CENTRE FOR EDUCATION ECONOMICS

A digital divide? Randomised evidence on the impact of computer-based assessment in PISA

CfEE Research Brief January 2018

Professor John Jerrim

About the Centre for Education Economics

The Centre for Education Economics (CfEE) is an independent think tank working to improve policy and practice in education through impartial economic research.

Founded in 2012, the Centre exists to research and disseminate evidence addressed to how to improve the quality and efficiency of education services; achieve optimal outcomes for young people; and maximise the benefits of education to society as a whole.

CfEE researchers monitor global research output in the economics of education, producing monthly and annual research digests to disseminate this literature. The Centre publishes in-depth policy studies, which frame and inform shorter reports and comment pieces on day-to-day education policy matters.

Working with research and event partners and sponsors across Westminster and central London, CfEE stages a variety of events to engage the public in the policy debate and inform its research.

We hope you'll visit us at one of these and participate in the exchange.

Read more: 'Why education economics?'

About the author

John Jerrim is a Professor of Education and Social Statistics at the UCL Institute of Education. John's research interests include the economics of education, access to higher education, intergenerational mobility, cross-national comparisons and educational inequalities. He has worked extensively with the OECD Programme for International Student Assessment (PISA) data, with this research reported widely in the British media.

About the CfEE Policy Papers

CfEE Policy Papers aim to make complex issues transparent and accessible. Focusing on key policy topics, we take high quality peer reviewed research and present it to our readers in a shorter format with a focus on policy outcomes and recommendations.

All Policy Papers are based on original peer review research. This Paper is based on:

Jerrim, J., Micklewright, J., Heine, J., Salzer, C., and McKeown, C. (forthcoming). *PISA 2015: how big is the 'mode effect' and what has been done about it?*

The information and views set out in this policy paper are those of the author(s) and do not necessarily reflect the official opinion of CfEE.

Contents

Foreword

Data from international large-scale assessments, such as PISA and TIMSS, has enabled policymakers to find new evidence with which to support decisions regarding their national education systems. The fact that the data is publicly available is also a great boon to researchers, since it allows them to rigorously scrutinise official explanations for why pupils perform differently in the tests. In this vein, Cambridge Assessment has previously used PISA data to investigate a variety of important issues, including the relationship between school autonomy/accountability and achievement; the relationship between time in education and achievement; and gender differences in attitudes to learning.

One of the OECD's main goals when creating the PISA survey was to allow performance comparisons between countries and within countries over time. Yet such comparisons may be problematic, for example because of changes to test administration between survey rounds that make scores incomparable. This report focuses on the potential impact of the recent change from paper-based to computer-based assessment that took place in PISA 2015. Analysing data from PISA 2015 field trial – in which pupils were randomly assigned to the different modes – the researchers are able to identify the causal impact of computer-based assessment on pupil performance. It is important to understand mode effects since assessment agencies are currently developing computerised versions of paper tests with the aim to assess different types of knowledge and skills and/or assess the same skills in a new and perhaps more valid way. While the agencies are well aware of the need to understand differences between paper-based and computer-based tests – and therefore carry out their own research in this respect – some may, rightly or wrongly, question their findings on the basis that they have a commercial interest in showing that there are no important differences between the different modes – or that computer-based assessment in some ways is preferable. The authors of this report are entirely independent from such commercial interests, which should make their findings particularly noteworthy to policymakers worldwide.

Tom Bramley

Director, ARD Research Division

Cambridge Assessment

Executive summary

- Since the Programme for International Student Assessment (PISA) first was carried out in 2000, it has increasingly come to dominate education-policy discussions worldwide. Educationalists and policymakers eagerly await the tri-annual results, with particular interest in whether their country has moved up or slid down the rankings.

- Yet there are many challenges to measuring trends using large-scale international assessments, such as PISA. This is because administration and analysis procedures may change between survey rounds, potentially influencing results. In this paper, we study the most important alteration: the move to computer-based assessment in 2015.

- Between 2000 and 2012, PISA was carried out as a regular paper-based assessment. However, in 2015, pupils in the great majority of countries instead took the test on a computer. Since the change to computer-based assessment could affect pupil performance by itself – in ways that differ between countries – it has the potential to reduce comparability of PISA test scores across countries and over time.

