1. Introduction

The Programme for International Student Assessment (PISA) is a major cross-national study of 15-year-olds' academic skills. Conducted by the Organisation for Economic Co-operation and Development (OECD), the results are now eagerly awaited by academics, policymakers, journalists, politicians and parents worldwide. In recent years, trends in PISA scores over time are of particular interest, with policymakers wondering if their country has risen up or slid down these international 'rankings'. Finland, Ireland, Australia, Sweden and England are prominent examples of where an apparent decline in PISA scores has received much attention (Ryan 2013; *author cite*). Conversely, countries such as Germany and Poland have been held up as 'successful reformers' due to improvements in their PISA scores (OECD 2011).

Yet there are many challenges to measuring trends using large-scale international assessments such as PISA. Administration and analysis procedures can change between survey rounds, potentially influencing results. For instance, in the case of Ireland, Cosgrove and Cartwright (2014) discussed differing response patterns, engagement of pupils and survey procedures as possible explanations as to why reading scores declined between 2006 and 2009. Similarly, *author cite* highlighted changing response rates, target populations, test month and number of countries as possible reasons why England fell in the PISA rankings between 2000 and 2009, despite simultaneously improving in another major international assessment, TIMSS (the Trends in International Mathematics and Science Study).

It is therefore pertinent that a major change was made to the administration of PISA in 2015; most (although not all) countries moved from paper-based assessment (PBA) to computer-based assessment (CBA)[1]. The use of computers in large-scale educational studies has several attractions, including the introduction of more interactive questions, efficiencies in processing and marking, tailoring questions to match pupils' ability[2] and enabling greater insights into test-taking behaviour. Moreover, with rapid technological innovation, a change in delivery mode was inevitable at some point. Yet, in the short-term, the move to CBA poses challenges, including the potential for 'mode effects' to influence the comparability of PISA scores over time. ('Mode effects' refer to whether questions designed to be delivered on paper are

[1] Of the PISA 2015 countries, 15 completed the paper test and 58 the computer test. The paper-based countries were Albania, Algeria, Argentina, Georgia, Indonesia, Jordan, Kazakhstan, Kosovo, Lebanon, Macedonia, Malta, Moldova, Romania, Trinidad and Tobago, and Viet Nam.
[2] PISA 2015 did not use a 'computer-adaptive' assessment design.

systematically easier or harder when delivered on computer). Moreover, these mode effects may differ between countries or groups (e.g. by gender or socio-economic status).

The primary aim of this paper is to quantify the magnitude of these mode effects using PISA field trial data. A secondary aim is to examine the methodology used by the PISA 2015 international consortium at the Educational Testing Service (ETS) (the Lead Contractor for the PISA 2015 international consortium, and the contractor responsible for the scaling of the data) have used to try and account for this problem, and to investigate whether the adjustment made to PISA 2015 scores has resolved this issue. Currently, little evidence exists on this issue in the public domain, and the information that has been presented by the OECD and ETS is often not clear and transparent. This is important as average scores in PISA 2015 were around eight points lower in science (on average across OECD countries) than they were in 2012. In reference to our countries of interest, there was a statistically significant 19 point decline in science in Ireland between 2012 and 2015, a significant 19 point decline in Germany and a non-significant 10 point increase in Sweden. More generally, among the top 30 countries on PISA 2012 science, 11 had a significant decline in achievement in 2015, and just one (Portugal) had a significant increase. It is possible that mode effects could be partially responsible for this change in performance over time. Our goal therefore is to provide some clear and transparent evidence on the extent of mode effects in the PISA 2015 field trial data for Germany, Ireland and Sweden, and to provide some independent insight on whether the OECD and ETS were able to fully correct for this issue in the main study data.

There already exists a sizable literature on the impact of test administration mode upon pupils' performance in cognitive tests. From meta-analyses it can be seen that the direction and strength of the mode effect depends on different factors, such as subject area (Kingston 2009), study design, sample size, computer delivery algorithm, computer practice (Wang, Jiao, Young, Brooks and Olson 2008) and question response format - which might also be altered with the change of assessment mode (Bennett, Braswell, Oranje, Sandene, Kaplan and Yan 2008). Consequently, each study should have its own empirical verification of mode effects, due to this mixture of possible causes (see e.g. Kroehne and Martens 2011).

We contribute to this literature via analysis of PISA 2015 field trial data for Sweden, Germany and Ireland. The pupils who took part in the field trial were randomly assigned to complete either PBA or CBA versions of the same PISA questions. This allows us to estimate the causal impact of assessment mode upon pupils' performance. Our analysis includes each core PISA

domain (reading, maths and science), along with a question-by-question assessment of each trend item. It builds upon *author cite*, which used the fact that pupils in 32 countries in the 2012 cycle completed a paper-based version of the PISA test in the morning and a computer-based assessment in the afternoon. Although *author cite* found some striking results (e.g. average scores in Shanghai were half a standard deviation lower on paper than computer, mode effects differed significantly by gender and across countries) the findings were limited due to the study design and the fact that different questions were used across the paper and computer assessments[3]. In contrast, this paper uses data from a randomised design, with one set of pupils randomly assigned to answer paper versions of the PISA questions, with the other answering the same questions on computer. The analysis presented in this article hence has greater internal validity than the evidence presented in *author cite*.

We find consistent evidence of mode effects, and that these are of greater magnitudes than typically reported elsewhere in the literature. Having established the presence of mode effects, we then describe how the international consortium responsible for PISA 2015 has attempted to account for this problem in constructing the PISA 2015 scores. We conclude by providing an independent assessment of whether this methodology is likely to have reduced the problem of mode effects.

This paper therefore contributes to the literature in several ways. First, we add new evidence on the impact of computer assessment at a time when this mode is becoming increasingly prevalent in developed countries. Second, it represents the only independent study of how change in assessment mode influences pupils' responses to PISA 2015 trend questions, and how this may have influenced the data collected in PISA 2015[4]. Third, we provide an independent view of the international consortium's methodology of adjusting for mode effects, with the evidence presented in the public domain by the OECD and ETS on this matter being rather limited.

The paper proceeds with section 2 describing the PISA field trial data and our empirical methodology. Initial results are presented in section 3. Section 4 turns to how the international

[3] In PISA 2012, as the paper test was typically conducted in the morning and the computer test in the afternoon, *author cite* could not rule out fatigue or test motivation as possible confounding explanations for the difference in paper and computer scores. Likewise, differences could have been due to the specific test questions posed, rather than being an effect of assessment mode per se.

[4] By 'independent', we are referring to outside the international consortium responsible for delivering PISA 2015. The Educational Testing Service (ETS), as part of the international consortium, have also investigated the issue of mode effects (ETS 2015).

consortium have attempted to account for mode effects in PISA 2015, while section 5 provides an independent investigation into how successful this adjustment is likely to have been. Conclusions are then provided in section 6.

