PISA 2012: How do results for the paper and computer tests compare?

John Jerrim

UCL Institute of Education

February 2015

http://johnjerrim.com/papers/

Abstract

The Programme for International Assessment (PISA) is an important cross-national study of 15 year olds academic achievement. Although it has traditionally been conducted using paper-and-pencil tests, the vast majority of countries will use computer-based assessment from 2015. In this paper we consider how cross-country comparisons of children's skills differ between paper and computer versions of the PISA mathematics test. Using data from PISA 2012, where more than 200,000 children from 32 economies completed both paper and computer versions of the mathematics assessment, we find important and interesting differences between the two sets of results. This includes a substantial drop of more than 50 PISA test points (half a standard deviation) in the average performance of children from Shanghai-China. Moreover, by considering children's responses to particular test items, we show how differences are unlikely to be solely due to the interactive nature of certain computer test questions. The paper concludes with a discussion of what the findings imply for interpretation of PISA results in 2015 and beyond.

Acknowledgements: I would like to thank John Micklewirght for the helpful comments.

Key Words: PISA, computer-based assessment, Shanghai-China, educational inequality.

Contact Details: John Jerrim (J.Jerrim@ioe.ac.uk) Department of Quantitative Social Science, Institute of Education, University of London, 20 Bedford Way London, WC1H 0AL

1. Introduction

The Programme for International Student Assessment (PISA) is an important cross-national study of 15 year olds educational achievement. Conducted every three years by the Organisation for Economic Co-Operation and Development (OECD), the tri-annual update of results is now eagerly awaited by educationalists and policymakers alike. Since its inception in 2000, the main study has been conducted using a traditional paper-and-pencil test. However, this will change from 2015, when the vast majority of countries will move to computer based assessment. Although testing using modern technology is becoming increasingly common (e.g. state-wide computer assessments are set to be delivered in Ontario-Canada from 2015/16 - EQAO 2014) there remains some uncertainty regarding the impact this has upon children's test scores (Wang et al 2007). Moreover, little is currently known about how such a change is likely to influence cross-country comparisons of educational achievement, including the PISA summary statistics widely reported by the OECD. The aim of this paper is to start to fill this important gap in the literature.

By moving to computer-based tests, PISA is making an important change to assessment practise. There are a number of reasons why this may have a subsequent influence upon the results (and why the impact may vary by country or demographic group). First, different cognitive processes are needed for reading on paper and computer (Mangen, Walgermo and Bronnick 2013; Jabr 2013) which may influence how children interpret and answer the PISA test questions. Indeed, it has been suggested that even subtle differences such as screen size and resolution can dramatically change the nature of computer-based tasks (McKee and Levinson 1990). Second, schools and teachers have less experience of conducting computer-based assessments, including dealing with the technical challenges involved (e.g. software crashing). Third, computer-based tests require all children to have a basic level of computing skill (e.g. to be able to type using a key board, move the mouse). Despite the widespread use of computers in modern society, such skills may still be lacking in certain countries or amongst certain demographic groups (Platt 2014:17). Fourth, unlike paper tests, most computer assessments do not allow children to review answers to previous questions (Vispoel et al 1992), thus removing this important test-taking strategy (Mueller and Wasser 1977). Fifth, the novelty of computer assessment could improve children's engagement with the test (Johnson and Geen 2006). Conversely, there may be more distractions (particularly if teacher's struggle to enforce

examination conditions) causing frustration amongst participants and reduced effort[1]. Finally, computer based tests may use different types of questions (or involve the use of interactive tools) which may advantage or disadvantage certain demographic groups.

Previous work has recognised that many of the factors listed above are test specific (Kolen and Brown 1995), and may vary across test settings. It is therefore difficult to generalise between studies or contexts. This perhaps explains why research investigating the link between test administration mode and children's test scores has produced somewhat inconsistent results. A summary of the evidence for upper secondary school mathematics (the focus of this paper) can be found in Table 1[2]. Most of this previous research has been conducted in the United States, with some studies using a randomised design (considered the 'gold standard' in this literature), and have typically found either a null or negative effect of computer-based assessment upon average test scores. Yet the magnitude of the effect report differs widely. For instance, whereas Ito and Sykes (2004) find average test scores to be around 0.24 standard deviations higher on paper than computer tests, Wang et al (2004) report there to be essentially no difference at all.

<< Table 1 >>

Although insightful, there are two notable limitations to the current evidence base (particularly regarding the implications for PISA). First, the studies reviewed in Table 1 almost exclusively focus upon differences in *average* test scores. In contrast, there is very little evidence on distributional effects (e.g. the spread of achievement or the proportion of children with high or low test scores) or co-variation with important demographic characteristics such as gender, family background and country of birth. Yet these are all examples of widely cited PISA statistics, with further work clearly needed to assess the likely impact of computer assessment upon such results. Second, the existing literature does not investigate how differences between paper and computer test scores vary across countries. For instance, do children in every country obtain lower test scores (on average) on computer based tests? Or is the use of computers advantageous in some nations but a disadvantage in others? Similarly, do computer-based tests strengthen or weaken the performance of low socio-economic groups, and is this a common pattern found across the developed world?

[1] For instance, students in the study by Fluck, Pullen and Harper (2009) reported that even the noise made by key boards distracted their focus on the test.

[2] The studies reviewed focus upon mathematics tests taken by children who were approximately between ages 15 and 18.

These questions are addressed in this paper. Specifically, we exploit the fact that more than children from 32 economies completed both a paper and a computer version of the PISA 2012 test, allowing us to compare children's achievement across the two modes of assessment. This includes mean sores, the spread of achievement, the distribution of children across PISA proficiency levels and co-variation with gender, country-of-birth and socio-economic status. Detailed analysis of two of the computer mathematics test questions follows – one that exploits the increased functionality of computer-based assessment (requiring students to interact with the software to reach the correct answer) and one that does not (such that the question could have been administered within a 'standard' paper-and-pencil PISA test). This provides some insight into whether differences between paper and computer scores can simply be attributed to the interactive nature of some of the computer test items (which will be increasingly used in future PISA assessments) or if significant 'mode effects' are observed even in traditional PISA-style tasks.

Our results provide evidence of important and interesting differences. Despite a strong cross-country correlation, mean paper and computer test scores differ by at least 10 PISA points in one-in-three economies. There is a particularly notable decline in Shanghai-China, where children score 50 points lower (on average) on the computer-based mathematics test. Moreover, in almost every country, both the spread of achievement and the socio-economic gradient is smaller in the computer test. In contrast, the association between test administration mode and the gender gap varies significantly across countries. Yet we find little evidence that these results are being driven just by the new 'interactive' PISA questions. Rather, substantial differences are observed across groups even upon questions that require minimal interaction with the computer. In other words, the mode of assessment also has an important (and differential) impact on children's performance in 'traditional' PISA mathematics tasks. We therefore conclude that the use of computer administration in PISA 2015 represents a major change from previous and could lead to important differences in the results (if not properly accounted for).

The paper proceeds as follows. Section 2 provides an overview of the PISA 2012 paper and computer mathematics tests. Our empirical methodology is outlined in section 3, with results presented in section 4. Discussion and conclusions follow in section 5.

2. The PISA sample design and mathematics tests

The Programme for International Student Assessment (PISA) is a cross-national study of 15 year olds skills in three academic domains (reading, maths and science). A total of 65 economies participated in the 2012 round, including all members of the OECD. Within each country, a two-stage sample design was used, with a minimum of 150 schools initially selected (with probability proportional to size) and then 35 children randomly selected from within. Average response rates at the school and pupil level were high (around 90 percent in the median country). To account for the complex survey design, including the clustering of pupils within schools, student and Balance-Repeated-Replication (BRR) weights are provided by the survey organisers. These are applied throughout the analysis.

The main PISA 2012 test, used to create the final country rankings, was conducted via a traditional paper-and-pencil assessment. It took children two hours to complete. Mathematics was the 'major domain', to which the majority of test questions were devoted. Consequently, mathematics is also the focus of this paper. The test required children to demonstrate their skills in four mathematics content areas ('quantity', 'space and shapes', 'change and relationships', 'uncertainty and data') using three different cognitive processes ('employ', 'formulate' and 'process'). Further details on these sub-domains can be found in OECD (2014: Chapter 2). Final proficiency estimates are available for each pupil in each content and process area, along with a score for the mathematics domain overall. These proficiency estimates are recorded as a set of five 'plausible values', created via a complex item-response theory (IRT) model (see Von Davier and Sinharay 2013). The intuition is that, as each child only answers a random sub-set of questions, their 'true' ability in mathematics cannot be directly observed, and must therefore be estimated from their answers to the test. Recommended practise is followed throughout this paper (OECD 2005), with all estimates produced five times (once using each plausible value) and then averaged to produce the final results.

In addition to this two hour paper-based assessment, 32 economies also conducted a 40 minute computer-based mathematics test. The paper assessment was typically administered in the morning, with the computer test following in the afternoon (ACER 2011:6). Within most countries, a random sub-sample of 18 children within each participating school were required to complete both versions of the PISA assessment (OECD 2014b:78) [3]. Brazil, Spain and Italy

[3] These 18 children were then further randomly assigned one of 24 test forms which contained two out of three possible 'clusters' of questions (mathematics, digital reading and problem solving).

were exceptions, where a probabilistically drawn sub-sample of schools were firstly selected, and then a sub-sample of approximately 18 students from within. In total, 126,126 children from across these 32 economies were assigned to complete both the paper and computer PISA tests. The number of children who skipped the computer test, or could not complete it due to technical problems, was low (approximately 95 percent of children sampled to complete the computer test did so). Appendix A provides further details on the computer test sample sizes and response rates by country.

The PISA computer and paper mathematics tests were designed using the same analytical framework (see OECD 2013). Table 2 provides a summary, illustrating their similarity in terms of mathematics content, question context and cognitive processes. In other words, the two tests were both attempting to capture the same latent mathematics trait. Moreover, the PISA 2012 technical report explicitly states that:

"*the computer-based mathematics scale was equated to the paper-based mathematical scale so the results could be compared for the two modes of assessment*" (OECD 2014:253).

The issue of comparability was further discussed in a short one-page annex to the official PISA 2012 report (OECD 2014:491). The key point to note is that the two tests have been designed and jointly scaled (equated) so that the paper and computer mathematics scores can be directly compared.

< Table 2 >

It is nevertheless important to discuss how findings from such a comparison should be interpreted. Although both tests were designed to measure the same underlying mathematics skills, there was no direct overlap of questions. (I.e. None of the paper items also appeared in the computer test, or vice-versa). Moreover, some of the computer items presented children with tools and tasks that would not be possible using traditional paper-based methods. (For instance, some of the computer mathematics questions asked children to draw shapes with a tool or use an interactive graph). The PISA mathematics test framework (OECD 2013:44) therefore describes how each of the computer-mathematics test questions has three elements:

- The mathematic competency being tested
- The basic ICT skills required to correctly answer the question
- The extent to which ICT needs to be used to solve the problem

The first element is tested by all PISA mathematics questions, and does not depend upon the mode of assessment (paper or computer). The second element is specific to computer-based testing. It reflects the fact that, in order to provide an answer, children must have a minimum level of computer skill. The test developers state that all computer items were designed to '*keep such skills to a minimum core level in every computer-based item*' (OECD 2013:44). It is thus the third and final element that differs substantially between the computer test items. Specifically, whereas some computer items were very similar to 'standard' paper PISA questions, others required more interaction with the software's tools. For instance, children may have to manipulate on-screen instruments such as interactive graphs, or create a chart using data.

