-one proctoring or that the items employed in PISA and PIAAC may differ along dimensions not considered in this work and that may render the PIAAC test easier for males.

Our results are a reminder that measurement matters: it is key for researchers and policy makers alike to consider what is being measured and how. While the role of test content and scoring methods has been identified before, our results suggest (but do not prove) that how a test is administered could have an important bearing on results and, certainly, on between-group differences. In particular, we believe that the type of proctoring and the setting in which people complete an achievement test might influence test-taking motivation, especially among males, and that this, in turn, could influence observed gender gaps in achievement.

Virtually all large-scale assessment studies of gender gaps in achievement are based on tests that are administered in school settings. This makes sense from a practical and theoretical standpoint: applying a two stage (or even a three stage) sampling design whereby schools are sampled first and (classes) students are sampled within participating schools reduces costs, ensures high levels of participation and facilitates the standardisation of testing settings.

Moreover, since schools are the institutions typically mandated to develop students' competencies and to certify these, administering tests in schools corresponds to standard educational scenarios.

Males and females have been found to hold different attitudes towards schools and, in particular, during the teenage years males are drawn to display confrontational attitudes towards school as a way to establish their position within peer groups (Eccles, Wigfield, Harold, & Blumenfeld, 1993; Jacobs, Lanza, Osgood, Eccles, & Wigfield, 2002; Kaplan, Gheen, & Midgley, 2002; Martin, 2007; Skinner, Furrer, Marchand, & Kindermann, 2008). It is therefore possible that in school settings male students may be motivated to display lack of engagement with the test and exert lower effort than female students. These differences may mean that males and females may not be equally motivated to do well in school-administered tests. As a result, estimates of gender gaps may not reveal what is intended. PISA states that its aim is to measure what 15-year-old students "can do with what they know" and students with low levels of achievement on the test are considered as not having acquired the skills and competences that are necessary to thrive in society (OECD, 2019a). Our work raises the possibility that some of these students may have acquired skills and competences but may not be showing this over the course of the PISA assessment and, in fact, in many school assessments. It is well-known that test taking motivation and test taking engagement are related to features of the test in general (Penk, Pöhlmann, & Roppelt, 2014) and proctoring specifically (Lau, et al., 2009). Our results apply to low-stakes assessments, but many school-level in-course assessments are relatively low-stakes individually, but, taken together, can be consequential in shaping students' educational progressions.

Our study suffers from a number of limitations that should be remedied in future research. First, the PIAAC sample was not designed to enable analyses of specific age groups but rather, to provide population level estimates. Therefore, studies that rely on specific age groups in PIAAC employ small samples and therefore estimates are imprecise. While in most analyses we

were able to reject the null hypothesis that the gender gap in literacy is the same across PISA and PIAAC, a larger sample would allow for better and more robust comparisons and would allow to account for multiple item characteristics at the same time. Second, we only focus on literacy and results may not apply across other subjects/assessment domains. The third and most important limitation of our study is that it relies on a comparison between existing large-scale assessments data rather than an experimental design. Ideally, future studies could attempt to identify the influence of different measurement approaches and administration conditions by randomly allocating to participants in the PISA and PIAAC tests under either PISA or PIAAC conditions. Such studies would allow to identify if the PIAAC item pool differs in ways that we could not account for and that render it more approachable to males or if, by contrast, administration conditions may shape the size of gender gaps in achievement tests.

Our work contributes new evidence to the growing body of work suggesting that the relative performance of males and females in achievement tests is highly dependent on the characteristics of these tests. Prior work investigated the extent to which gender differences depend on the consequences tests have for test takers (Braun et al., 2011; Coffman & Klinowski, 2018; Gneezy, et al., 2019); test length (Balart & Oosterveen, 2019; Borgonovi & Biecek, 2016); types of tasks included in the assessment (Taube & Munck, 1996); mode of delivery of the assessment tasks (Borgonovi, 2016); and response format (Baldiga, 2014; Reardon, Kalogirdes, Fahle, Podolsky, & Zarate, 2018). Our work suggests that even when tests are very similar, administration conditions, or subtle differences in the types of tasks used in assessments, could influence the relative achievement of males and females by shaping their engagement and motivation.

Before embarking on major policy reforms designed to ensure that boys acquire literacy skills based on their poor showing in the context of large-scale assessments as well as school tests,

it would be important to evaluate if and how assessments reflect all of what boys "know and can do", if assessments are comprehensive enough to capture dimensions of literacy that boys may be more proficient in and, crucially, if the assessments provide incentives for boys to show test administrators what they "know and can do".

Changes in instructional materials, pedagogical approaches, school-level policies and institutional frameworks to support the learning of boys may be unnecessary and counterproductive if boys are learning but are not willing or are not able to display what they learnt in the context of assessments. If gender differences are due to differences in effort rather than differences in competencies, policies that promote test-taking motivation either by redesigning assessments or changing the incentives and information boys have on the importance of investing effort in the assessments would be more appropriate.