2. Data and methods

The PISA 2015 field trial design

The PISA 2015 field trial was conducted in spring 2014. All countries making the switch from paper to computer assessment took part. This paper is based upon data from Germany, Sweden and Ireland; countries where data is available.

The design of the field trial was led by ETS. Each participating country was asked to recruit pupils and schools using one of the following designs:

- Design A = Recruit 25 schools and 78 pupils within each
- Design B = Recruit 39 schools and 52 pupils within each
- Design C = Recruit 54 schools and 36 pupils within each

There was no requirement for these schools to be randomly (or probabilistically) sampled[5]. With respect to our countries, Ireland followed design A, Germany design C and Sweden design C. However, due to pupil exemptions (e.g. for pupils with special educational needs who would not be able to complete the test without assistance and pupils with limited language experience), logistical restrictions and challenges with recruitment, these criteria were not always met. Final achieved sample sizes were therefore:

- Germany = 62 schools and 2,341 pupils
- Ireland = 25 schools and 1,503 pupils
- Sweden = 54 schools and 2,141 pupils

All pupils who completed the field trial were randomly assigned to one of three groups:

- Group 1: Paper-based Assessment (PBA) of trend PISA items (23 percent)
- Group 2: Computer-based Assessment (CBA) of trend PISA items (35 percent)
- Group 3: CBA of new PISA science items (42 percent)

[5] While convenience samples were drawn, the field trial did attempt to mirror the population, through use of stratifying variables.

Our focus is groups 1 and 2. These pupils sat *identical* tests consisting of only PISA 'trend' items (questions that have been used in previous PISA cycles and are the basis for linking the PISA test scale over time) with the only difference being assessment mode[6]. This leaves the following working samples:

- Germany = 1,240 pupils (517 paper based and 723 computer based)
- Ireland = 966 pupils (382 paper based and 583 computer based)
- Sweden = 1,232 pupils (515 paper based and 717 computer based)

Pupils within the paper and computer groups were then randomly assigned one of 18 'booklets', each containing four different clusters of test questions. Each cluster contained test questions from only one subject area. Across the three countries combined, 1,149 pupils completed only science and maths questions (those assigned booklets 1 to 6), 1,156 pupils only reading and maths questions (booklets 7 to 12) and 1,133 pupils only reading and science questions (booklets 13 to 18). Moreover, the different subject 'clusters' also contained different test questions (i.e. maths cluster M1 included different questions to maths cluster M3). Sample sizes at the question level are therefore more limited – typically around 200 observations per item per country.

Methodology

The basic feature we exploit is that of a randomised control trial (RCT); pupils have been randomly allocated within schools to either the PBA or CBA group[7]. The PBA and CBA groups should therefore be equivalent in terms of observable and unobservable characteristics. Within the impact evaluation literature, it is considered good practise to test for 'balance' between the randomly assigned groups. We have conducted such balance tests and found that the PBA and CBA groups are generally similar in terms of demographic characteristics[8]. For instance, in both countries the distribution of gender, immigrant status, Special Educational Needs (SEN), school grade, language and mobile phone use are very similar across the PBA and CBA groups. Indeed, the only statistically significant differences are for computer use in Germany and

[6] In contrast, pupils in group 3 were assigned new science questions never administered as part of PISA before, and have therefore been excluded from our analysis.
[7] We have excluded one German school from the analysis, where only computer-based testing was used in the field trial. As randomisation of pupils occurs within schools, this does not pose a threat to the internal validity of our results.
[8] Further information available upon request.

internet use in Sweden and Ireland[9]. Nevertheless, our overall interpretation is that randomisation of pupils to the different test administration modes has functioned adequately.

Next, total reading, science and mathematics scores were derived for all pupils in Groups 1 and 2 in these countries who participated in the field trial. This was done by first converting all test questions into binary format; items were coded as 1 for a correct answer and 0 for an incorrect answer[10]. A one-parameter item-response-theory (IRT) model was then used to derive an overall score for each pupil within each PISA domain. These have then been converted into z-scores; all estimates can therefore be interpreted as effect sizes. The steps outlined above are followed separately for each country and each PISA domain. We experimented with alternative methods of deriving overall test scores (e.g. a simple summative index, a two-parameter IRT model) and have found little substantive change in our results.

In the following section, our analysis will investigate how these overall test scores vary between the 'treatment' (CBA) and 'control' (PBA) groups. As pupils have been randomly assigned, we are able to estimate the causal impact of test administration mode by simply comparing mean PBA and CBA scores. However, we focus upon results from the following OLS regression model, which has slight advantages in statistical power[11].

$$Y_{ij} = \alpha + \beta.M_i + \gamma.D_i + \delta.B_i + \mu_j + \varepsilon_{ij} \qquad (1)$$

Where:

Y = child's test score in a given domain (standardised to mean 0 and standard deviation 1)

M = A dummy variable for assessment mode (0 = PBA; 1 = CBA)

D = A vector of demographic characteristics (e.g. Gender, School Grade)

B = A vector of dummy variables for test booklet (reference = Booklet 1)

μ_j= school level fixed effect

i = Pupil i

j = School j

[9] Pupils answered these questions *after* completing the test. Their responses could therefore be affected by assessment mode e.g. pupils more likely to report internet access and computer use if they had just completed the CBA version of the test.

[10] The few partial credit items were coded as zero for any incorrect or partially correct answer and one for fully correct responses. If a child did not reach or respond to an item, they were awarded a zero for the question.

[11] The results based upon differences in average scores between groups are very similar to those presented, and available upon request.

Throughout our analysis, we take the clustering of pupils within schools into account by making Huber-White adjustments to the standard errors (Huber, 1967; White, 1980).

Heterogeneous effects

The model presented in equation (1) has specified a common treatment effect; the impact of test administration mode will be the same across different groups. Yet, in reality, mode effects may be greater for pupils with certain characteristics (e.g. boys versus girls), or for high versus low ability pupils. We therefore present evidence on possible heterogeneous effects in two ways. First we investigate whether the mode effect differs according to gender, by including an interaction in equation (1) between gender and administration mode[12]. Second, we examine possible heterogeneity in the mode effect across the achievement distribution (i.e. whether the mode effect is bigger for high or low ability pupils) by re-estimating equation (1) using quantile regression[13]. This will reveal whether there are potentially important effects of test administration mode on statistics other than the mean score.

Question-level analysis

Finally, we turn to the impact of test administration mode upon individual test items. Specifically, for each question, we compare percentage correct under the PBA and CBA modes. This may be viewed as an investigation of whether Differential Item Functioning (DIF) is present. Our primary interest is whether a small sub-set of trend questions suffer a particularly large mode effect, or if small mode differences are found across a number of items (resulting in a large cumulative impact). Due to our limited sample size at the test question level, this part of our analysis will be based upon data pooled across the three countries. Hence we have around 700 observations (300 PBA pupils and 400 CBA pupils) per item. As the methodology used to account for mode effects is based upon selecting, adjusting and removing particular questions, it is important that the impact of administration mode upon individual items is explored (further details will be provided in section 4).