Consequently, any difference in overall paper and computer mathematics scores could be due to either:

- Differences across groups (including countries) in children's basic ICT skills
- Differences across groups (including countries) in the extent children can use ICT to solve mathematics problems
- The impact of "mode effects", where questions become easier or harder depending upon whether they are presented in a paper or computer test environment. (Such effects occur even when children do not need to interact with the computer to reach the correct answer). These may also differ between groups (including countries)
- Test fatigue or boredom – as the paper test was always conducted first (in the morning) with the computer test following after lunch.

Unfortunately, when considering differences in overall performance between the paper and computer mathematics tests, it is not possible to distinguish between these completing explanations (at least with the data available). However, indicative evidence can be provided by exploring children's responses to particular test questions. In particular, the OECD has released some of the PISA 2012 mathematics items (see http://erasq.acer.edu.au/) which fundamentally differ in their required level of computer skill.

We consider children's responses to two particular test questions within our analysis. The first asks children to interpret a simple table and graph about Body Mass Index (BMI) in a fictional country. This can be found in Appendix B, with an online version available from

http://erasq.acer.edu.au/index.php?cmd=cbaItemPreview&unitId=25&item=2[4]. This question is very similar to a 'standard' PISA paper item, with it possible to directly answer the question from the information presented. No further use of the computer or manipulation of the data is required. In other words, it largely rules out the second bullet point above. It thus represents a good example of a question where 'mode effects' can be explored: do children perform much better or worse on this question than one would anticipate given their scores on the paper-based test (and does this differ between demographic groups)? If so, this will suggest that mode of assessment per se is important, and that overall differences between paper and computer scores is not simply picking up differences in children's ability to solve mathematics problems interactively using ICT.

In the second item, children are asked to use an interactive graph and price calculator to derive a formula for CD production. This question can be found at http://erasq.acer.edu.au/index.php?cmd=cbaItemPreview&unitId=23&item=2 and has been reproduced in Appendix C[5]. There are clear differences to the BMI question discussed above. It is not possible to ask children this question using a traditional paper-and-pencil assessment, as the test-taker must use the computer tools (e.g. the interactive graph) to reach the correct answer. It will thus be influenced by all the factors listed above; children's ICT skills, their ability to use computers to address maths questions and mode effects (along with their proficiency in mathematics). It therefore represents a good exemplar for how children perform on computer mathematics questions that are quite different to 'standard' paper PISA items.

3. Empirical methodology

The empirical analysis proceeds in two stages. To begin, cross-country variation in children's computer and paper mathematics test scores are compared. (We remind readers that the published rankings for PISA 2012 were based only upon the paper test, with little attention given by the OECD to children's computer scores). This closely follows the OECD's presentation of PISA results in its widely cited international reports (e.g. OECD 2013b), including a comparison of mean scores, educational inequality (as measured by the standard deviation) and the distribution of children's achievement (as measured by the proportion of children within each proficiency level). Co-variation of scores with key demographic

[4] Following the terms used in Table 2, the mathematical content of this question is "uncertainty and data", the process is "interpret" and the context is "societal"

[5] Following the terms used in Table 2, the mathematical content of this question is "change and relationships", the process is "formulate" and the context is "occupational".

characteristics (gender, socio-economic status and country of birth) is also considered. Socio-economic status is measured using the PISA Economic, Social and Cultural Status (ESCS) index; this has been derived by the survey organisers from a principal components analysis of parental education, occupation and household possession, and has been standardised to a mean of zero and a standard deviation of one across participating countries (higher values indicate a more advantaged family background). Differences by country-of-birth are summarised as differences between first-generation immigrants and country natives.

Although such summary statistics are frequently reported by the OECD, there has been no direct comparison of how they differ between the paper and computer mathematics tests at the cross-country level. Our primary interest is therefore threefold. First, how strong is the cross-country correlation between the paper and computer results (i.e. do countries with high scores in the paper mathematics test also achieve high scores on the computer mathematics test)? Second, are there certain countries that buck the trend, with non-trivial differences in achievement scores between the paper and computer tests (whether on average, or at certain points along the proficiency distribution)? Finally, do certain demographic groups perform systematically better or worse on the computer based assessment, and to what extent is this a common pattern found across countries[6]?

The paper then turns to analysis of the released computer items. Our goal is to provide some indicative evidence as to whether differences between the paper and computer mathematics test can simply be attributed to the different (interactive) nature of some of the computer-based items. Or are substantial mode effects observed (and differ by demographic group) even upon questions that could be administered using either a paper or computer assessment?

We begin by considering the BMI question presented in Appendix B. Recall that this question is not interactive, requires minimal computer skills, and could easily be delivered on either a paper or computer test. The question we ask is, *conditional upon* children's performance in the paper PISA test, is there still an association between demographic group (e.g. gender, country, SES) and the probability of correct response? If so, we take this as evidence that test administration mode matters per se, and that differences in overall paper and

[6] For example, one may be concerned that low SES children do not have access to computing facilities, that this holds true in every country, and therefore the SES gap is always bigger on the computer than paper assessment

computer scores is unlikely to be solely due to the interactive nature of a sub-set of the computer-based tasks.

This issue is investigated via the following linear probability model[7]:

$$Y_{ijk} = \alpha + \beta.C_k + \gamma.P_{ijk} + \delta.U_{ijk} + \delta.I_{ijk} + \tau.A_{ijk} + \rho.F_{ijk} + \varepsilon_{ijk} \quad (1)$$

Where:

Y_{ijk} = Children's coded response to the BMI test question (0 = incorrect; 1 = correct)

C_k= A vector of country dummy variables (Reference: Shanghai-China)

P_{ijk}= Children's overall score on the PISA 2012 mathematics scale

U_{ijk}= Children's score on the 'uncertainty and data' mathematics sub-scale

I_{ijk} = Children's score on the 'interpret' mathematics sub-scale

A_{ijk}= A vector of variables capturing children's access to computers

F_{ijk} = A variable capturing the effort children put into completing the PISA mathematics test

ε = Error terms. (All elements of the PISA complex survey design, including the clustering of children with schools, is account for by application of the BRR replicate weights).

i = Student i

j = School j

k = Country k

The parameters of interest are captured in β – are there differences across countries in the probability of correct response (conditional upon the other factors included in the model)?

It is important to recognise that model 1 includes an extensive set of controls; estimates are conditional upon children's performance on the paper mathematics test overall (P_{ijk}), along with the 'uncertainty and data' (U_{ijk}) and 'interpret' (I_{ijk}) sub-domains[8]. In other words, we are controlling for children's performance on very similar questions administered in the paper version of the test. Moreover, as minimal interaction with the computer is required to reach the

[7] A linear probability model is used for ease of interpretation and due to the challenges associated with comparing parameter estimates across binary response models – see Mood (2010). We have nevertheless reproduced all estimates using a logit model as well, and obtained similar substantive results.

[8] Recall from section 2 that the BMI question belonged to these 'content' and 'process' areas. Hence why we control for these domains in our analysis.

correct answer, it is difficult to argue that any remaining differences between countries could be due to variation in children's ability to use computers to address mathematical problems. Nevertheless, we also include controls for children's access to computers and educational software at home (E_{ijk}) to further account for this possibility[9]. Finally, although there was a break between the paper and computer assessments, one may argue that test fatigue may be an issue, or that children may not exert equal effort on the paper and computer tests. We therefore also control for the PISA 'test effort' scale[10], though a limitation of this study is that we still cannot fully rule the possibility of test fatigue or differential effort out. After conditioning upon the above, we interpret the β coefficients as estimates of whether the effect of test administration mode varies across countries.

To explore whether there is also variation in the probability of correct response across demographic groups, we estimate a second set of regression models where gender, socio-economic status and immigrant status are included as covariates (these variables are added one at a time and do not appear in the model simultaneously). Formally, these models can be represented as:

$$Y_{ij} = \alpha + \beta.D_{ij} + \gamma.P_{ij} + \delta.U_{ij} + \delta.I_{ij} + \tau.A_{ij} + \rho.F_{ij} + \varepsilon_{ij} \qquad \forall K \qquad (2)$$

Where

D_{ij} = A variable capturing the demographic characteristic of interest (e.g. gender)

Note that $\forall K$ refers to the fact that model 2 will be estimated separately for each economy. The estimated parameters (β) will capture differences in the probability of correct response between demographic group (e.g. between males and females) within each country.

The modelling process outlined above is then repeated for the second exemplar computer test item (CD production – see Appendix C). The only change to models 1 and 2 is that the 'change and relationships' and 'formulate' sub-scales are now included as controls (instead of 'uncertainty and data' and 'interpret'). Our goal is to illustrate whether similar patterns hold for a question that requires much more interaction with the computer.

[9] As part of the PISA background questionnaire, children were asked whether they had certain household possessions, including a computer, educational software and access to the internet. They were also asked to report the number of computers at home. All these variables are included as controls in the analysis.

[10] This scale is based upon children's self-reports of how much effort they put into the PISA test and whether this would change had it contributed to their final school grades. See Jerrm 2014 for further details.

4. Results

Figure 1 illustrates the cross-country correlation between mean paper (x-axis) and mean computer (y-axis) test scores. The 45 degree line is where the two are identical. The association between the two sets of results is strong (Pearson r = 0.96; Spearman's rank = 0.90), with the ranking of most countries stable to whether the paper or computer test results are used. There are, however, some notable exceptions. For instance, children from Shanghai-China score around half a standard deviation lower on the computer-based mathematics test than on the paper based equivalent (562 points versus 613). Thus, although Shanghai is consistently a high-performing jurisdiction, it is only the paper mathematics test where children's scores are truly exceptional.

<< Figure 1 >>

Table 3 formally tests for differences between mean computer and mean paper test scores. Despite the strong correlation observed in Figure 1, differences are typically significant at conventional thresholds. Specifically, average mathematics test scores are significantly lower under the computer based assessment in 11 of the 32 economies at the five percent level (results for a further two countries reach significance at the ten percent level), including four of the seven high-performing East Asian jurisdictions (Shanghai, Chinese-Taipei, Hong Kong and Singapore). In contrast, there are 13 countries where children's test scores are significantly higher for the computer-based PISA test, including each of the three lowest performers (Columbia, Brazil and Chile). Although the magnitude of these differences is often modest (for 19 of the 32 countries the difference is less than 0.10 standard deviations or 10 PISA test points) there are notable exceptions. Some of the world's major economies, such as the United States (498 on computer versus 481 on paper), France (508 versus 495), Sweden (490 versus 478) and Italy (499 versus 488), are prime examples. As differences of this magnitude have previously been highlighted by the OECD as representing an important change (e.g. the OECD 2011:201 described Germany as showing 'rapid improvement' after its PISA maths scores increased from 503 in 2003 to 513 in 2009) we believe this result is of practical importance. In additional analysis (available from the authors upon request) we have investigated the correlation between the results reported in Table 3 and seven scales capturing children's use of ICT at schools and at home (this information is based upon children's self-reports collected as part of the PISA background questionnaires). The cross-country association was generally weak (Spearman's rank was below 0.4 for comparisons with six of the seven scales), although Shanghai did stand out as an economy with comparatively low scores on most scales.