REFERENCES

Azmat, G., Bagues, M., & Cabrales, A. (2016). *What You Don't Know... Can't Hurt You? A Field Experiment on Relative Performance Feedback in Higher Education*, IZA Discussion Papers, No. 9853, Institute for the Study of Labor (IZA), Bonn.

Balart, P., & Oosterveen, M. (2019). Females show more sustained performance during test-taking than males. *Nature Communications, 10*(1). https://doi.org/10.1038/s41467-019-11691-y

Baldiga, K. (2014). Gender differences in willingness to guess. *Management Science, 60*(2), 434–448. https://doi.org/10.1287/mnsc.2013.1776

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? a high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342–363. https://doi.org/10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462. https://doi.org/10.1007/BF03173192

Baumeister, R. E., Heatherton, T. E., & Tice, D. M. (1994). *Losing control: How and why people fail at self-regulation.* San Diego: Academic Press.

Borghans, L., & Schils, T. (2012). *Decomposing achievement test scores into measures of cognitive and noncognitive skills*. Available online at: http://www.sole-jole.org/13260.pdf

Borgonovi, F. (2016). Video gaming and gender differences in digital and printed reading performance among 15-year-olds students in 26 countries. *Journal of Adolescence, 48,* 45-61.

Borgonovi, F., & Biecek, P. (2016). An international comparison of students' ability to endure fatigue and maintain motivation during a low-stakes test. *Learning and Individual*

Differences, *49*, 128–137. https://doi.org/10.1016/j.lindif.2016.06.001

Boys' Reading Commission (2012) The report of the All-Party Parliamentary Literacy Group Commission. Report compiled by the National Literacy Trust. Retrieved November 12, 2019, from https://literacytrust.org.uk/policy-and-campaigns/all-party-parliamentary-group-literacy/boys-reading-commission/

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teachers College Record*, *113*(11), 2309–2344.

Breda, T., Jouini, E., & Napp, C. (2018). Societal inequalities amplify gender gaps in math. *Science*, *359*(6381), 1219–1220. https://doi.org/10.1126/science.aar2307

Buchmann, C., DiPrete, T. A., & McDaniel, A. (2008). Gender Inequalities in Education. *Annual Review of Sociology*, *34*(1), 319–337. https://doi.org/10.1146/annurev.soc.34.040507.134719

Chan, D., Schmitt, N., DeShon, R. P., Clause, C. S., & Delbridge, K. (1997). Reactions to cognitive ability tests: The relationships between race, test performance, face validity perceptions, and test-taking motivation. *Journal of Applied Psychology*, *82*(2), 300–310. https://doi.org/10.1037/0021-9010.82.2.300

Coffman, K. B., & Klinowski, D. (2018). *The Impact of Penalties for Wrong Answers on the Gender Gap in Test Scores*. Harvard Business School Working Paper 19-017. Boston: Harvard Business School.

Cole, N. S. (1997). The ETS Gender Study: How Females and Males Perform in Educational Settings. Princeton: Education Testing Service. Retrieved from https://files.eric.ed.gov/fulltext/ED424337.pdf

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–

624. https://doi.org/10.1016/j.cedpsych.2007.10.002

Cunningham, A. E., & Stanovich, K. E. (1998). What reading does to the mind. *American Educator*, *22*, 1–8.

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

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*, 198–208. https://doi.org/10.1037/0022-0663.98.1.198

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. https://doi.org/10.1146/annurev.psych.53.100901.135153

Eccles, J. S., Wigfield, A., Harold, R. D., & Blumenfeld, P. (1993). Age and gender differences in children's self- and task perceptions during elementary School. *Child Development*, *64*(3), 830. https://doi.org/10.2307/1131221

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy and Practice*, *17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127. https://doi.org/10.1037/a0018053

Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age.* New York: Springer.

Gal, I., & Tout, D. (2014). Comparison of PIAAC and PISA frameworks for numeracy and mathematical literacy. *OECD Education Working Papers 102.* Paris: OECD Publishing.

Gneezy, J. List, J. A., Livingston J. A., Qin, X., Sadoff, S., & Xu Y. (2019). Measuring success in education: The role of effort on the test itself. *American Econmic Review: Insights, 1*(3), 291-308.

Grotlüschen, A., Mallows, D., Reder, S., & Sabatini, J. (2016). Adults with low proficiency in literacy or numeracy. *OECD Education Working Papers*, No. 131, OECD Publishing, Paris, https://doi.org/10.1787/5jm0v44bnmnx-en.

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science, 320*(5880).

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*(5220), 41–45. https://doi.org/10.1126/science.7604277

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems. *Journal of Economic Perspectives, 30*(3), 85-108.