[12] It has not been possible to investigate heterogeneity for other groups (e.g. by socio-economic status) as the field trial did not collect this information for both the PBA and CBA groups.

[13] When estimating these quantile regression models, we have also chosen to exclude the control variables (the vectors D, B and μ_j in equation 1). Hence all estimates refer to unconditional differences at each achievement percentile.

3. Results 1: The impact of test administration mode on pupils' scores

Mean scores

Figure 1 documents the impact of administration mode on pupils' average mathematics, reading and science scores. All figures can be interpreted as an 'effect size'. These results show that administration mode has a substantial impact upon pupils' performance, with mean scores from the computer-based tests below those from the paper-based tests. This holds true across countries and each PISA domain. For instance, in mathematics, pupils who took the computer version scored, on average, around 0.10 to 0.20 standard deviations lower than their peers who took the paper test[14]. Similar findings hold for reading, where computer-assessed pupils scored between 0.15 and 0.20 standard deviations lower than the paper assessed pupils in each of the three countries[15]. In terms of cross-national variation, the most pronounced difference is in science, where the negative effect of taking the test on computer is three times larger in Germany than in Sweden (-0.25 versus -0.07)[16]. To put these figures in context, an effect size of 0.2 is roughly equivalent to 20 PISA test points. There is consistent evidence that pupils perform less well on average on the computer versions of the tests. Moreover, although the mode effect is of a similar magnitude across the three countries in reading and mathematics, the same appears unlikely to be true in science.

< Figure 1 >

Heterogeneous effects

Table 1 investigates whether the impact of test mode differs by gender. Overall, the mode effect tends to be similar for boys and girls. For instance, in science (panel C) both German boys and girls who took the computer version of the test scored significantly lower than their peers who took the paper version (-0.32 for boys versus -0.29 for girls). The same is true in Ireland, though with somewhat smaller effects (-0.09 versus -0.16). Indeed, on only two occasions is the gender-by-mode interaction statistically significant. The first is the reading domain in Ireland,

[14] Average PISA maths scores fell by eight point in Germany between 2012 and 2015 and three point in Ireland. In contrast, the average in Sweden rose by 16 points.
[15] Average reading scores were broadly stable in Ireland and Germany between 2012 and 2015. In Sweden they increased by 17 points, taking the average back to the 2009 level. However, in the case of Ireland, overall performance in reading masks changes in the performance of boys and girls, which may be related to mode.
[16] Comparing PISA 2015 to 2012 results, average science scores fell in Ireland by 17 points and Germany by 15 points. In contrast, they rose in Sweden by eight points.

where there is essentially no effect for boys (-0.03), but a negative effect for girls (i.e. girls who took the computer version scored -0.31 standard deviations lower than girls who took the paper version). This is consistent with the results of the main study where the gender difference in Ireland on computer-based reading was among the smallest across all participating countries; the difference in mean scores for girls and boys in Ireland was less than half that of the OECD average (12 points compared to 27 points)[17]. The second is the science domain in Sweden, where boys who took the computer version of the test scoring lower than those who took the paper version (-0.23 effect size), though the same does not hold true for girls (+0.08 effect). Nevertheless, with these exceptions, there is little evidence that assessment mode has a differential impact by gender on the PISA trend items.

< Table 1 >

A similar finding holds with respect to differences for high and low achievers (based upon our quantile regression results). We found no consistent evidence that administration mode has a different impact upon pupils at the top or bottom of the achievement distribution in reading, science or mathematics. Indeed, the results for the three countries tend to be similar, with the exception of science where the mode effect is always estimated to be of greater magnitude in Germany than Sweden (further details on these results available from the authors upon request).

Item-level analysis

Figure 2 compares 'percent correct' statistics for each question across the two administration modes. Each data point refers to an individual question, with PBA results plotted on the y-axis, and CBA plotted on the x-axis. If a point falls on the 45 degree line, then the percentage correct is equal across the PBA and CBA groups. In contrast, points above this line indicate the item was 'harder' (i.e. the percentage correct was lower) when administered on computer. Recall that, due to the limited sample size at the question level, this part of our analysis is based upon data from trend items pooled across the three countries.

<< Figure 2 >>

Two features of Figure 2 stand out. First, across all three domains, most data points sit above the 45 degree line. Reflecting the findings from the previous sub-section, this indicates that

[17] In Ireland, comparing PISA 2015 to 2012 results, a convergence in scores was observed for girls and boys, average reading scores for girls fell in Ireland by 11 points, and average reading scores for boys in Ireland increased by six points.

PISA trend questions are generally harder to answer correctly on computer than paper. Second, the correlation of the percent correct statistics across administration modes is nevertheless high (Pearson's $r \approx 0.95$ in each domain). Consequently, although computer administered questions are harder, the ordering of test-items (in terms of difficulty) remains largely unchanged.

The top half of Table 2 provides further insight into this issue. In particular, it illustrates how around two-thirds to four-fifths of mathematics, reading and science questions are harder to answer using computer rather than paper assessment. Moreover, a statistically significant difference is found for approximately one-in-three items at the ten percent level. This is many more than the one-in-ten expected to occur by 'chance' when using the ten percent significance threshold. In other words, there are clear signs of substantial mode effects for individual test-items.

<< Table 2 >>

The lower half of Table 2 provides a summary of item-level mode effects. Specifically, each item has been placed into one of five categories, based upon the difference in probability between PBA and CBA groups of the question being answered correctly. (See the supplementary material for why these particular thresholds were chosen).

- Negligible (0 to 3 percentage point difference)
- Small (3 to 5 percentage point difference)
- Moderate (5 to 10 percentage point difference)
- Large (10 to 15 percentage point difference)
- Substantial (more than a 15 percentage point difference)

The mode effect for 7 out of 67 maths items (11 percent), 6 out of 90 science items (7 percent) and 7 out of 83 reading items (8 percent) fall into either the 'large' or 'substantial' category. In contrast, 'small' or 'negligible' differences can be observed for around 60 to 70 percent of questions within each of the three domains. There is thus a tentative suggestion that relatively small mode effects occur across a large number of trend items.

4. The methodology used to account for mode effects in PISA 2015

Ensuring comparability between PISA 2015 and scores from previous cycles is the responsibility of ETS. However, the evidence that has been presented in the public domain by the OECD and ETS about the PISA field trial is limited, and does not take into account the

possibility of differential mode effects by country as we do in the present study. Therefore, we now try to provide a clear and transparent discussion of how the OECD and ETS have tried to take mode effects into account.