- We investigate this issue using data from the OECD field trial, which was carried out in the spring of 2014 in all countries making the switch from paper-based to computer-based assessment. Since pupils taking part in the field trial were randomly assigned to complete the same PISA questions on a computer or using paper and pen, we are able to draw causal inferences. We have access to data from three countries: Germany, Sweden and Ireland.

- The results show that pupils completing the computer-based test performed substantially worse than pupils completing the paper-based test in all three countries. The difference is most pronounced in Germany (up to 26 PISA points), followed by Ireland (up to 18 PISA points) and Sweden (up to 15 PISA points). Also, we find little evidence of systematic gender differences in the impact of computers.

- Once we apply the method used to account for mode effects in PISA 2015, the differences decrease in all three countries. However, there is important heterogeneity in this respect. Whereas no statistically significant differences in performance remain in Sweden, pupils sitting the computer-based test in Ireland and Germany still perform 11 and 15 points lower in science respectively.

- Our key conclusion is the adjustment made in PISA 2015 does not overcome all the potential challenges of switching to computer-based tests, but that it represents an improvement compared with not making any adjustment at all.

- The results show that policymakers should take great care when comparing the results across and within countries obtained through different modes.

1. Introduction

The Programme for International Student Assessment (PISA) is a major cross-national study of 15-year-olds' academic skills, which has increasingly come to dominate education-policy discussions worldwide since it was first carried out in 2000. In recent years, policymakers have come to benchmark the success or otherwise of their policies by the changes in PISA performance over time. Australia, England, Finland, Ireland, and Sweden are prominent examples of countries where declining PISA scores have received much attention (Ryan 2013; Jerrim 2013). Conversely, countries such as Germany and Poland have been held up as 'successful reformers' due to their improvements in PISA (OECD 2011).

Yet there are many challenges to measuring trends using large-scale international assessments such as PISA. For example, administration and analysis procedures can change between survey rounds, which could potentially contaminate the results and make them difficult to interpret (see Cosgrove and Cartwright 2014; Jerrim 2013). It is therefore important to study to what extent such changes affect the headline results in the survey.

In this paper, we study a major change to the administration of PISA 2015: the move from paper-based assessment to computer-based assessment.[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, and enabling greater insights into test-taking behaviour.

With rapid technological innovation, it appears inevitable that computer-based assessment would be introduced in PISA at some point. Yet, in the short-term, the change poses challenges, including the potential for so-called 'mode effects' to influence the comparability of PISA scores over time. 'Mode effects' refer to whether questions designed to be delivered on paper are systematically easier or harder when delivered on computer. Moreover, such effects may differ between countries or groups, for example by gender or socio-economic status.

Previous research finds that the direction and strength of mode effects depend on different factors, including subject area, study design, and question response format (Bennett et al. 2008; Kingston 2009; Wang et al. 2008). Consequently, it is important to empirically verify the presence or otherwise of mode effects in each test separately (Kroehne and Martens 2011).

[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 Vietnam.

Since the PISA 2015 scores were released in December 2016, there has been considerable speculation about the possible impact of introducing computers (see Ward 2017). The average OECD score in PISA 2015 was around eight points lower in science than in 2012 – and 11 of the top 30 countries in 2012 saw a significant decline in achievement in 2015, while just one saw a significant improvement. For example, Hong Kong fell fully 32 PISA points in science between 2012 and 2015, a decline that at face value would represent learning worth about one full school year. While it is technically possible that such extreme changes could occur in such a short period of time, it is also plausible that mode effects have at least contributed to them.

To study potential mode effects, we use data from the PISA field trial in Germany, Ireland, and Sweden. In the field trial, pupils were randomly assigned to complete the same PISA questions on a computer or using paper and pen. This means that any difference between the pupils can be causally attributed to mode effects alone.[2] We also examine the methodology used in PISA 2015 to account for potential mode effects – and whether or not the adjustment made has worked. Our analysis includes each core PISA domain (reading, maths and science), along with a question-by-question assessment of each trend item.

The countries under investigation represent interesting case studies for the purposes of analysing mode effects, as all countries saw changes in their performance between 2012 and 2015. While science scores decreased by 15 points in Germany and Ireland, they improved by 10 points in Sweden. Analysing the experimental data at hand enables us to study to what extent these changes may have been due to the change in mode rather than bona fide changes in cognitive performance.

We find consistent evidence of mode effects and that these are of greater magnitudes than typically reported elsewhere in the literature. The results show that pupils completing the computer-based test performed worse than pupils completing the paper-based test in all three countries. The difference is most pronounced in Germany (up to 26 PISA points in science), followed by Ireland (up to 18 PISA points in reading) and Sweden (up to 15 PISA points in reading).