Nevertheless, the overall pattern suggested that the cross-country variation observed in Table 3 was largely unrelated to the availability and use of ICT at schools and in homes.

<< Table 3 >>

Figure 2 turns to inequality in educational achievement, as measured by the standard deviation. There are three important points to note. First, 26 of the 32 countries sit below the 45 degree line. This suggests that educational inequality is lower in most countries under the computer-based mathematics assessment. (The standard deviation in the median country declines from 93 PISA points in the paper assessment to 87 points in the computer assessment). Second, there is a positive association in the cross-country inequality rankings (Person r = 0.69, Spearman's rank = 0.73), though this is clearly weaker than the relationship for the mean. Finally, for certain countries the level of educational inequality depends heavily upon whether the computer or paper assessment data are used. For instance, Chinese-Taipei is by far the most unequal country in the paper mathematics test (the standard deviation equals 116 test points) but is around the international median in the computer assessment (standard deviation of 89 points). Similar differences occur when drawing pair-wise comparisons between countries. For instance, although the spread of achievement is almost identical in Denmark and the Slovak Republic in the computer-based mathematics assessment, there is a substantial difference (approximately 20 test points) in the paper mathematics test. In additional analysis (available from the authors upon request) we reach similar substantive conclusions when using alternative measures of educational inequality, such as the difference between the 10^{th} and 90^{th} percentile. Moreover, we find that the reduction in the standard deviation seems to be driven by there being less inequality in the *top* half of the achievement distribution in the computer based test[11]. Consequently, in contrast to the strong cross-country correlation found for the mean, the distribution of mathematics achievement seems more sensitive to whether the paper or computer PISA test is used.

<< Figure 2 >>

PISA results are also presented is in terms of 'proficiency levels' – the proportion of children within each country who display a certain level of mathematics competency. How do these results differ between assessment modes? Findings for three purposefully selected economies

[11] Specifically, we find that the difference between the 90^{th} and the 50^{th} percentile is much smaller in the computer assessment than the paper assessment. In contrast, there is a more varied cross-country pattern for the difference between P50 and P10.

(Shanghai-China, Chinese-Taipei and the United States) can be found in Figure 3[12]. The uppermost bars provide results for Shanghai-China. Only 12 percent of children in this economy reached the top proficiency level in the computer mathematics test, compared to 31 percent on the paper based assessment. This suggests that the large difference between average computer and paper test scores for Shanghai in Table 3 is being driven by a decline in performance amongst the highest achievers. (Although there are also some differences in the bottom half of the Shanghai achievement distribution, these are not nearly as stark)[13]. Very similar results are obtained for Chinese-Taipei (middle set of bars), where one-in-five children reach the top proficiency level in the paper assessment compared to one-in-twenty children on the computer based test (by way of comparison, around a quarter of Chinese-Taipei children score below level 3 in both assessment modes). In contrast, there is little difference in the proportion of children reaching the top two proficiency levels in the United States (10 percent in both paper and computer) though with more notable positive change in the lower half of the proficiency distribution (e.g. 43 percent of US children score below level 3 on the computer test versus 52 percent on paper). This again illustrates how, for some countries, there are important differences in results between the two versions of the PISA mathematics test.

Next, we turn to co-variation between PISA scores and three important demographic characteristics: socio-economic status (SES), gender and country of birth. Figures 3, 4 and 5 illustrate how the association between these variables and children's test scores differ across the computer and paper mathematics tests. These are supplemented by Table 4, which formally tests for significant differences between modes for each country.

<< Table 4 >>

<< Figures 3 to 5 >>

Figure 3 illustrates the association between a one standard deviation increase in the ESCS index and children's paper (x-axis) and computer (y-axis) mathematics scores. (Recall that ESCS is the OECD's preferred measure of children's socio-economic background). It is striking that every country except Brazil falls below the 45 degree line; in almost every country, socio-economic status has a weaker association with children's computer mathematics scores (compared to the association with paper scores). Although the magnitude of this difference is

[12] These countries have been chosen for further exploration given the large and statistically significant change in their mean test scores.

[13] For instance, 11 percent of Shanghai children reach level 2 on the computer assessment compared to 20 percent on the paper test.

often modest (Table 4 suggests that, in the median country, the socio-economic gradient is approximately 5 PISA test points lower in the computer test) statistical significance is still reached in 27 economies at the five percent level (results for Germany are additionally significant at the ten percent level). Yet, in terms of cross-country rankings, there remains a high degree of consistency between the two sets of results (Pearson correlation = 0.94; Spearman's rank = 0.94). This is because, as Figure 3 illustrates, the vast majority of countries are simply 'shifted' by a uniform amount. This is nevertheless an important difference – highlighting how measures of inequality of educational opportunity may decline as PISA moves towards computer based assessment.

A somewhat different pattern emerges for gender differences in Figure 4. (Results refer to the mean score of boys minus the mean score of girls). First, in the vast majority of countries (28 out of 32), the average score of boys is higher than the average for girls under both assessment modes. Second, Table 4 suggests that the gender gap tends to be somewhat bigger in the computer-based assessment, with the difference between modes reaching significance at the five percent level in 20 out of 32 economies (Belgium additionally reaches significance at the ten percent level). Finally, perhaps the most notable feature of Figure 4 is the wide dispersion of data points, suggesting some disagreement in the cross-country rankings. The correlation is moderate (Pearson correlation = 0.60; Spearman's rank = 0.59) with dramatically different conclusions reached for certain pairwise country comparisons. For instance, consider Shanghai and the United States. Gender differences in the paper assessment are almost identical in these two economies (approximately 5 test points). Yet whereas the gender gap falls to zero in the US in the computer mathematics test, it increases in Shanghai-China to almost 20 PISA test points. This illustrates how cross-country comparisons of gender differences depend greatly upon which version of the PISA mathematics test is used.

To conclude this sub-section, Figure 6 provides analogous results for country of birth, focusing upon differences in mathematics scores between natives and first-generation immigrants. Once more, there is a strong cross-national correlation (Pearson r = 0.86; Spearman's rank = 0.89) with the broad ranking of countries similar under both the paper and computer assessments. However, as Table 4 illustrates, there are some notable exceptions. In Germany, the immigrant-native gap is 24 points higher in the computer mathematics test than the paper test, with similar differences observed in Sweden (25 points) and Shanghai (30 points). Overall, a statistically significant difference is observed between the paper and computer results on 12 occasions at the five percent level (Ireland is additionally significant at

the ten percent level). This suggests that, for certain countries, estimates of the immigrant-native gap differ non-trivially across the two test modes. At the same time, neither the pattern nor the strength of these results are as pronounced as those for gender or family background.

Item-level analysis

The previous sub-section highlighted important differences between the PISA paper and computer test results. But is this simply due to differences in the mathematics tasks children were asked to complete on the computer assessment? Or is there evidence that certain groups perform better (or worse) than one would predict upon computer-administered questions, even when they are very similar to paper based equivalents (and require minimal interaction with the computer)? We now provide evidence on this matter via analysis of two of the released computer items.

Results for the BMI question are presented in the left-hand side of Table 5. (These are the results from model 1 presented at the end of section 3). Estimates refer to percentage point differences in the probability of correct response relative to children in Shanghai-China (reference group). Recall that all figures are conditional upon test effort, access to computers and performance in the paper mathematics test (both overall and achievement within specific sub-domains). There are two key points to note. First, differences between Shanghai and most other economies are large, positive and statistically significant at the five percent level. This is therefore an example of a computer test item where children in Shanghai do not perform as strongly as one would expect, given their scores on the paper assessment. (It therefore also contributes to the lower average computer test scores of this economy highlighted in Table 3). Second, there is also a great deal of variation in the conditional probability of success across the other 31 countries. For instance, the parameter estimate for Italy (31 percentage points) is almost twice that of Spain (15 percentage points). Given that this item requires minimal interaction with the computer, these results suggest that the mode of administration is likely to matter per se to children's responses (with the impact varying across countries).

<< Table 5 >>

Further insight is provided in Table 6, where differences in the probability of correct response is compared across demographic groups *within* each of the 32 economies. (These are the results from model 2 presented at the end of section 3). The left and right hand columns suggest that, after controlling for children's paper test scores, family background and country of birth no longer have a significant impact upon the probability of success. In contrast, the middle

columns indicate that there is a strong and statistically significant association with gender in 14 economies. Specifically, boys are up to 16 percentage points less likely to answer the question correctly than girls (after controlling for their paper test scores). Again, there is evidence of modest variation across countries, with effect sizes ranging from essentially zero in Chinese-Taipei and Slovenia to 15 percentage points or more in Belgium, Ireland and South Korea. (In a pooled regression, combining data from all 32 economies, we find boys are six percent percentage points less likely to provide the correct response than girls, and that this difference is statistically significant at the five percent level). Together, this provides further evidence that even 'traditional' PISA style questions seem to be affected by assessment mode (with evidence again suggesting the impact varies across demographic groups).

<< Table 6 >>

The right-hand column of Table 5 turns to the question on CD production. Recall that this is an example of a 'new-style' PISA mathematics question, requiring children to interact with the testing software and its tools. Estimates for all country coefficients are negative, with 26 out of 32 reaching significance at the five percent level (a further two are significant at the ten percent level). This is therefore an example of a question where children from Shanghai perform better than children in most other countries – above and beyond what one would expect given their performance on the paper mathematics test. Less variation is observed, however, across the 31 other economies, with most parameter estimates falling between five and 15 percentage points. These findings are clearly rather different to those for the BMI question presented in the left-hand side of Table 5. Indeed, there is almost no association between the BMI and CD production results (Pearson correlation = 0.10). Consequently, a common cross-national pattern does not seem to hold across all computer test items – children in certain countries do better than predicted in some question but a lot worse in others[14].

Finally, Table 7 presents differences in the probability of correct response to the CD production question across demographic groups. The left and right hand columns once more suggest little association with family background and immigrant status. However, there is notable variation by gender; in half the economies the male advantage is significantly bigger than one would predict given children's scores on the paper mathematics test (nine estimates are significant at the five percent level and a further seven at the ten percent level). Although

[14] Of course, given the limited number of items released, it is not possible to generalizable findings to all computer test questions. But these results do suggest important and interesting variation across both across items countries, and is an issue that should be explored further by the OECD (who have access to all test items).

this conditional gender gap is relatively small in most countries (typically less than five percentage points) these results provide another example of how certain types of computer based questions may advantage certain demographic groups[15].