ETS have described the analyses undertaken using field trial data in Annex A5 and A6 of the PISA 2015 Technical Report (OECD 2016). Critically, this includes how they have tried to model the mode effect to achieve comparability between PISA 2015 and previous cycles. (And between the 58 countries who took the PISA 2015 assessment on computer, and the remainder who completed the test using paper). The rest of this section explains this methodology.

The starting point of the ETS analysis is an IRT equivalent version of our Figure 2. They note that the very strong correlation in such graphics '*shows that the two modes are measuring the same constructs*' and that therefore '*a statistical link can be established such that the CBA and PBA countries' results can be reported on the same scales for 2015'* (OECD 2016).

ETS then consider three different Item-Response Theory (IRT) models that may be used to adjust for mode effects, thus creating a statistical link between paper and computer versions of the PISA test. The first is the 'constant shift' model, which assumes the change of mode has the same impact for all groups (e.g. boys and girls, each different country) and for every question (i.e. all items change in difficulty *by the same amount*). Formally, the probability of a child answering a question, *i*, correctly is given by the following IRT model:

$$P(X = 1|\, \theta, \alpha, \beta_i, \delta) = \frac{\exp(\alpha_i.\theta + \beta_i - \delta)}{1 + \exp(\alpha_i.\theta + \beta_i - \delta)} \qquad \text{(Model 1)}$$

where:

θ = Pupil's latent ability in the subject in question

α_i = item discrimination

β_i = item difficulty

δ = the mode effect (a constant assumed to be the same across every test question)

Returning to our analysis for Germany, Sweden and Ireland, this model is analogous to simply shifting the 45 degree lines in Figure 2 upwards or downwards by a certain amount in order to deal with the mode effect. (Under this adjustment method it is assumed that, were it not for sampling variation, all data points in Figure 2 would then sit upon this shifted line). However, we have already demonstrated in Table 2 and Figure 2 that the size and direction of the mode

effect varies significantly across items. Hence such a simple, homogeneous shift in difficultly does not seem realistic.

Model 2 allows a separate mode effect to be estimated for each test item. This 'heterogeneous shift' model is formally defined:

$$P(X = 1|\, \theta, \alpha, \beta_i, \delta) = \frac{\exp(\alpha_i.\theta + \beta_i - \delta_i)}{1 + \exp(\alpha_i.\theta + \beta_i - \delta_i)} \qquad \text{(Model 2)}$$

The difference between model (2) and model (1) is that a subscript 'i' has now been added to δ (i.e. δ has become δ_i). This indicates that some questions will be harder under CBA, some will be harder under PBA, and for some questions there will be no difference at all. In other words, mode effects are now allowed to differ by question. Furthermore, for some questions it may be reasonable to assume that the mode effect is zero (i.e. there will be some questions where there is little difference in the percentage correct, or IRT parameter estimates, between the PBA and CBA groups). ETS describe model (2) as representing a model of 'weak factorial invariance', which becomes stronger the greater the number of questions it is possible to assume that the mode effect is zero. Importantly, model 2 continues to assume that the size of the mode effect is the same across different groups. This implies that the mode effect for any given question is the same for pupils from low and high income backgrounds, and is the same across countries (for example).

The third model considered by ETS relaxes this assumption by allowing a second latent trait (e.g. computer ability) to also influence the probability of answering each question correctly. Consequently, the size of the mode effect may vary across different types of individual (e.g. those with more or less computer skill):

$$P(X = 1|\, \theta, \alpha, \beta_i, \delta) = \frac{\exp(\alpha_i.\theta + \beta_i - \alpha_{mi}.\pi)}{1 + \exp(\alpha_i.\theta + \beta_i - \alpha_{mi}.\pi)} \qquad \text{(Model 3)}$$

where:

m = the mode of test administration

π = latent ability in computer assessment skills

Of these models, which is the most appropriate to account for mode effects? ETS has made its decision by trading-off parsimony (i.e. simplicity) and how well the model 'fits' the data (and thus complexity). This was done through examination of 'model fit' indices, such as the Akaike Information Criterion (AIC – Akaike, 1974) and Bayesian Information Criterion (BIC –

Schwarz, 1978). AIC and BIC both summarise the relative fit of a set of models (i.e. how well the model fits the observed data) against the number of parameters estimated. Interested readers are directed to Dziak et al (2012) for further discussion.

OECD (2016: Annex A6) presents the AIC and BIC statistics for the three models described above. Model 1 clearly had the weakest fit (largest values of AIC and BIC throughout), suggesting that making a common adjustment across all test items for every participating pupil is insufficient. The second model, where the size of the mode effect is allowed to vary by test question (but is the same for every participating pupil) represented a substantial improvement. Both the AIC and BIC are then marginally lower for model 3 than model 2. However, the difference in these fit criteria between models 2 and 3 is small. Moreover, ETS (2015:44) reports that the correlation between estimated PISA scores from model 2 and model 3 is above 0.99. They thus conclude that '*there is little added utility in using model 3'*, and that model 2 '*describes the data sufficiently well*.' They thus choose model 2 to account for the mode effect. It is hence being assumed that there is no person (or country) specific mode effect – and that the statistical adjustment that needs to be made to the PISA questions is the same across all demographic groups (e.g. boys versus girls; high versus low SES pupils) and across all countries.

Proceeding under the assumption that model 2 fits the observed data sufficiently well, ETS then describe how comparisons will be made between PISA 2015 and previous cycles. The intuition of their approach is as follows. Not all trend questions are subject to mode effects; there will be a subset of items which are not affected by the change of assessment mode. (That is, for these items it can reasonably be assumed in model 2 that $\delta_i = 0$). These items have the property of 'strong measurement invariance', and are the basis for linking computer based PISA 2015 scores with those from previous cycles. In other words, questions thought to be equally difficult across paper and computer tests (based upon evidence from the field trial) form the key link between the two assessment modes (including changes over time). Critically, the more items that have this property of strong measurement invariance (i.e. the more items that seem unaffected by mode) the more reliable these comparisons will be. Overall, ETS have concluded that 61 science items, 51 mathematics items and 65 reading items have this property, with a full list provided in Annex A of OECD (2016). It is hence only this subset of questions,

which appear to function equivalently on paper and computer, that have been used to locate pupils' position performance in 2015 on the PISA 2000-2012 scale [18].

This process does, of course, lead one to consider whether PISA should be 'modelling' rather than 'measuring' children's achievement. By incorporating item-by-mode interactions, one may argue PISA is moving further away from measuring and towards modelling. Such issues have been widely discussed in the educational measurement literature (e.g. Goldstein 1979; Goldstein 1980; Andrich 2004; Panayides, Robinson and Tymms 2010) and continue to be relevant here[19]. Given PISA's main purpose, we believe it should strive to *measure* (rather than model) pupils' achievement, in order for comparisons across groups and across countries to be made in an objective manner. Against the backdrop of the current PISA modelling approach, one may even argue that children in different countries are now taking different tests, given that some continued to use a paper based assessment (see footnote 1). This, in some ways, undercuts a key principle of international comparative studies; children coming from different countries are no longer running the same race (and may also be running a different race to before).