Jacobs, J. E., Lanza, S., Osgood, D. W., Eccles, J. S., & Wigfield, A. (2002). Changes in children's self-competence and values: Gender and domain differences across grades one through twelve. *Child Development, 73*(2), 509–527. https://doi.org/10.1111/1467-8624.00421

Kaplan, A., Gheen, M., & Midgley, C. (2002). Classroom goal structure and student disruptive behaviour. *British Journal of Educational Psychology*, *72*(2), 191–211. https://doi.org/10.1348/000709902158847

Kunnskapsdepartementet. (2019). *Nye sjanser-bedre laering*. Retrieved from www.regjeringen.no

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58*(3), 196–217. https://doi.org/10.1353/jge.0.0045

Legewie, J., & DiPrete, T. A. (2012). School context and the gender gap in educational achievement. *American Sociological Review*, *77*(3), 463–485. https://doi.org/10.1177/0003122412440802

Lietz, P. (2006). A meta-analusis of gender differences in reading achiebement at the secondary level. *Studies in Educational Evaluation*, *32*(4), 317–344. https://doi.org/10.1016/j.stueduc.2006.10.002

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Lynn, R., & Mikk, J. (2009). Sex differences in reading achievement. Trames, 13(1), 3-13. doi:10.3176/tr.2009.1.01

Machin, S., & Pekkarinen, T. (2008). Assessment: Global sex differences in test score variability. *Science*, 322, 1331–1332. https://doi.org/10.1126/science.1162573

Martin, A. J. (2007). Examining a multidimensional model of student motivation and engagement using a construct validation approach. *British Journal of Educational Psychology*, *77*(2), 413–440. https://doi.org/10.1348/000709906X118036

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology*, *101*(3), 689–704. https://doi.org/10.1037/a0014240

Mood, C. (2009). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, 26(1), 67–82.

Nollenberger, N., Rodríguez-Planas, N., & Sevilla, A. (2016). The math gender gap: The role of culture. *American Economic Review*, *106*(5), 257–261. https://doi.org/10.1257/aer.p20161121

OECD. (2001a). *Literacy Skills for the World of Tomorrow: Further Results from PISA 2000 - Publications 2000*. Paris: OECD Publishing.

OECD (2010a). *PISA 2009 Assessment Framework.* Paris: OECD Publishing.

OECD (2010b). *PISA 2009 Results: What students know and can do. (Volume I). Paris: OECD Publishing*. http://dx.doi.org/10.1787/9789264091450-en

OECD. (2012). *Learning beyond Fifteen Ten Years after PISA.* Paris: OECD Publishing. https://doi.org/10.1787/9789264172104-en

OECD. (2013a). *Technical Report of the Survey of Adult Skills (PIAAC). First Edition.* Available online at https://www.oecd.org/skills/piaac/_Technical%20Report_17OCT13.pdf

OECD. (2013b). *PISA 2012 Assessment and Analytical Framework PISA 2012 Assessment and Analytical Framework.* Paris: OECD Publishing. https://doi.org/10.1787/9789264190511-en

OECD. (2013c). *PISA 2012 Results: What Students Know and Can Do.* Paris: OECD Publishing.

OECD. (2013d). *The Survey of Adult Skills: Reader's companion.* OECD Publishing.Paris: OECD Publishing.

OECD. (2015). *The ABC of Gender Equality in Education.* Paris: OECD Publishing.

OECD. (2016) *The Survey of Adult Skills - Reader's Companion, Second Edition.* Paris: OECD Publishing.

OECD. (2017). *The Pursuit of Gender Equality*. Paris: OECD Publishing.

OECD. (2019a). *PISA 2018 Results (Volume I). What Students Know and Can Do.* Paris: OECD Publishing.

OECD. (2019b). *Technical Report of the Survey of Adult Skills (PIAAC). Third Edition.* Available online at https://www.oecd.org/skills/piaac/publications/PIAAC_Technical_Report_2019.pdf

OECD (2020). PISA 2018 Technical Report. Paris: OECD Publishing.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students'

performance in low-stakes assessments: an investigation of school-track-specific differences. *Large-Scale Assessments in Education, 2*(1). https://doi.org/10.1186/s40536-014-0005-4

PIAAC Literacy Expert Group. (2009), *PIAAC Literacy: A Conceptual Framework*. OECD Education Working Papers, No. 34, Paris: OECD Publishing.

Pokropek, A., & Borgonovi, F. (2019). Linking via pseudo-equivalent group design: Methodological considerations and an application to the PISA and PIAAC assessments. *Journal of Educational Measurement*. Doi: 10.1111/jedm.12261

Reardon, S. F., Kalogirdes, D., Fahle, E. M., Podolsky, A., & Zarate, R. C. (2018). The relationship between test item format and gender achievement gaps on math and ELA tests in fourth and eight grades. Educational Researcher, 47(5), 284-294.