<< Table 7 >>

5. Discussion and conclusions

The Programme for International Student Assessment (PISA) study has been conducted to-date using paper-and-pencil tests. This is set to change, however, from 2015 when most countries will start to utilise computer-based assessments. There are a number of reasons why this may influence children's (and countries) results, including challenges with reading questions on-screen to different types of distraction during test administration. Using data from PISA 2012, where more than 200,000 children from 32 economies completed both a paper and computer version of the mathematics test, this work has examined mode differences in children's test scores. The results provide a first insight into how the introduction of computer-based testing may influence the PISA rankings.

We find a very strong cross-national correlation between average paper and computer test scores: the same group of countries are identified as high and low performers under both modes of the assessment. However, in a third of economies, mean paper and mean computer PISA scores were also found to differ by at least ten points (0.1 standard deviations) - a magnitude previously described as substantial (e.g. OECD 2011:201). Common measures of educational inequality were consistently lower under the computer-based mathematics assessment, both in terms of the spread of achievement and the impact of family background. On the other hand, cross-country comparisons of the gender gap were particularly sensitive to the switch between paper and computer assessment. By analysing two of the released items, we demonstrate that these results are unlikely to be solely driven by the 'interactive' questions introduced into parts of the computer assessment. Rather, notable differences across demographic groups (including across countries) occur even in computer test questions that are very similar to 'standard' paper-and-pencil items.

What do these findings imply for analysis and interpretation of future PISA waves? Clearly, the move to computer-based testing is an important change to assessment procedures, potentially influencing how children interpret, engage and answer test questions. Our analysis

[15] Again, although Table 7 shows that boys then to do better than girls on this particular interactive computer question, it is not possible for us to say whether this is a generalizable to all interactive computer items).

indicate that this is likely to lead to important differences to country-level results. Although this may only be an interpretational issue when drawing comparisons *within* future PISA assessments, it is likely to be of substantive importance when looking at trends *across* PISA cycles (including change in country scores over time). At the time of writing, the survey organisers are aware of the challenges such 'mode effects' present, and have been exploring this issue using field trial data from PISA 2015. It seems likely that a statistical adjustment to the 2015 results will be proposed, in an attempt to facilitate comparisons to previous (2000 to 2012) PISA waves. Whether statistical procedures can fully account for mode effects is an important issue, though unfortunately one that is beyond the scope of this paper. Nevertheless, close scrutiny of the methodology proposed is planned in future work (Jerrim, Micklewright and Ploubidis *forthcoming*). In the meantime, readers should take heart from some of our findings while exercising caution given others. While cross-country rankings are, on the whole, quite consistent across paper and computer versions of the PISA mathematics test, there are also some important differences. Therefore, based upon current evidence, we advise academics, policymakers and journalists to take great care when interpreting results from PISA 2015.

These findings should, of course, be considered in light of the limitations of this study. First, a randomised design has not used to allocate children to paper and computer tests (what many consider to be the 'gold standard' approach in this literature). Rather, children first completed the paper PISA test then, after a break, completed the computer based assessment. It is therefore impossible to rule out test order (and associated factors such as fatigue) as an alternative explanation for our results. Second, although the paper and computer tests were designed according to the same framework, and had similar content coverage, the test questions were not identical. Although our analysis of two of the released items has provided some insight into this issue, our ability to identify 'pure' mode effects (i.e. differences occurring simply due to computer administration of the test) remains limited. Third, in the introduction we described why differences between computer and paper test scores may occur (and why this may differ between groups). While our analysis has indeed established the existence of such group differences, it has not been possible to identify the specific driving force. Finally, the analysis has considered just one of the core PISA domains. Yet both the reading and science will also be assessed via computer from 2015. Future work should consider whether similar findings hold for these subject areas as well.

Despite these limitations, this paper has the potential to make an important contribution to our understanding of mode effects and our interpretation of international educational

achievement rankings. It is the first study to illustrate how PISA rankings differ between paper and computer versions of the mathematics test, not only in terms of mean scores, but also co-variation with key demographic characteristics. Given the prominence of such statistics in the OECD's international PISA reports, it provides an important first insight into how computer testing may influence the country-level results.

References

ACER. 2011. 'PISA 2012 main survey test administrator's manual.' Accessed 13/02/2015 from

Bennett, Randy; James Braswell; Andreas Oranje; Brent Sandene; Bruce Kaplan and Fred Yan. 2008. 'Does it Matter if I Take My Mathematics Test on Computer? A Second Empirical Study of Mode Effects in NAEP.' *The Journal of Technology, Learning, and Assessment* 6(9).

Davis, J. (2003). *Comparability study of CBT and PBT test for Mississippi DOE 2003* (Research Report). San Antonio, TX: Harcourt Assessment, Inc.

Davis, J. (2004). *Comparability study for algebra, biology, English, and history: Fall 2003 online version administered in spring 2004* (Research Report). San Antonio, TX: Harcourt Assessment

EQAO. 2014. 'EQAO to offer provincial student assessments online.' Accessed 05/02/2015 from www.eqao.com/NR/ReleaseViewer.aspx?Lang=E&release=b14R008

Fluck, A; D. Pullen and C. Harper. 2009. 'Case study of a computer based examination system.' *Australian Journal of Education al Technology* 25(4): 509-523.

Ito, K., & Sykes, R. C. (2004, April). *Comparability of scores from norm-reference paper-and-pencil and web-based linear tests for grades 4-12*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Jabr, F. 2013. 'The reading brain in the digital age: the science of paper versus screens.' *Scientific American.* Accessed 06/02/2015 from http://www.scientificamerican.com/article/reading-paper-screens/

Jerrim, J. 2015. 'Why do East Asian children perform so well in PISA? An investigation of Western-born children of East Asian descent.' *Oxford Review of Education.* Forthcoming.

Jiao, H., & Vukmirovic, Z. (2005). *Comparability between the spring 2005 online and paper-and-pencil tests of the Mississippi Subject Area Testing Program* (Research Report). San Antonio, TX: Harcourt Assessment.

Johnson, M. & Green, S. 2006. 'On-Line Mathematics Assessment: The Impact of Mode on Performance and Question Answering Strategies.' *Journal of Technology, Learning, and Assessment, 4*(5).

Mangen, A.; B. Walgermo and K. Bronnick. 2013. 'Reading linear texts on paper versus computer screen: Effects on reading comprehension'. *International Journal of Educational Research* 58: 61-68.

McKee, L & E. Levinson. 1990. 'A review of the computerized version of the Self-Directed Search'. *Career Development Quarterly 38*(4): 325-333.

Mood, Carina. 2010. "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.

Mueller, D. and V. Wasser. 1977. 'Implications of changing answers on objective test items.' *Journal of Educational Measurement 14*(1): 9-14.

OECD. 2005. "PISA 2003 Data Analysis Manual." Paris: OECD.

OECD (2011), *Lessons from PISA for the United States*, Strong Performers and Successful Reformers in Education, OECD Publishing. *http://dx.doi.org/10.1787/9789264096660-en*

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

OECD (2014), *PISA 2012 Results: What Students Know and Can Do – Student Performance in Mathematics, Reading and Science (Volume I, Revised edition, February 2014*), PISA, OECD Publishing. *http://dx.doi.org/10.1787/9789264201118-en*

OECD (2014b), *PISA 2012 Technical Report*, PISA, OECD Publishing.

Platt, L. 2014. 'Millennium Cohort Study: Initial findings from the age 11 survey.' London: Centre for Longitudinal Studies.

Poggio, J., Glasnapp, D.,Yang, X., Beauchamp, A. and Dunham, M. 2005. *A study of comparability of item and score results from computerized and paper and pencil mathematics testing in a state large scale assessment program*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Vispoel, W; T. Wang; R. de la Torre; T. Bleiler and J. Dings. 1992. *How review options and administration mode influence scores on computerized vocabulary tests*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Von Davier, Matthias and Sandip Sinharay. 2013. "Analytics in International Large-Scale Assessments: Item Response Theory and Population Models." Pp 155 – 174 in Leslie Rutkowski, Matthais von Davier and David Rutkowski (eds) *Handbook of International Large-Scale Assessment*. CRC Press.

Wang, Shudong; Hong Jiao, Michael J. Young, Thomas Brooks and John Olson. 2007. 'A Meta-Analysis of Testing Mode Effects in Grade K-12 Mathematics Tests.' *Educational and Psychological Measurement* 67: 219

Table 1. The effect of test administration mode upon upper secondary school students mathematics test scores: a summary of the literature

	Effect size	Upper CI	Lower CI
Davis (2003)	-0.41	-0.50	-0.31
Ito and Sykes (2004)	-0.24	-0.44	-0.08
Davis (2004)	-0.23	-0.36	-0.11
Bennett et al (2008)	-0.14	-0.02	-0.28
Poggio et al. (2005)	-0.07	-0.16	0.02
Wang et al. (2004)	0.00	-0.12	0.12
Jiao and Vukmirovic (2005)	0.04	-0.20	0.20

Notes: Studies restricted to those focusing upon 15 to 18 year olds scores in mathematics test. Figures refer to differences in mean test scores (computer minus paper). Negative figures indicate children perform worse on the computer test. All figures except Bennett et al taken directly from Wang et al (2007).

Table 2. Mathematics content, process and context of the PISA 2012 computer and paper based exam

	Computer	Paper
Content		
Quantity %	22	26
Space and shape %	29	25
Change and relationships %	27	27
Uncertainty and data %	22	23
Process		
Employ %	54	46
Formulate %	22	29
Interpret %	24	25
Context		
Personal %	32	19
Societal %	27	33
Occupational %	22	22
Scientific %	20	26
Number of items	**41**	**109**

Source: Authors calculations using data from OECD (2014: Annex A6).

Table 3. Difference between mean paper and mean computer maths test score by country

Country	Abbreviation	Difference	SE	SIG
Shanghai-China	QC	-50	2.6	**
Poland	PL	-28	2.5	**
Chinese-Taipei	TP	-23	2.4	**
Israel	IS	-20	2.9	**
Slovenia	SL	-14	0.7	**
Hong Kong	HK	-12	2.6	**
Ireland	IE	-8	2.5	**
Spain	ES	-8	3.0	**
Singapore	SG	-7	0.8	**
Hungary	HU	-7	2.4	**
Estonia	ET	-4	2.0	**
Germany	DE	-4	2.2	*
Denmark	DK	-4	2.0	*
Belgium	BE	-2	1.7	-
Korea	KR	-1	2.5	-
Argentina	AR	0	1.9	-
Austria	AT	2	2.6	-
Portugal	PT	2	2.6	-
Japan	JP	3	1.9	-
Australia	AU	4	1.3	**
Macao-China	MA	5	0.9	**
Canada	CA	5	1.6	**
Russia	RU	7	2.5	**
Norway	NO	8	2.7	**
Chile	CL	9	2.5	**
Italy	IT	11	3.2	**
Sweden	SE	12	2.6	**
France	FR	13	2.9	**
Slovak Republic	SK	16	2.0	**
United States	US	17	2.2	**
Columbia	CO	20	2.5	**
Brazil	BR	25	3.2	**

Notes: Author's calculations using the PISA 2012 dataset. Difference = computer maths score minus paper maths score. Figures refer to difference in terms of PISA test points. * and ** indicates statistical significance at the ten percent and five percent levels. Stratification and clustering of pupils within schools accounted for by application of the PISA BRR weights.