5. Results 2. How do our findings change when only 'strongly invariant' items are considered?

How well is the solution outlined above likely to work? To provide some evidence on this issue, we now replicate the analysis presented in section 3 with one key change – we derive pupils' total test score using only questions where 'strong measurement invariance' is thought by ETS to hold (i.e. questions that are thought to function the same whether they are delivered on paper or computer). In other words, do the results presented in section 3 change when we focus only upon the sub-set of questions that have been used to locate pupils' position on the PISA scale?

In Figure 3 we replicate our analysis of the impact of administration mode upon average scores (previously presented in Figure 1), having removed those questions thought to be influenced by the change of mode. Compared to Figure 1, the estimates are now all lower, with most not reaching statistical significance at conventional levels. Indeed, once we have constructed our

[18] Test questions where $\delta_i \neq 0$ contribute to measurement precision, but do not provide information about the location of pupils on the PISA scale.

[19] Indeed, even more fundamentally, one may question whether any educational (or psychological) assessment can be of the measurement kind (see e.g. Michel, 2008).

scores using 'strongly invariant' questions only, children taking the computer version of the test score only around 0.1 standard deviations lower than the paper-based group in reading and mathematics. However, the biggest remaining issue is in science. Despite now using strongly invariant items only, overall scores on the computer test remain around 0.2 standard deviations below scores on the paper test in Germany. The difference is lower in Ireland (0.1 standard deviations) and Sweden (0.03 standard deviations).

Overall, Figure 3 suggests the method ETS has used to account for mode effects has been beneficial; paper and computer scores are more comparable relative to if no change had been made at all. In mathematics, we have evidence from three countries that the adjustment has worked sufficiently well that any residual impact of mode effects upon average scores is likely to be small[20]. Yet even after restricting the science scale to questions which are meant to be invariant across modes, we continue to see the computer-based group scoring below the paper-based group for the PISA trend items, at least in Ireland and Germany. Consequently, we suggest some caution is needed when interpreting how average science scores have changed in PISA 2015 from previous cycles.

6. Conclusions

PISA is an important international study of school pupils' academic skills. A key change was made in 2015, when most countries moved from paper to computer-based assessment. In this study we have used field trial data for Germany, Sweden and Ireland to investigate the impact changing test modes has upon pupils' responses to trend reading, science and mathematics questions originally designed for administration on paper. The presence of mode effects has been established in all three cognitive domains which, if not accounted for, could limit the comparability of PISA 2015 scores with previous cycles. Consequently, we have also described how the international contractors (ETS) have tried to account for this problem, and documented the evidence they provided to support their approach. We have also investigated how our findings change once we focus only on those test questions thought to function equivalently across different test administration modes. Our interpretation is that the methodology used to produce the PISA 2015 test scores is likely to have reduced the impact of mode effects. However, it also seems unlikely that all the challenges associated with the transition from paper

[20] However, we are unable to comment upon any residual impact upon other key statistics, such as the distribution of performance (standard deviation, percentiles) or co-variation with demographic characteristics such as socio-economic status.

to computer testing have been resolved. Moreover, the issues below stand out as areas in need of further research.

First, our study is based upon data from only three participating countries. Yet more than 50 educational systems made the transition from paper to computer assessment in PISA 2015. Although our estimates of the mode effect are often similar for the three countries for whom we have data available, we are unable to comment upon the extent to which our key findings hold within the other educational systems that participated in the PISA 2015 assessment.

Second, our ability to investigate possible heterogeneous mode effects has been limited by both the field trial sample size and the lack of information regarding pupils' demographic characteristics. For instance, we have been unable to investigate potential interactions with other important variables such as ethnicity and socio-economic status. It is therefore vital that any future data collection gathers more detailed background information on participants so that a more comprehensive analysis of possible heterogeneous mode effects can be undertaken.

Third, likewise, we have been unable to empirically investigate the potential mechanisms that may be causing the mode effects within the PISA study. For instance, as detailed by *author cite*, there are a number of potential causes of mode effects. This includes differences between reading on paper versus on screen, differences in computing skills, test-taking strategies, technical challenges of dealing with computer administration of tests in schools and pupil engagement in the test. All of these factors could be playing some role in driving the mode effects this study has revealed. Unfortunately, the data available do not allow us to consider which of these potential mechanisms are the most important, with this therefore remaining an important direction for future research.

Fourth, this paper has focused upon 'mode effects' for PISA paper-based (trend) items that have been converted into a computer-based format. We have not considered the impact of PISA introducing new interactive questions, which occurred for the science domain only in 2015, with reading and mathematics due to follow in 2018 and 2021 respectively. One may argue that this in itself changes the very construct underpinning PISA and, by 2021, each domain will have a construct reinforced by computer-based skills and technology. While the OECD (2016) frame the concept of scientific literacy as a knowledge of both science and science-based technology, the combination of a switch in mode and the inclusion of new test item types

designed specifically for computer-based delivery, increased the complexity in unpicking PISA 2015 results. Further work on the impact of introducing such questions into PISA is needed. Our advice is therefore that an entire future cycle of PISA should be devoted to a full mode effect study, in order to provide important new evidence on this increasingly important issue.

Fifth, it is not only the PISA 2015 cognitive test that has moved to computer administration, but also the background questionnaire. Yet this aspect of PISA was not included in the field trial mode effects study. Consequently, we do not know how pupils' responses to the background questionnaire may have been influenced by the move from paper to computer administration. This, in turn, brings additional complexity to comparisons of certain important statistics over time (e.g. socio-economic inequality in test scores).

In conclusion, the data and evidence that we have access to suggest that pupils' responses to some of the PISA test have been impacted by the change of administration mode. This is likely to bring important challenges in the short-run and, inevitably, more uncertainty surrounding the comparability of PISA 2015 scores to previous cycles. The good news is that the survey organisers have taken this issue seriously, and have developed a strategy to minimise (though not eliminate) the disruption this is likely to cause.

References

Andrich, D. 2004 "Controversy and the Rasch Model: A Characteristic of Incompatible Paradigms?" Medical Care 42 (1) 17–116.

Akaike, H. 1974. 'A new look at the statistical model identification'. *IEEE Transactions on Automatic Control*, *AC19*(6), 716–723.

Bennett, R.; Braswell, J.; Oranje, A.; Sandene, B.; Kaplan, B. and Yan, F. 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). Retrieved from: https://136.167.2.46/ojs/index.php/jtla/article/view/1639

Cosgrove, J. and Cartwright, F. 2014. 'Changes in achievement on PISA: the case of Ireland and implications for international assessment practice.' *Large Scale Assessments in Education* 2(2): 1-17.