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

Rubin, D. B., Wiley, J., York, N., Brisbane, C., & Singapore, T. (1987). *Multiple Imputation for Nonresponse in Surveys*.

Schenker, N., & Gentleman, J. F. (2001). On judging the significance of differences by examining the overlap between confidence intervals. *American Statistician, 55*(3), 182–186.

Schunk, D. H., Pintrich, P. R., Meece, J. L., & Pintrich, P. R. (2008). *Motivation in education : theory, research, and applications*. Pearson/Merrill Prentice Hall.

Skinner, E., Furrer, C., Marchand, G., & Kindermann, T. (2008). Engagement and disaffection in the classroom: Part of a larger motivational dynamic? *Journal of Educational Psychology*, *100*(4), 765–781. https://doi.org/10.1037/a0012840

Smith, M. C., Mikulecky, L., Kibby, M. W., Dreher, M. J., & Dole, J. A. (2000). What will be the demands of literacy in the workplace in the next millennium? *Reading Research Quarterly*, *35*(3), 378–383. https://doi.org/10.1598/rrq.35.3.3

Smith, M. W., & Wilhelm, J. D. (2009). Boys and literacy. Complexity and Multiplicity. In L. Christenbury, R. Bomer, & P. Smagorinsky (Eds.), *Handbook of Adolescent Literacy Research* (pp. 360-372.).

Solheim, O. J., & Lundetræ, 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 & Practice, 25*(1), 107-126.

Støle, H., Mangen, A., & Schwippert, K. (2020). Assessing children's reading comprehension on paper and screen: A mode-effect study. *Computers & Education, 151,* 103861.

Stoet, G., & Geary, D. C. (2018). The Gender-Equality Paradox in Science, Technology, Engineering, and Mathematics Education. *Psychological Science*, *29*(4), 581–593. https://doi.org/10.1177/0956797617741719

Taube, K., & Munck, I. (1996). Gender differences at item level. In H. Wagemaker, K. Taube, I. Munck, G. Kontogiannopoulou-Polydorides, & M. Martin (Eds.), *Are girls better readers? Gender differences in 32 countries* (pp. 53–98). Amsterdam: International Association for the Evalutation of Educational Achievement.

UNESCO. (2019). Global Education Monitoring Report – Gender Report: Building bridges for gender equality. In *Accountability in Education: Meeting Our Commitments*. Retrieved from https://en.unesco.org/gem-report/sites/gemreport/files/References_GenderReport2019.pdf

Van Hek, M., Buchmann, C., & Kraaykamp, G. (2019). Educational systems and gender differences in reading: A comparative multilevel analysis. *European Sociological Review, 35*(2), 169–186. https://doi.org/10.1093/esr/jcy054

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17.

https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*(1), 27–41. https://doi.org/10.1080/10627191003673216

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test, motivation, and mntally taxing items. *Applied Measurement in Education, 8*(4), 341–351. https://doi.org/10.1207/s15324818ame0804_4

Wolfe, R., & Hanley, J. (2002). If we're so different, why do we keep overlapping? When 1 plus 1 doesn't make 2. *CMAJ : Canadian Medical Association Journal*. *Journal de l'Association Medicale Canadienne, 166*(1), 65–66.

Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital, 13*(4), 519-552.

Zhou, M. (2014). Gender difference in web search perceptions and behavior: does it vary by task performance? *Computers and Education, 78,* 174-184.

Table 1

Descriptive statistics

	PISA		PIAAC	
Statistics	Boys	Girls	Boys	Girls
N	113590	112343	3715	3780
Mean performance	488	521	263	268
Standard Deviation	99	89	46	41
25th percentile	422	463	234	241
75th percentile	559	583	296	296

Source: PISA 2000 and PIAAC databases.