Table 4. Inequalities in educational opportunity by gender, immigrant and socio-economic status: differences between computer and paper tests

	SES gradient			Gender			Immigrant		
	Beta	**SE**	**SIG**	**Beta**	**SE**	**SIG**	**Beta**	**SE**	**SIG**
Chinese-Taipei	-16	1.7	**	10	3.0	**	-38	32.9	-
Sweden	-11	1.7	**	16	1.6	**	25	4.4	**
France	-10	1.8	**	7	2.2	**	9	6.5	-
Hong Kong	-8	2.1	**	2	2.8	-	1	3.4	-
Slovak Republic	-8	2.6	**	2	2.2	-	50	19.2	**
Denmark	-7	1.3	**	6	1.4	**	8	4.0	**
Portugal	-7	1.2	**	9	1.6	**	8	6.9	-
Austria	-7	1.5	**	-2	2.7	-	1	5.1	-
Columbia	-7	1.2	**	-13	2.1	**	25	21.8	-
Australia	-7	0.9	**	-3	1.2	**	-7	2.5	**
Slovenia	-7	0.7	**	-1	1.2	-	26	5.1	**
Belgium	-7	1.2	**	3	1.9	*	13	4.6	**
Japan	-6	1.8	**	-3	2.3	-	24	23.6	-
Chile	-6	1.4	**	-6	2.4	**	12	9.9	-
Spain	-6	1.6	**	0	2.2	-	-5	5.1	-
Poland	-6	1.4	**	7	1.6	**	-99	27.9	**
Hungary	-5	1.9	**	3	2.6	-	-38	22.9	-
Ireland	-5	1.6	**	3	3.3	-	6	3.2	*
Russia	-5	2.2	**	16	1.7	**	1	4.5	-
Canada	-5	1.3	**	7	1.1	**	6	3.0	**
Italy	-5	1.6	**	8	2.9	**	1	5.4	-
Singapore	-5	0.7	**	4	1.1	**	-10	1.8	**
Macao-China	-5	0.7	**	10	1.6	**	-5	2.8	-
Israel	-4	2.1	**	-9	3.9	**	4	6.0	-
Norway	-4	1.3	**	1	1.6	-	4	4.0	-
United States	-4	1.4	**	-4	1.8	**	2	4.7	-
Argentina	-4	1.5	**	-8	3.6	**	-15	2.8	**
Korea	-3	1.9	-	0	3.7	-	40	117.9	-
Germany	-3	1.6	*	-4	1.5	**	24	6.0	**
Shanghai-China	-2	1.9	-	13	1.8	**	30	10.2	**
Estonia	-1	1.2	-	4	1.6	**	-6	9.2	-
Brazil	2	1.6	-	1	1.7	-	-16	19.4	-

Notes: Author's calculations using the PISA 2012 dataset. Difference = gradient using computer maths score minus gradient using paper maths score. Figures refer to difference in terms of PISA test points. * and ** indicates statistical significance at the ten percent and five percent levels. Stratification and clustering of pupils within schools accounted for by application of the PISA BRR weights.

Table 5. Differences across countries in children providing correct responses to two PISA computer mathematics items

	BMI			CD Production		
	Percentage point difference	SE	SIG	Percentage point difference	SE	SIG
Shanghai (REF)	-	-	-	-	-	-
Chinese-Taipei	0	2.7	-	-5	2.9	*
Poland	1	2.2	-	-15	2.5	**
Spain	15	3.1	**	-12	2.5	**
Macao-China	16	2.4	**	-8	2.7	**
Slovenia	16	2.9	**	-13	2.5	**
Japan	17	2.7	**	-11	2.7	**
Russia	18	3.7	**	-10	2.8	**
Austria	19	3.1	**	-12	2.8	**
Denmark	20	3.5	**	-12	2.9	**
Hungary	20	2.9	**	-12	2.6	**
Israel	20	3.1	**	-7	2.9	**
Argentina	22	2.6	**	-8	2.5	**
Columbia	22	3.0	**	-3	3.1	-
Norway	23	3.2	**	-11	2.9	**
Sweden	23	3.2	**	-12	2.8	**
Germany	23	3.0	**	-13	2.8	**
Estonia	24	3.1	**	-16	2.7	**
Chile	24	3.1	**	-6	2.7	**
United States	26	3.5	**	-6	3.0	*
Portugal	26	2.8	**	-10	2.5	**
Brazil	27	3.3	**	-3	3.2	-
Belgium	28	2.9	**	-10	2.6	**
Australia	29	2.6	**	-10	2.6	**
Italy	31	3.8	**	-11	2.7	**
Slovak Republic	32	3.2	**	-11	2.3	**
Korea	33	3.0	**	-7	3.0	**
Canada	34	3.1	**	-6	2.9	**
Ireland	35	3.5	**	-16	2.6	**
Hong Kong	37	3.1	**	-13	3.0	**
Singapore	39	3.4	**	-5	2.9	-
France	43	3.0	**	-9	2.7	**

Notes: Author's calculations using the PISA 2012 database. Results from a linear probability model controlling for children's score on PISA paper mathematics test, test effort scale and information on computer possessions at home. All results refer to percentage point differences relative to children from Shanghai-China (reference group). BMI results additionally control for the "uncertainty and data" and "interpret" paper mathematics sub-scale. CD results additionally control for the "change and relationships" and "formulate" paper mathematics sub-scale. * and ** indicates probability of children providing the correct answer is significantly different to Shanghai-China at the ten percent and five percent levels. Stratification and clustering of pupils within schools accounted for by application of the PISA BRR weights.

Table 6. The association between demographic characteristics and the probability of correct response to the *BMI* computer test item

	SES gradient			Gender			Country of birth		
	% point diff	SE	SIG	% point diff	SE	SIG	% point diff	SE	SIG
Korea	-4	3.8	-	-16	4.6	**	-	-	-
Ireland	4	2.9	-	-15	5.0	**	-5	10.5	-
Belgium	-2	2.1	-	-15	3.3	**	-3	6.9	-
Brazil	0	2.2	-	-13	4.0	**	-	-	-
Hong Kong	6	2.6	**	-13	4.3	**	-17	5.9	**
Germany	1	3.6	-	-13	4.0	**	3	9.1	-
Portugal	5	2.9	-	-12	4.7	**	7	7.8	-
United States	6	2.7	**	-12	4.3	**	6	7.7	-
Norway	0	3.4	-	-11	4.8	**	4	10.1	-
Australia	-1	1.8	-	-10	2.2	**	-1	4.1	-
France	3	3.3	-	-10	4.1	**	-9	9.1	-
Israel	0	2.7	-	-10	4.0	**	-3	8.7	-
Singapore	-3	2.3	-	-10	4.4	**	-1	7.5	-
Estonia	-3	3.2	-	-7	4.6	-	-14	5.6	**
Argentina	-1	1.8	-	-5	2.1	**	9	3.3	**
Chile	5	2.3	**	-5	3.7	-	-15	6.5	**
Austria	0	2.9	-	-5	4.2	-	3	7.7	-
Sweden	-2	2.6	-	-4	4.5	-	20	8.2	**
Canada	-1	2.4	-	-4	3.6	-	6	4.9	-
Hungary	2	2.4	-	-4	4.2	-	-10	6.0	-
Slovak Republic	0	2.5	-	-4	4.0	-	-	-	-
Russia	5	2.7	*	-4	3.5	-	10	11.5	-
Spain	-2	2.1	-	-3	2.8	-	1	5.6	-
Columbia	-1	1.2	-	-3	2.7	-	1	3.3	-
Shanghai-China	6	2.2	**	-3	4.1	-	-16	8.9	*
Macao-China	-3	2.8	-	-2	3.3	-	-4	5.6	-
Japan	-4	2.2	*	-2	2.7	-	-14	11.1	-
Denmark	0	2.6	-	-2	5.0	-	6	11.0	-
Slovenia	1	2.8	-	-1	3.5	-	-1	5.4	-
Chinese-Taipei	0	2.3	-	0	3.1	-	-4	2.9	-
Poland	1	1.8	-	4	3.2	-	-	-	-
Italy	-2	3.6	-	9	6.0	-	2	11.4	-

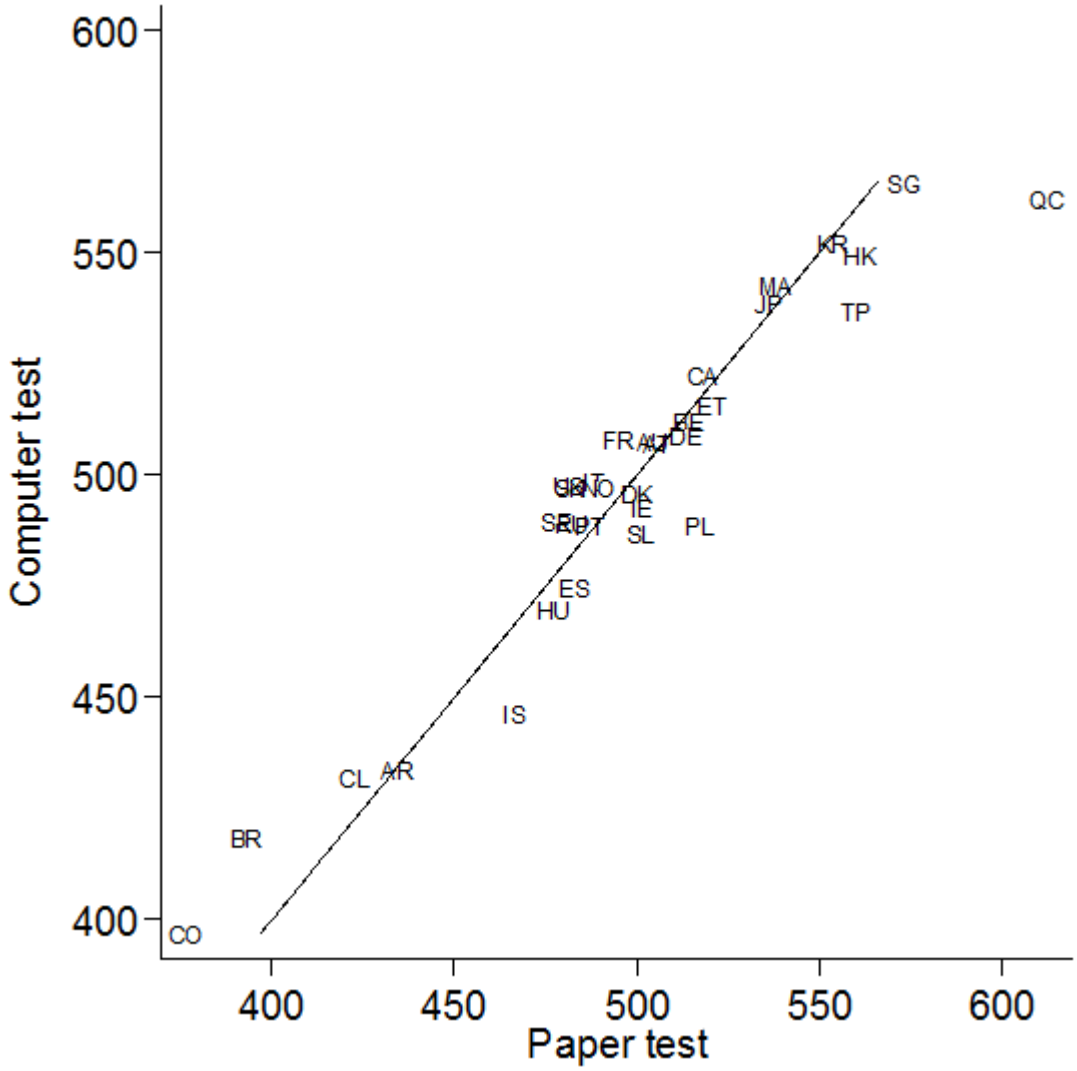
Notes: Author's calculations using the PISA 2012 database. Results from a linear probability model controlling for children's score on PISA paper mathematics test, "uncertainty and data" and "interpret" paper mathematics sub-scales, test effort scale and information on computer possessions at home. All results refer to percentage point differences. Reference groups are female (gender) and country native (country of birth). * and ** indicate significant differences relative to the reference group at the ten percent and five percent levels. Stratification and clustering of pupils within schools accounted for by application of the PISA BRR weights.