Dziak, J.; Coffman, D.; Lanza, S. and Li, R. 2012. 'Sensitivity and specificity of information criteria.' *The Methodology Center Technical Report Series 12-119*. Accessed 07/12/2014 from http://methodology.psu.edu/media/techreports/12-119.pdf

Educational Testing Service. 2014. 'PISA 2015 field trial analysis report: outcomes of the cognitive assessment. Draft 1.'

Educational Testing Service. 2015. 'PISA 2015 field trial analysis report: outcomes of the cognitive assessment. Draft 2'

Educational Testing Service. 2014b. 'Scaling the cognitive items in the PISA 2015 field trial.' *Presentation to the PISA National Project Managers meeting in Dublin (November 2014).*

Goldstein, H. 1979. 'Consequences of using the Rasch model for educational assessment.' *British Educational Research Journal* 5(2): 211–220.

Goldstein, H. 1980. 'Dimensionality, bias, independence and measurement scale problems in latent trait test score models.' *British Journal of Mathematical and Statistical Psychology* 33(2): 234–246.

Huber, P. 1967. 'The behaviour of maximum likelihood estimates under nonstandard conditions'. Presented at the Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, The Regents of the University of California. Retrieved from http://projecteuclid.org/euclid.bsmsp/1200512988

Ito, K. and Sykes, R. 2004. 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.

Kingston, N. M. 2008. 'Comparability of computer and paper administered multiple-choice tests for K–12 populations. A synthesis.' *Applied Measurement in Education*, *22*(1), 22–37. Retrieved from: https://doi.org/10.1080/08957340802558326

Kroehne, U. and Martens, T. 2011. 'Computer-based competence tests in the national educational panel study: The challenge of mode effects'. *Zeitschrift Für Erziehungswissenschaft*, *14*(2), 169. Retrieved from: https://doi.org/10.1007/s11618-011-0185-4

Michell, J. 2008. 'Is psychometrics pathological science?' *Measurement: Interdisciplinary Research and Perspectives*, *6*(1–2), 7–24. https://doi.org/10.1080/15366360802035489

Organisation for Economic Co-Operation and Development. 2016. *PISA 2015 Technical Report*. Paris: OECD.

Organisation for Economic Co-Operation and Development. 2016. PISA 2015 Assessment and Analytical Framework.

Organisation for Economic Co-Operation and Development. 2010. *PISA Computer-based assessment of student skills in science*. Paris: OECD.

Organisation for Economic Co-Operation and Development. 2011. *Strong performers and successful reformers in education. Lessons from PISA for the United States*. Paris: OECD.

Panayides, P., Robinson, C. and Tymms, P. 2010.' The assessment revolution that has passed England by. Rasch measurement.' *British Educational Research Journal* 36(4): 611–626.

Ryan, C. 2013. 'What is behind the decline in student achievement in Australia?' *Economics of Education Review* 37: 226-39.

Schwarz, G. 1978. 'Estimating the dimension of a model.' *The Annals of Statistics*, *6*(2), 461–464.

Wang, S.; Jiao, H.; Young, M.; Brooks, T. and Olson, J. 2007. 'Comparability of computer-based and paper-and-pencil testing in K12 reading assessments: A meta-analysis of testing mode effects'. *Educational and Psychological Measurement*. Retrieved from: https://doi.org/10.1177/0013164407305592

Wang, S.; Young, M. and Brooks, T. 2004. *Administration mode comparability study for Stanford Diagnostic Reading and Mathematics Tests* (Research Report). San Antonio, TX: Harcourt Assessment.

Wang, S; Jiao, H.; Young, M.; Brooks, T. and Olson, J. 2007. 'A meta-analysis of testing mode effects in grade K-12 mathematics tests.' *Educational and Psychological Measurement* 67: 219

White, H. 1980. 'A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity'. *Econometrica*, *48*(4), 817.

Table 1. Estimates of whether the mode effect differs by gender

(a) Mathematics

	Germany		Ireland		Sweden	
	Mean	**SE**	**Mean**	**SE**	**Mean**	**SE**
Boys paper	0.2	0.12	0.13	0.11	0.13	0.11
Boys computer	-0.03	0.09	-0.08	0.09	0.00	0.08
Mode effect boys	-0.22**	0.09	-0.22*	0.13	-0.13	0.10
Girls paper	0.02	0.12	0.12	0.11	-0.02	0.07
Girls computer	-0.13	0.10	-0.09	0.10	-0.08	0.07
Mode effect girls	-0.15	0.09	-0.21**	0.08	-0.06	0.09
Mode effect * Gender interaction	**-0.08**	**0.10**	**-0.01**	**0.15**	**-0.07**	**0.12**

(b) Reading

	Germany		Ireland		Sweden	
	Mean	**SE**	**Mean**	**SE**	**Mean**	**SE**
Boys paper	0.1	0.13	-0.09	0.12	0	0.11
Boys computer	-0.14	0.1	-0.12	0.12	-0.23	0.09
Mode effect boys	*-0.24***	*0.1*	*-0.03*	*0.13*	-0.22	0.10
Girls paper	0.17	0.12	0.29	0.12	0.18	0.09
Girls computer	-0.05	0.11	-0.02	0.12	0.09	0.06
Mode effect girls	*-0.23***	*0.11*	*-0.31***	*0.11*	-0.08	0.1
Mode effect * Gender interaction	**-0.01**	**0.1**	**0.28****	**0.14**	**-0.14**	**0.13**

(c) Science

	Germany		Ireland		Sweden	
	Mean	**SE**	**Mean**	**SE**	**Mean**	**SE**
Boys paper	0.32	0.12	-0.01	0.11	0.13	0.1
Boys computer	0	0.09	-0.1	0.09	-0.09	0.09
Mode effect boys	*-0.32***	*0.11*	*-0.09*	*0.1*	-0.23	0.13
Girls paper	0.02	0.12	0.16	0.11	-0.05	0.09
Girls computer	-0.26	0.09	0	0.1	0.03	0.08
Mode effect girls	*-0.29***	*0.1*	*-0.16***	*0.07*	0.08	0.09
Mode effect * Gender interaction	**-0.03**	**0.12**	**0.07**	**0.14**	**-0.30****	**0.14**