Table 2

Testing for differences in gender gaps across specifications

	PISA		PIAAC		Testing for difference between PISA and PIAAC		
Type of analysis	Standardised gender gap (Girls-Boys)	SE	Standardised gender gap (Girls-Boys)	SE	Difference (PISA-PIAAC)	SE	t test
PISA main vs. PIAAC adjusted for out of school	0.347	(0.012)	0.085	(0.002)	0.262	(0.012)	22.488
PISA adjusted for out of school vs. PIAAC main	0.336	(0.012)	0.085	(0.002)	0.251	(0.012)	21.201
PISA adjusted for response rate differences vs. PIAAC main	0.298	(0.010)	0.104	(0.002)	0.193	(0.010)	19.390
PISA adjusted for response rate differences vs. PIAAC adjusted for out of school	0.298	(0.010)	0.085	(0.002)	0.213	(0.010)	21.528
	PISA		PIAAC		Testing for difference between PISA and PIAAC		
	AME (Girls-Boys)	SE	AME (Girls-Boys)	SE	Difference (PISA-PIAAC)	SE	t test
Scoring effects							
All test: Missing treated as wrong	0.056	0.003	-0.016	0.027	0.072	(0.027)	2.639
All test:Missing treated as missing	0.040	(0.003)	-0.023	0.027	0.063	(0.027)	2.338
Test length effects							
PISA 1 vs. PIAAC 1: missing as wrong	0.044	(0.004)	-0.037	(0.039)	0.081	(0.040)	2.046
PISA 1 vs. PIAAC 1: missing as missing	0.033	(0.004)	-0.048	(0.039)	0.081	(0.039)	2.079
PISA 1 vs. PIAAC 2: missing as wrong	0.044	(0.004)	0.002	(0.036)	0.042	(0.036)	1.177
PISA 1 vs. PIAAC 2: missing as missing	0.033	(0.004)	0.002	(0.036)	0.031	(0.037)	0.857
Mode of delivery effects							
PISA digital PIAAC all: missing as wrong	0.052	(0.005)	-0.016	0.027	0.067	(0.027)	2.451
PISA digital PIAAC all: missing as missing	0.043	(0.004)	-0.023	(0.027)	0.066	(0.027)	2.436
PISA paper PIAAC paper: missing as wrong	0.056	0.003	-0.040	0.032	0.096	(0.032)	2.953

PISA paper PIAAC paper: missing as missing	0.040	(0.003)	-0.029	0.033	0.069	(0.033)	2.103
Item characteristics effects							
Access and retrieve: missing as wrong	0.049	0.003	-0.030	0.033	0.080	(0.033)	2.394
Access and retrieve: missing as missing	0.034	0.003	-0.040	0.033	0.073	(0.033)	2.235
Mixed texts: missing as wrong	0.032	0.005	-0.050	0.031	0.082	(0.032)	2.587
Mixed texts: missing as missing	0.024	0.005	-0.061	0.032	0.085	(0.032)	2.615

Note: Standardised gender gaps represent Cohen's d (a positive sign indicates a gap in favour of females and a negative sign indicates a gap in favour of males). PIAAC out of school population adjustment: removal of individuals not in education at the time of the PIAAC test. PISA out of school population adjustment: removal of the bottom 5% of the country specific and gender specific distribution of literacy achievement. PISA response rate adjustment: removal of the bottom 25% of the country specific and gender specific distribution of literacy achievement. Results robust to alternative thresholds and specifications.
Average Marginal Effects (AME) obtained from a logistic regression model where 1=correct and 0=incorrect answer to a particular test question. In the 'missing as missing specification' unattempted answers are considered as missing while in the 'missing as wrong specification' unattempted answers are considered as wrong. All models include country fixed effects. PISA 1 and PIAAC 1 refer to test material at the start of the test, PISA 2 and PIAAC 2 refer to test material in the second part of the test. Digital refers to computer-delivered test. Paper refers to paper-and-pencil material. PIAAC estimates are based on results that account for item difficulty to adjust for the adaptive nature of the test. Results are robust to the inclusion of controls for item difficulty.

Table 3

The literacy gender gap in Canada

Study population	Boys' mean literacy achievement	Girls' mean literacy achievement	Gender difference in literacy achievement (Girls-Boys)	Cohen's d
Canada PISA 2000	519	551	32	0.32
Canada PISA 2009	507	542	34	0.34
YITS participants 2000	526	558	32	0.32
YITS participants 2009	590	608	18	0.18
Canada PISA 2012	506	541	35	0.35