Table 7. The association between demographic characteristics and the probability of correct response to the *CD production* computer test item

	SES gradient			Gender			Country of birth		
	% point diff	SE	SIG	% point diff	SE	SIG	% point diff	SE	SIG
Sweden	2	1.0	-	-1	2.3	-	6	4.0	-
Norway	-1	1.9	-	-1	2.1	-	3	5.4	-
United States	-1	1.2	-	0	2.6	-	9	5.2	*
Chile	1	0.4	**	0	1.1	-	0	1.9	-
Columbia	1	0.6	-	0	1.0	-	4	2.0	*
Brazil	1	0.6	**	0	1.3	-	0	0.0	-
Australia	-2	1.0	*	0	1.2	-	1	2.6	-
Belgium	-1	1.4	-	1	2.5	-	3	4.8	-
Spain	-1	0.9	-	2	1.6	-	0	1.0	-
Japan	-1	1.6	-	2	1.8	-	3	13.8	-
Slovak Republic	2	1.6	-	2	1.6	-	0	0.0	-
Hungary	0	0.7	-	2	1.4	*	0	1.3	-
Argentina	-1	1.1	-	2	1.1	**	-1	1.7	-
Israel	4	1.6	**	3	3.1	-	-5	4.4	-
Hong Kong	-4	1.6	**	3	2.6	-	0	3.9	-
Russia	1	1.3	-	3	1.8	*	3	6.8	-
Germany	2	1.5	-	3	2.4	-	15	8.0	*
Slovenia	2	1.2	*	3	2.0	-	8	5.4	-
Ireland	1	0.7	-	3	1.5	**	-2	1.0	-
Estonia	-2	1.8	-	4	1.8	*	3	2.5	-
Austria	-2	1.5	-	4	2.1	*	3	3.1	-
Korea	0	2.4	-	4	3.4	-	0	0.0	-
Canada	-1	1.4	-	4	2.3	*	6	3.1	*
Italy	1	1.4	-	5	2.0	**	-1	1.7	-
Poland	1	1.3	-	5	2.0	**	0	0.0	-
Macao-China	-2	2.0	-	5	2.8	*	0	3.7	-
Portugal	0	1.4	-	5	2.4	**	1	8.2	-
Denmark	0	1.3	-	6	2.8	*	3	2.6	-
France	0	1.7	-	6	2.3	**	1	3.0	-
Chinese-Taipei	0	2.3	-	9	3.7	**	-1	5.6	-
Singapore	1	2.4	-	9	3.6	**	17	7.4	**
Shanghai-China	2	2.8	-	10	4.4	**	-22	15.8	-

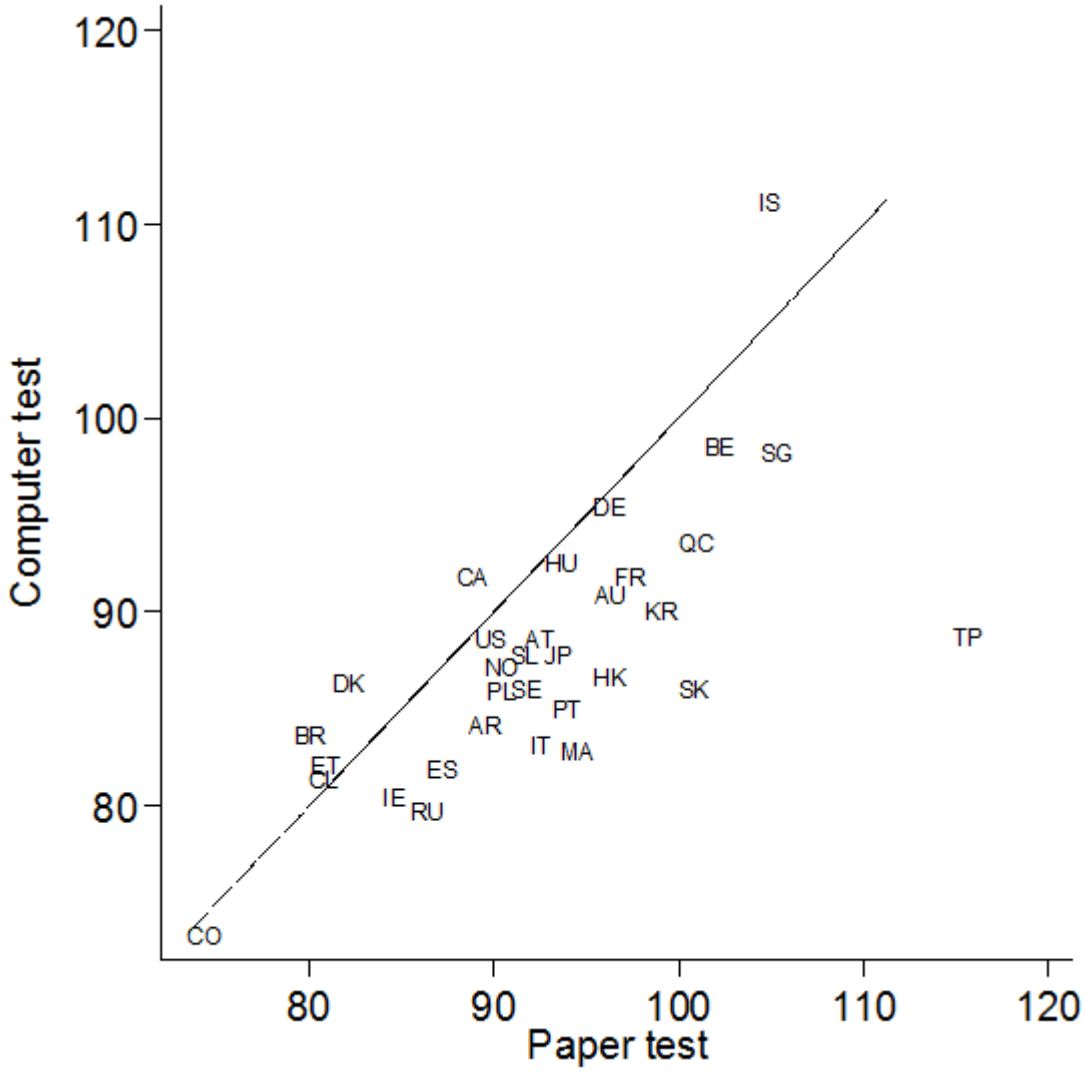
Notes: Author's calculations using the PISA 2012 database. Results from a linear probability model controlling for children's score on PISA paper mathematics test, "change and relationships" and "formulate" paper mathematics sub-scales, test effort scale and information on computer possessions at home. All results refer to percentage point differences. Reference groups are female (gender) and country native (country of birth). * and ** indicate significant differences relative to the reference group at the ten percent and five percent levels. Stratification and clustering of pupils within schools accounted for by application of the PISA BRR weights.

Figure 1. Mean country scores on the PISA 2012 computer and paper maths tests



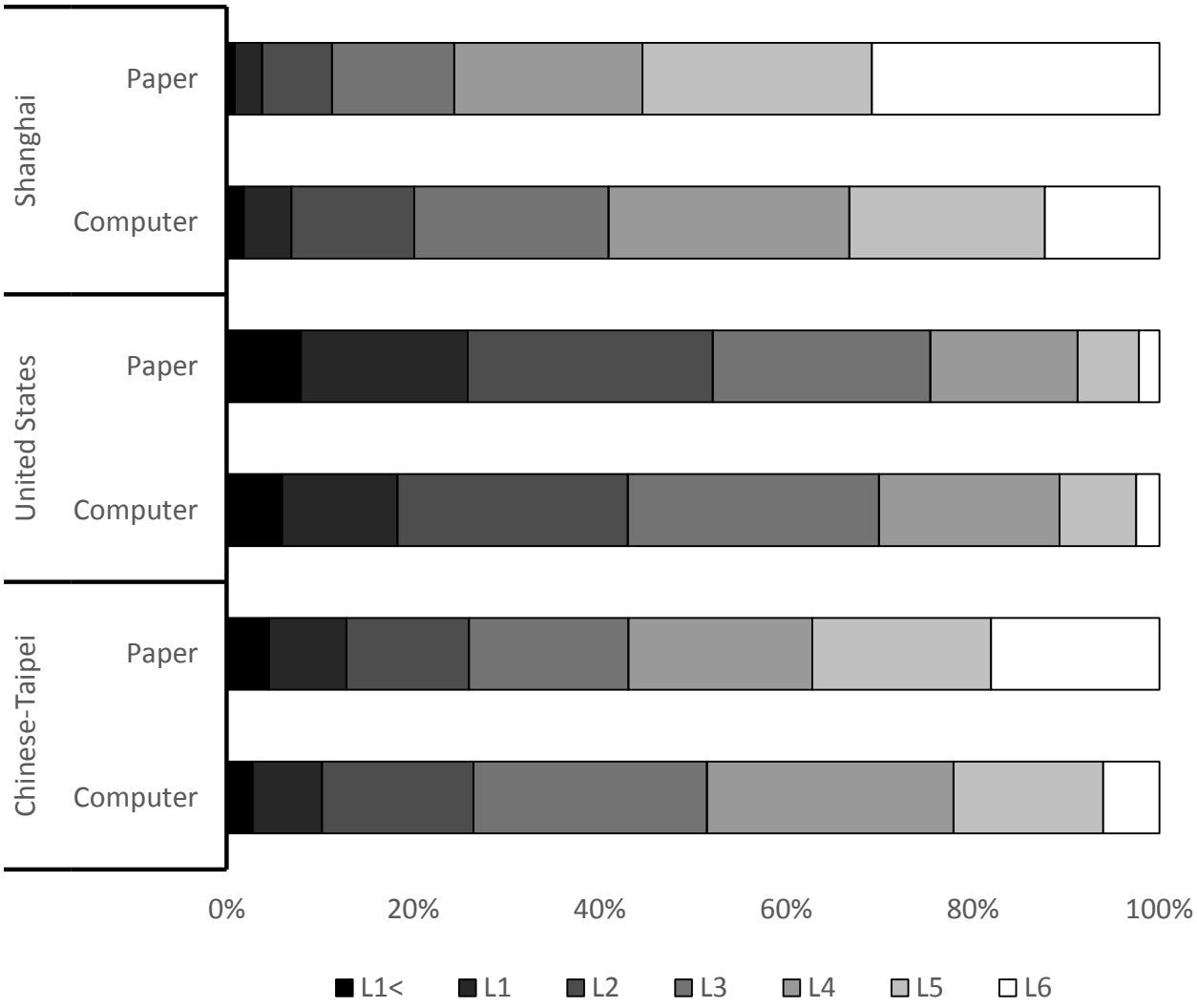
Notes: Author's calculations using the PISA 2012 dataset. Countries can be identified using their official two letter code (see Table 3). Correlation = 0.96 (Spearman rank = 0.90).