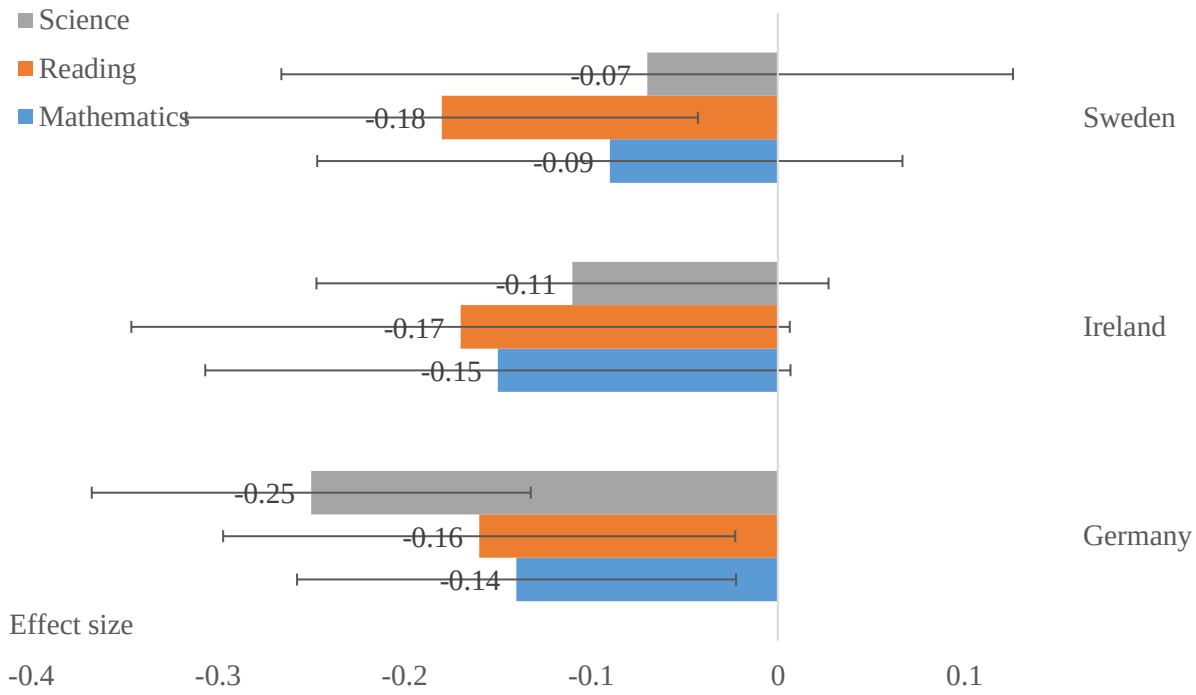
Notes: Authors' calculations using the PISA 2015 field trial dataset. Test scores have been created using all items administered to pupils in each country. Huber-White adjustments have been made to all standard errors. * and ** Indicates a statistically significant interaction at the ten and five percent level.

Table 2. Number of questions where there is a difference in performance across paper and computer assessment modes

	Maths	Reading	Science
CBA perform less well than PBA	46 (69%)	56 (67%)	67 (74%)
Significant at 5 percent level	18 (27%)	22 (27%)	19 (21%)
Significant at 10 percent level	23 (34%)	26 (31%)	22 (24%)
Difference in percent correct between modes			
% 0 to 3 percentage points (negligible)	28 (43%)	39 (48%)	34 (38%)
% 3 to 5 percentage points (small)	13 (20%)	13 (16%)	27 (30%)
% 5 to 10 percentage points (moderate)	17 (26%)	23 (28%)	23 (26%)
% 10 to 15 percentage points (large)	5 (8%)	5 (6%)	6 (7%)
% More than 15 percentage points (substantial)	2 (3%)	2 (2%)	0 (0%)
Total number of items considered	**67**	**83**	**90**

Notes: Authors' calculations using the pooled PISA 2015 field trial database. Analysis based upon all items administered to pupils in all three countries. Figures refer to the number of test questions, with the associated percentages provided in parentheses.

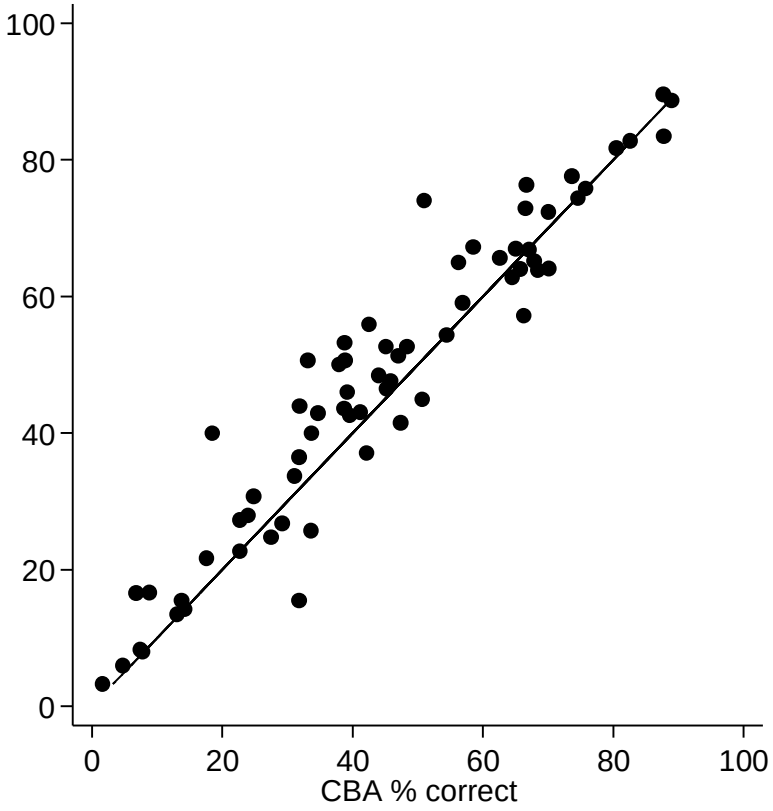
Figure 1. The impact of test administration mode upon pupil's scores



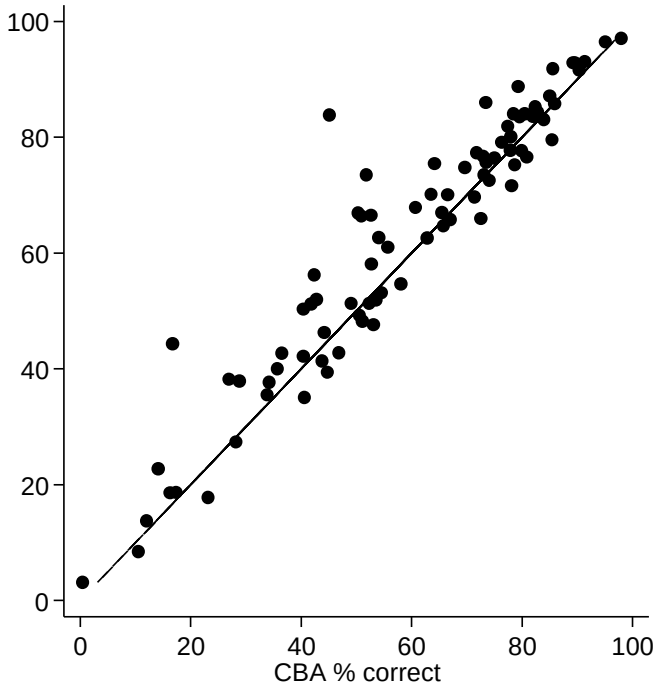
Notes: Authors' calculations using the PISA field trial dataset. Test scores in each domain have been created using all test items administered to pupils in each country. All figures refer to effect sizes. Negative coefficients indicate pupils who completed the computer test obtained lower scores than pupils who completed the paper test. Thin black represents the estimated 95 percent confidence interval. Huber-White adjustments have been made to standard errors. Pupils with a zero score have been excluded.