Table 4

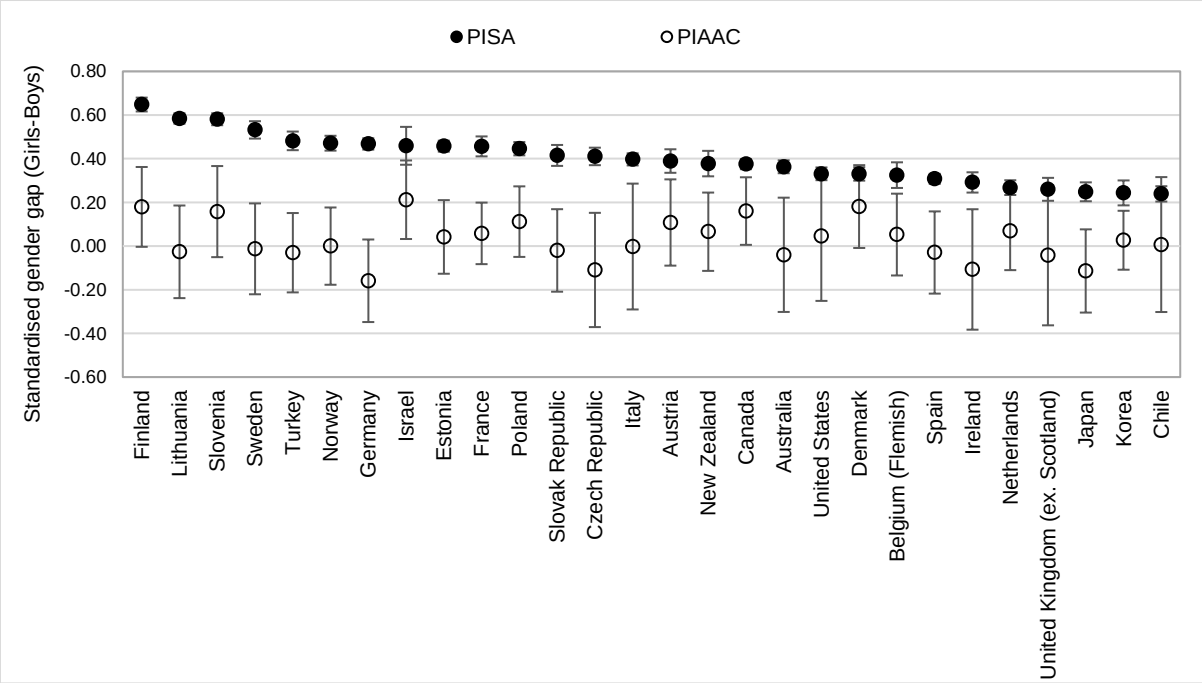
Literacy achievement as a function of attitudes towards school, by gender

	Males		Females		Gender gap		
	b	SE	b	SE	b	SE	*Standardised gap d (Females-Males)*
Share of students who disagree or strongly disagree that school a waste of time	0.851	0.003	0.930	0.003	0.079	0.004	
Literacy achievement if the student considers school NOT to be a waste of time	500	1.587	527	1.460	27	1.535	0.272
Literacy achievement if the student considers school to be a waste of time	442	2.827	478	3.493	36	4.256	0.359
Difference	57	2.679	49	3.430	9	4.825	0.087

Source: PISA 2012 database.

Figure 1

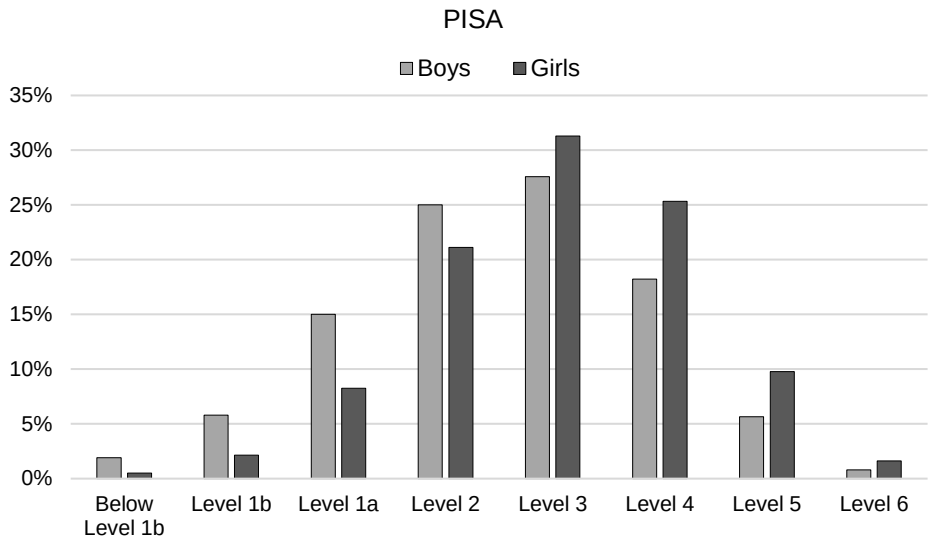
The gender gap in reading in PISA and in literacy in PIAAC