Figure 2. The correlation between country standard deviation on the PISA 2012 computer and paper maths tests



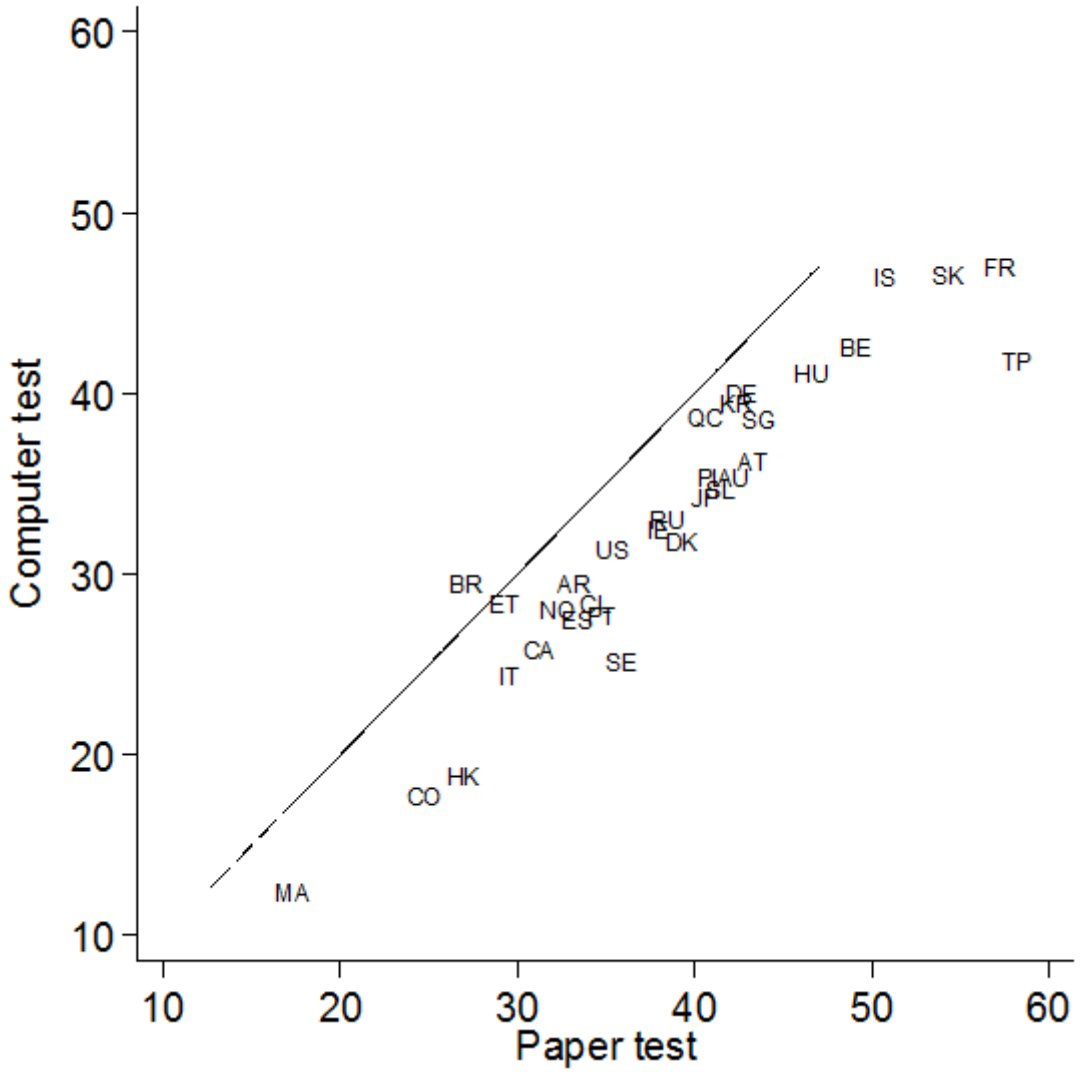
Notes: Author's calculations using the PISA 2012 dataset. Countries can be identified using their official two letter code (see Table 3). Correlation = 0.69 (Spearman rank = 0.74). The 45 degree line indicates where results are equal between the paper and computer based tests.

Figure 3. Estimated mathematic proficiency levels for selected countries: a comparison of paper and computer tests



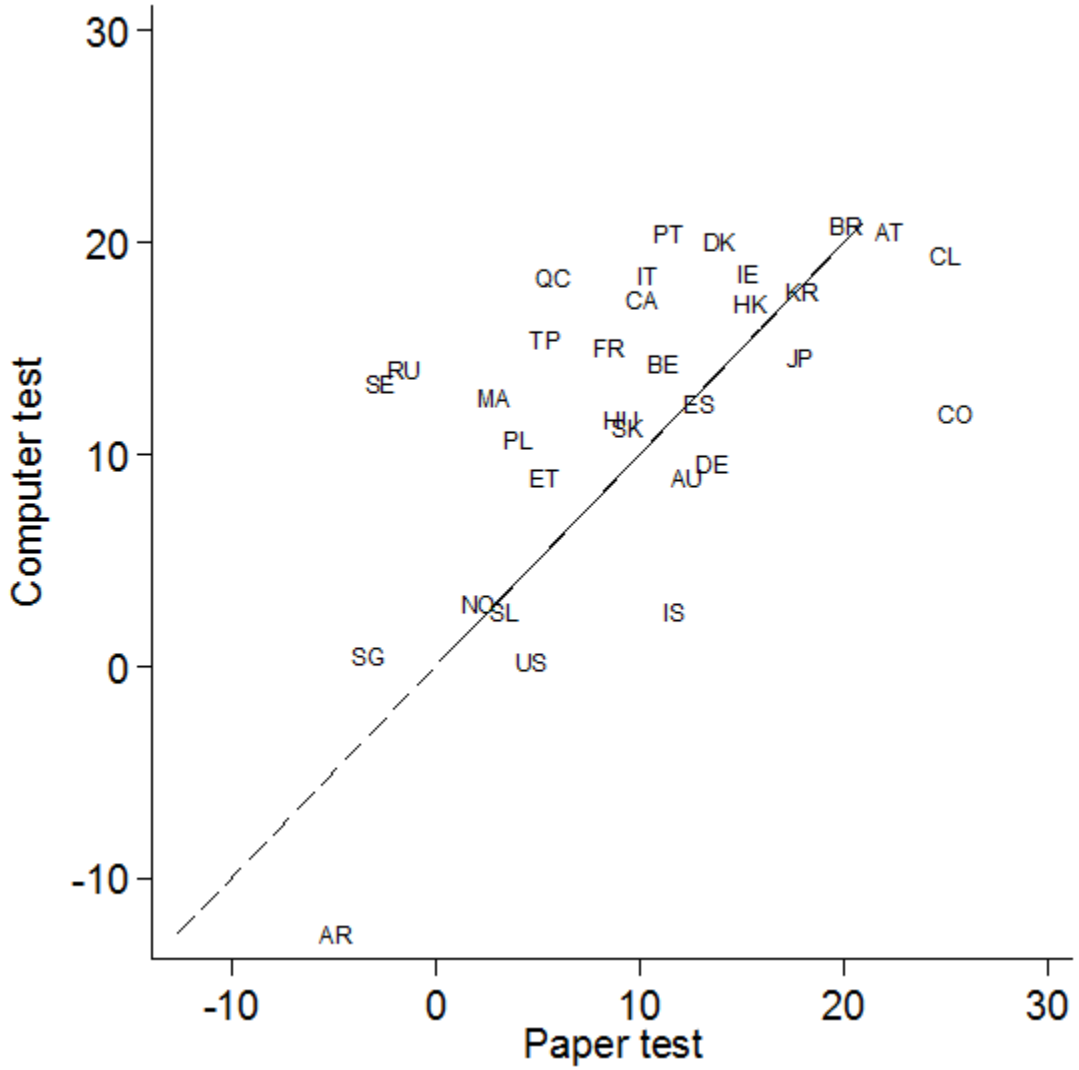
Notes: Author's calculations using the PISA 2012 dataset. Shading of bars refers to the seven PISA competency levels (ranging from below level 1 to level 6). Further description of these levels (along with example questions they correspond to) can be found in OECD (2014:61).