Figure 2. Percent of correct responses per item by test administration mode

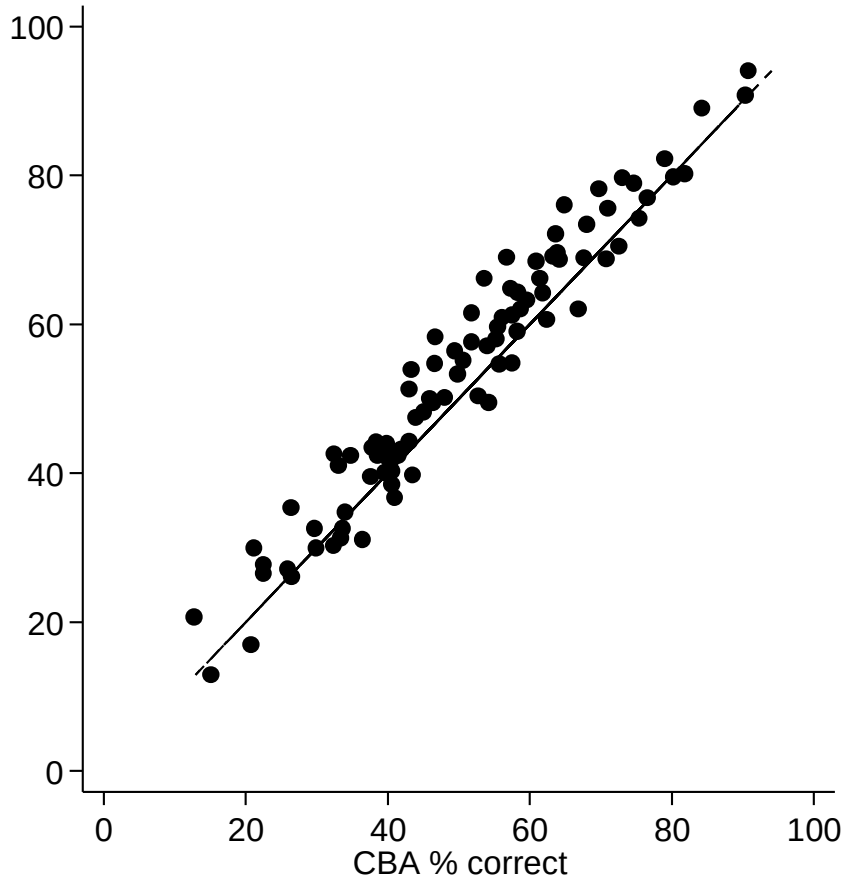
(a) **Maths**



(b) Reading

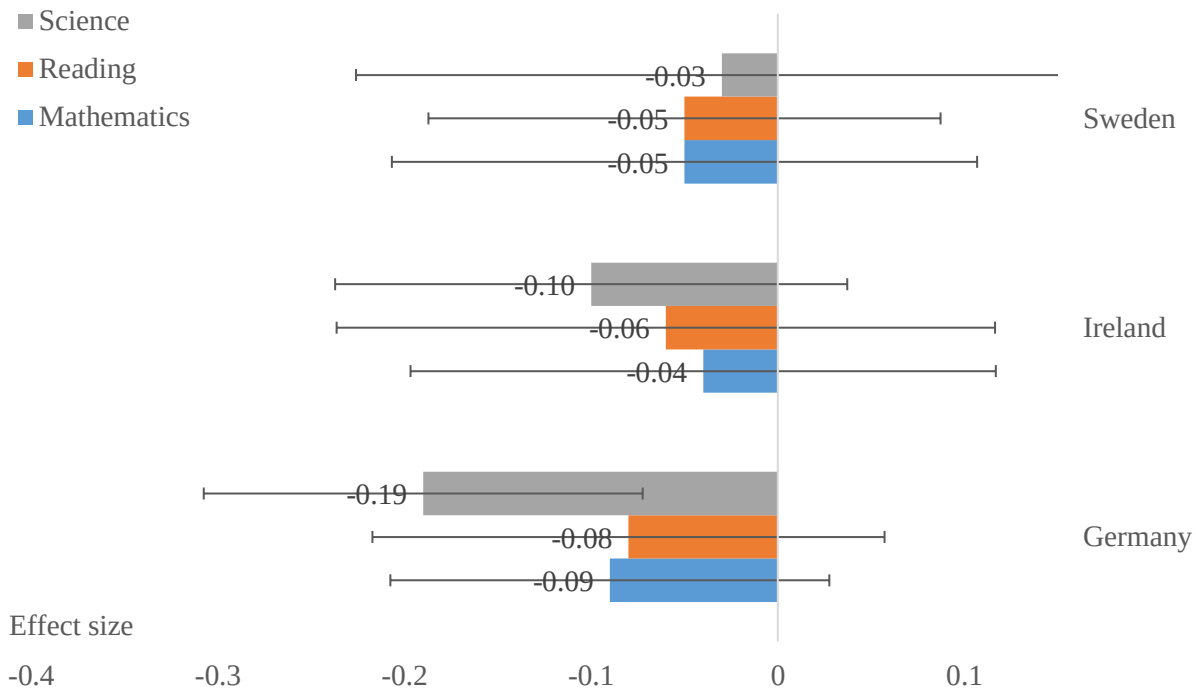


(c) **Science**



Notes: Authors' calculations using the PISA 2015 field trial database. Each data point refers to a test question. Diagonal 45 degree line illustrates where the percentage correct is equal across test administration modes. Points above this line are where pupils sitting the paper version of the question performed better. Pearson correlation equals 0.96 for maths, 0.94 for reading and 0.97 for science.

Figure 3. The estimated impact of test administration mode upon pupil's test scores. Data restricted to strongly invariant items only



Notes: Authors' calculations using PISA field trial. Test scores in each domain have been created using only those items ETS have assumed to function equivalently across paper and computer modes. All figures refer to differences in effect sizes. Negative coefficients indicate pupils who completed the computer test obtained lower scores than pupils who completed the paper test. Thin black line represents the 95 percent confidence interval. Huber-White adjustments have been made to all standard errors. Pupils with a 0 score have been excluded.