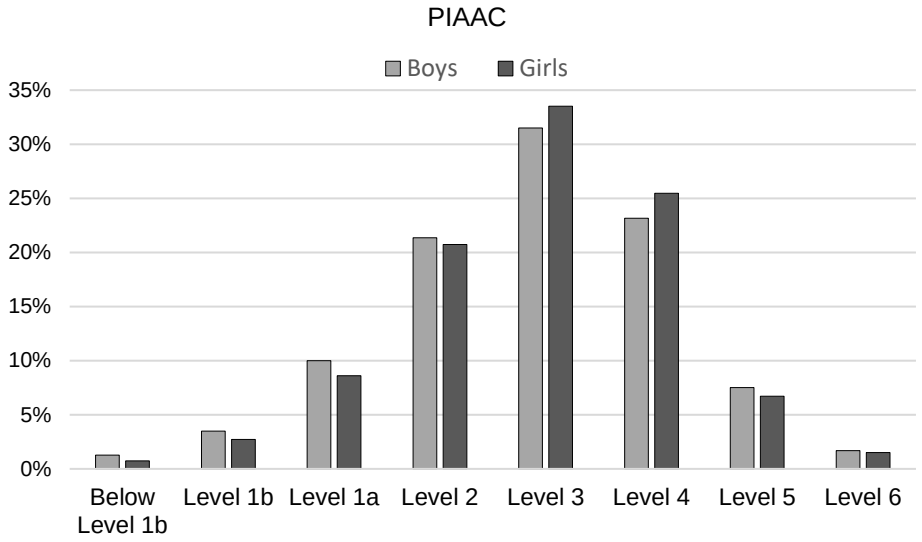


Note: The PISA main sample refers to 15 and 16-year-olds. The PIAAC main sample refers to 16-20 year-olds (to increase the country specific sample size for PIAAC). The dark dot represents the estimated gender gap (F-M) expressed as a Cohen's d in PISA. The light dot represents the estimated gender gap (F-M) expressed as a Cohen's d in PIAAC. Confidence Intervals at the 95% level for each estimate are presented.
Source: OECD, PISA 2012 and PIAAC Databases.

Figure 2

Percentage of boys and girls at each proficiency level in literacy in PISA and PIAAC

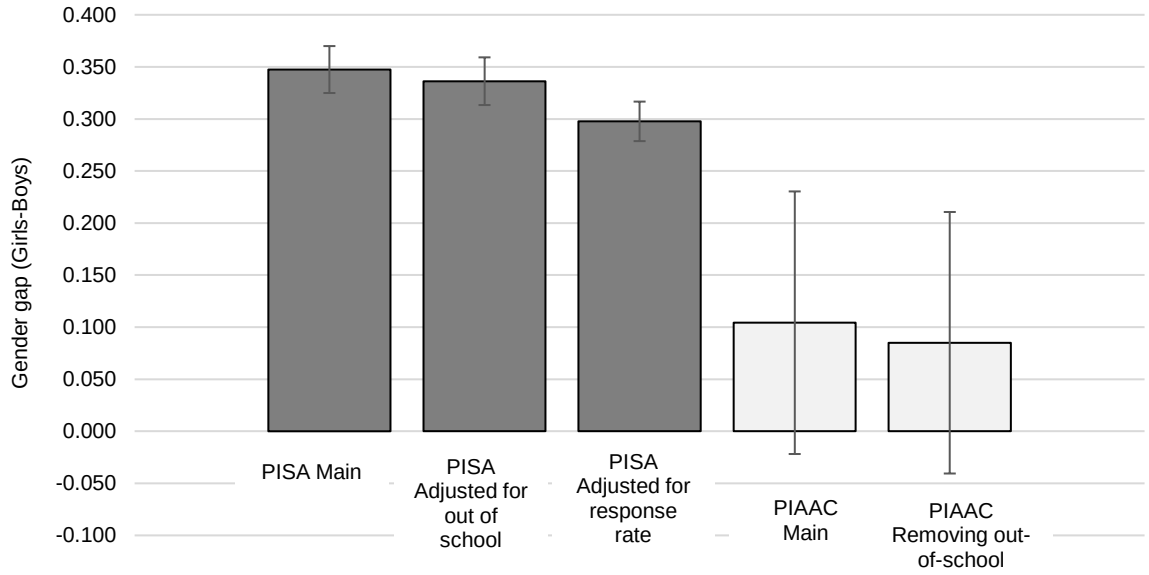




Notes: PISA: 15-16 year-olds. PIAAC: 16-17 year-olds. Results for PIAAC reflect PISA proficiency scores and are based on concordance scores estimated by Pokropek & Borgonovi (2019) and reported in Annex Table A3.