Figure 4. The correlation between the estimated socio-economic gradient using the PISA 2012 computer and paper maths tests



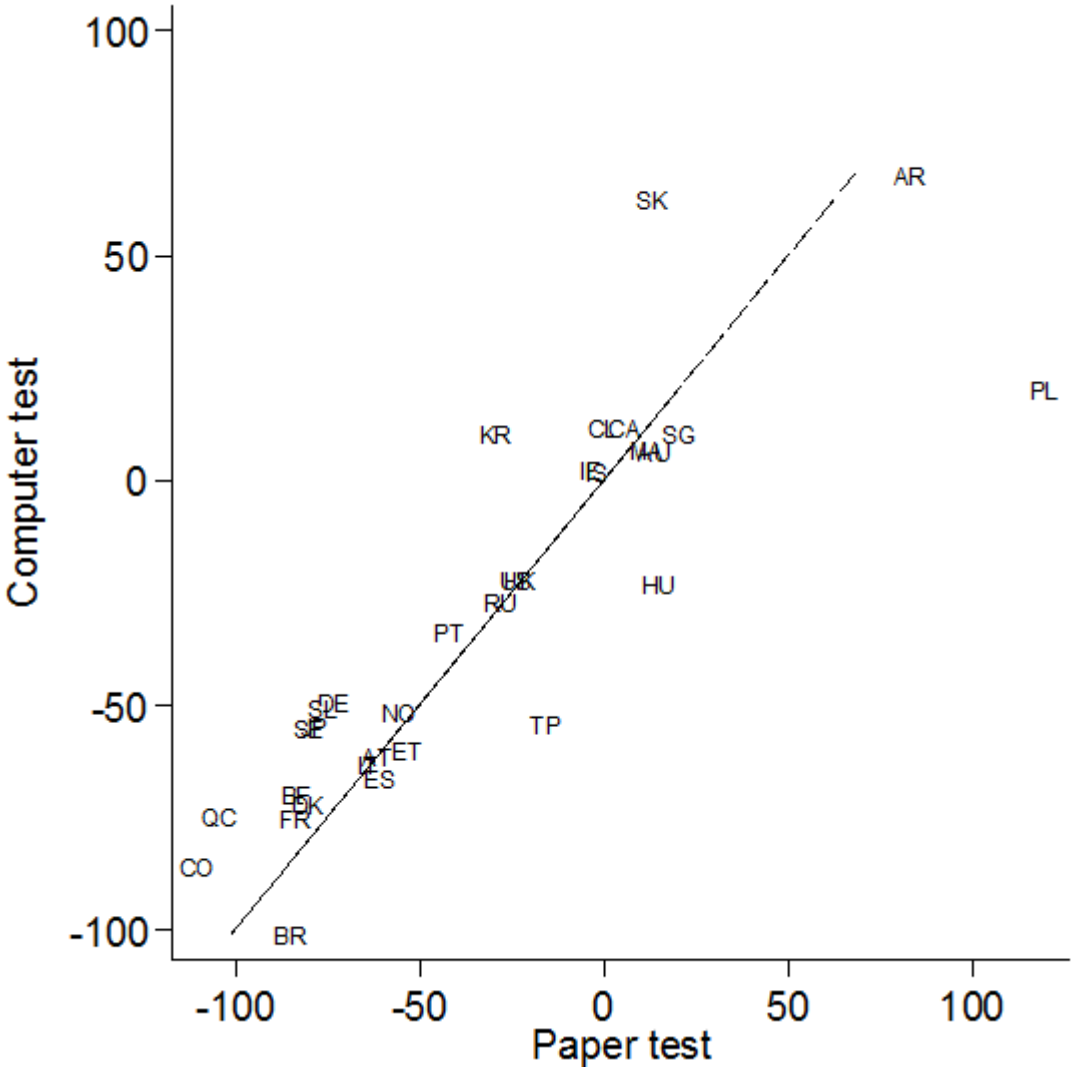
Notes: Author's calculations using the PISA 2012 dataset. Countries can be identified using their official two letter code (see Table 3). Correlation = 0.94 (Spearman rank = 0.94). The 45 degree line indicates where results are equal between the paper and computer based tests. Figures can be cross-referenced with Table 4 (left-hand results column).

Figure 5. The correlation between gender differences using the PISA 2012 computer and paper maths tests



Notes: Author's calculations using the PISA 2012 dataset. Countries can be identified using their official two letter code (see Table 3). Correlation = 0.60 (Spearman rank =0.59), falling to 0.52 (0.55) once AR has been excluded as an outlier. The 45 degree line indicates where results are equal between the paper and computer based tests. Figures can be cross-referenced with Table 4 (middle results column).

Figure 6. The correlation between <u>immigrant differences</u> using the PISA 2012 computer and paper maths tests



Notes: Author's calculations using the PISA 2012 dataset. Countries can be identified using their official two letter code (see Table 3). Correlation = 0.86 (Spearman rank = 0.89). The 45 degree line indicates where results are equal between the paper and computer based tests. Figures can be cross-referenced with Table 4 (right-hand results column).

Appendix A. Sample sizes and response rates for the computer based test by country

	Paper sample size	Computer test sample size	Assigned computer maths questions	Computer assessment response rate
Canada	21,544	10,817	5,410	85
Australia	14,481	11,834	5,921	94
Argentina	11,500	6,732	3,367	95
Spain	10,175	5,751	2,878	95
Columbia	9,073	5,173	2,618	88
Belgium	8,597	4,617	2,299	93
Denmark	7,481	4,149	2,048	93
Chile	6,856	3,341	1,675	94
Japan	6,351	6,351	3,166	95
Chinese-Taipei	6,046	3,063	1,512	98
Slovenia	5,911	4,385	2,188	95
Portugal	5,722	3,272	1,630	88
Singapore	5,546	2,873	1,436	97
Brazil	5,506	3,172	1,574	91
Italy	5,495	3,089	1,529	88
Macao-China	5,335	3,147	1,558	99
Russia	5,231	3,186	1,610	98
Shanghai-China	5,177	2,409	1,190	99
Israel	5,055	2,677	1,333	93
South Korea	5,033	2,675	1,324	99
Ireland	5,016	2,613	1,303	91
Germany	5,001	2,881	1,446	94
United States	4,978	2,572	1,276	98
Hungary	4,810	2,746	1,377	96
Estonia	4,779	2,837	1,405	97
Austria	4,755	2,731	1,367	96
Sweden	4,736	2,671	1,314	94
Norway	4,686	2,924	1,463	85
Slovak Republic	4,678	3,145	1,559	93
Hong Kong	4,670	2,714	1,356	96
France	4,613	3,012	1,512	88
Poland	4,607	2,567	1,290	99
Total	**213,444**	**126,126**	**62,934**	

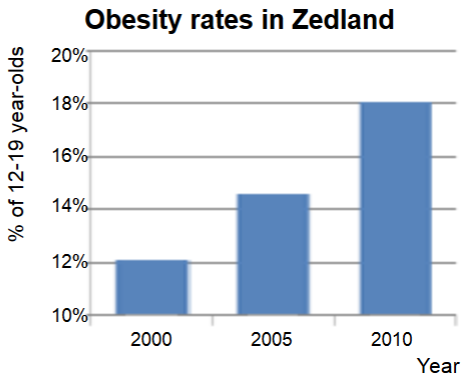
Notes: Authors' calculations using the PISA 2012.

Appendix B. Body Mass Index released PISA 2012 computer item

Question

What is one major change in the BMI classifications for 12-19 year-olds in Zedland between 2000 and 2010? Justify your answer based on value(s) from the data table.

BMI Classification:	Percentage of 12-19 year-olds		
	2000	2005	2010
Underweight	14.1%	14.8%	14.3%
Normal	57.5%	53.5%	51.5%
Overweight	16.3%	17.1%	16.1%
Obese	12.1%	14.6%	18.1%



Answer

A statement that shows a correct understanding of at least one of the two major changes given Below. Students must provide *both the size and direction* of the change:

- The percentage of 12-19 year olds within the normal weight range has decreased from 57.5% in 2000 to 51.5% in 2010 or decreased by 6% (points)
- The percentage of 12-19 year olds who are obese has increased by from 12.1% in 2000 to 18.1% in 2010 or increased by 6% (points).

Full Question link (requires Firefox):
http://erasq.acer.edu.au/index.php?cmd=cbaItemPreview&unitId=25&item=2

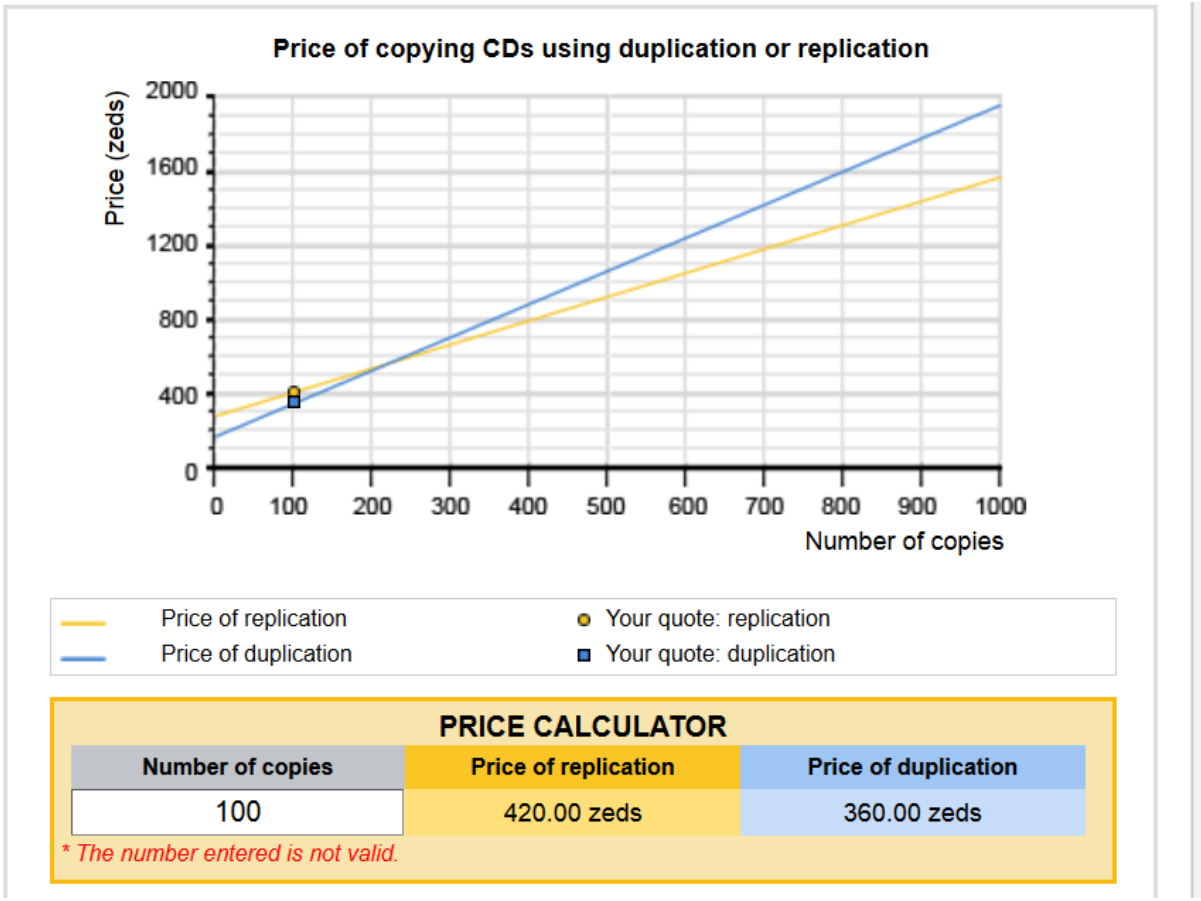
Appendix C. CD production released PISA 2012 computer item

Question

Use the graphs and price calculator to find the rule for how the price of **replication** is determined.

Write the two missing values in the rule below to show how price, *P*, relates to number of copies made, *n*, for replication.

P = _ _ _ _ n + _ _ _ _



Full credit answer

P = 1.3n + 290

Full Question link (requires Firefox):
http://erasq.acer.edu.au/index.php?cmd=cbaItemPreview&unitId=23&item=2