Figure 3

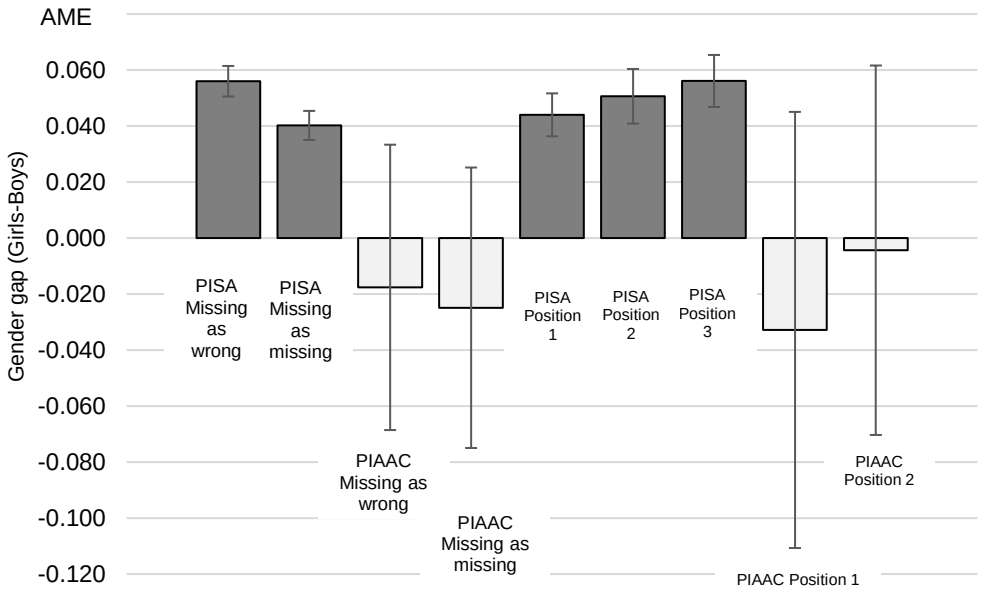
The role of differences in target population and response rates in explaining differences in gender gaps in literacy in PISA and PIAAC



Notes: PISA: 15-16-year-olds. PIAAC: 16-17 year-olds. The adjustment for the out-of-school population in PISA involves removing from the sample students in the gender specific bottom 5 percent of literacy performance. The adjustment for the response rate in PISA involves removing from the sample students in the gender specific bottom 25 percent of literacy performance. The adjustment for individuals not in education in PIAAC involves removing from the sample individuals who reported not being in education at the time of testing (6%).

Figure 4

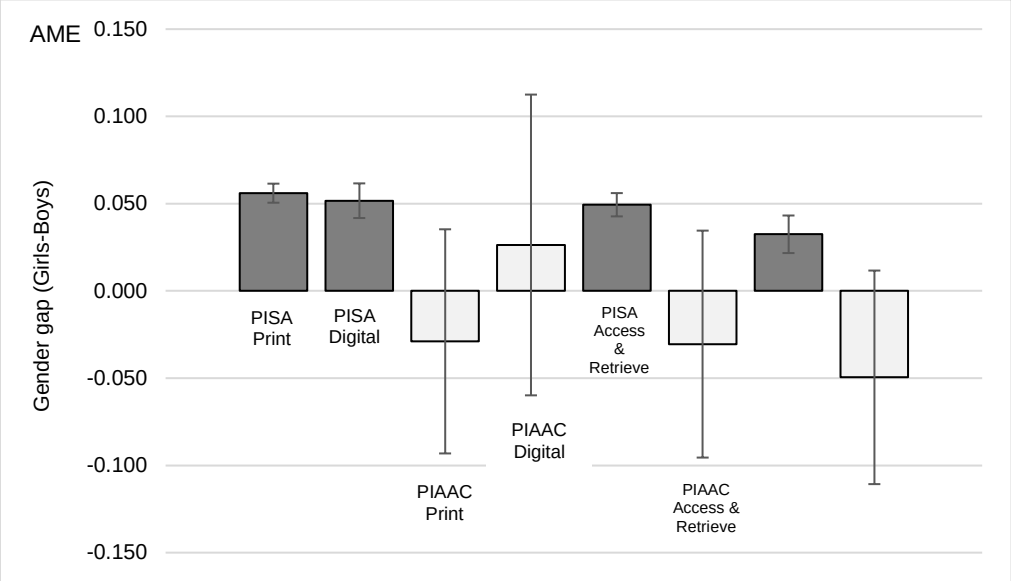
The role of scoring and test length in explaining differences in gender differences in literacy in PISA and PIAAC



Note: Average Marginal Effects (AME). In the 'missing as missing specification' unattempted answers are considered as missing while in the 'missing as wrong specification' unattempted answers are considered as wrong. PISA 1 and PIAAC 1 refer to test material at the start of the test, PISA 2 and PIAAC 2 refer to test material in the second part of the test. PISA 3 refers to test material in the third part of the test. Results for test length are based on the missing as wrong specification. PIAAC estimates are based on results that account for item difficulty to adjust for the adaptive nature of the test. Results are robust to the inclusion of controls for item difficulty. All models include country fixed effects.

Figure 5

The role of mode of delivery and item characteristics in explaining differences in gender differences in literacy in PISA and PIAAC



Note: Average Marginal Effects (AME). All results are based on the missing as wrong specification. PISA print refers to the main paper-based assessment. PISA digital refers to the computer-based optional assessment. PIAAC print refers to computer-delivered test questions that were originally developed for paper-based administration. PIAAC digital refers to computer-delivered and developed for computer delivery. Access and retrieve type of items are items requiring the application of these cognitive processes to be solved. Mixed refers to test items containing mixed texts. Results are robust to the inclusion of controls for item difficulty. All models include country fixed effects.