Once we apply the method used to account for mode effects in PISA 2015, the difference decreases in all three countries. However, there is heterogeneity in this respect. Whereas no statistically significant differences in performance remain in Sweden, pupils sitting the computer-based test in Ireland and Germany still perform 11 and 15 points lower in science respectively. Also, while not

[2] In a previous analysis of mode effects in PISA 2012, Jerrim (2016) compared pupil performance in the paper-based version with performance in a special computer-based version. He found strong evidence of mode effects that differed across countries, but the findings were limited due to the study design and the fact that different questions were used across the paper and computer assessments.

statistically significant, the differences in reading and mathematics indicate that some differences remain also in these subjects, especially in Germany.

Overall, therefore, the adjustment made in PISA 2015 certainly represents an improvement to not making any adjustment at all, but it does not overcome all the potential challenges of switching to computer-based tests. Policymakers should therefore take great care when comparing the results across and within countries obtained through different modes. More generally, our findings clearly highlight the importance of mode effects in education. Further research is necessary to analyse to what extent computer-based or paper-based assessments have greater predictive value for longer-term outcomes.

2. Data

We use data from the PISA 2015 field trial, which was conducted in the spring of 2014. All countries making the switch from paper to computer assessment took part. The design of the field trial was led by the Educational Testing Service (ETS), a global assessment provider. Each participating country was asked to recruit pupils and schools using one of three 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

With respect to our countries, Ireland followed design A, Germany design C, and Sweden design C. However, due to pupil exemptions, logistical restrictions, and challenges with recruitment, these criteria were not always met. The final 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

There was no requirement for these schools to be randomly sampled, so we cannot be sure that the sample is representative of the countries' pupil population. However, all pupils who completed the field trial were randomly assigned within schools to one of three groups:

- Group 1: Paper-based assessment of trend PISA items (23 per cent)
- Group 2: Computer-based assessment of trend PISA items (35 per cent)
- Group 3: Computer-based assessment of new PISA science items (42 per cent)

In this paper, we focus on groups 1 and 2. These pupils sat identical tests consisting of only questions that have been used in previous PISA cycles, which are the basis for linking the PISA test scale over time – with the only difference being assessment mode. This leaves us with the following working samples:

- Germany = 1,240 pupils (517 paper based and 723 computer based)[3]
- Ireland = 966 pupils (382 paper based and 583 computer based)

[3] 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.

- Sweden = 1,232 pupils (515 paper based and 717 computer based)[4]

3. Methodology

The basic feature we exploit is that of a randomised control trial. As noted in Section 2, pupils were randomly allocated within schools to either the paper-based or computer-based assessment, which means that the groups should be equivalent in terms of observable and unobservable characteristics. We tested this by conducting balance tests on background characteristics, the results of which are reported in Table A1, and found that the treatment and control groups are indeed very similar: the distributions of boys, pupils with an immigration and language background, and pupils with special educational needs mostly do not differ with a statistically significant margin in the different groups. Our overall interpretation is therefore that randomisation of pupils worked adequately.

Next, we derive total reading, science, and mathematics scores for all pupils. 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.[5] We then derived an overall score for each pupil within each PISA domain.

We use these scores to study whether achievement differs between the 'treatment' (computer-based assessment) group and the 'control' (paper-based assessment) group. As pupils have been randomly assigned to each of these groups, we estimate the causal impact of test administration mode by simply comparing the mean scores in the different groups[6]. Jerrim et al (forthcoming) also illustrate very similar results when an OLS regression model is used instead. We also use a regression model to present evidence to what extent the mode effect differs depending on gender in Section 4.1, by including an interaction between the gender and treatment indicators.

3.1. Question-level analysis

As noted in Section 4, the methodology used to account for mode effects in PISA is based upon selecting, adjusting and removing particular questions – and it is therefore important to explore the impact of administration mode upon individual items. To do so, we simply compare percentage correct answers under each mode. Our primary interest is whether there is particularly large mode

[4] Pupils within the paper and computer groups were 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, 1,156 pupils only reading and maths questions, and 1,133 pupils only reading and science questions. Moreover, the different subject 'clusters' also contained different test questions. Sample sizes at the question level are therefore more limited – typically around 200 observations per item per country.

[5] 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.

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

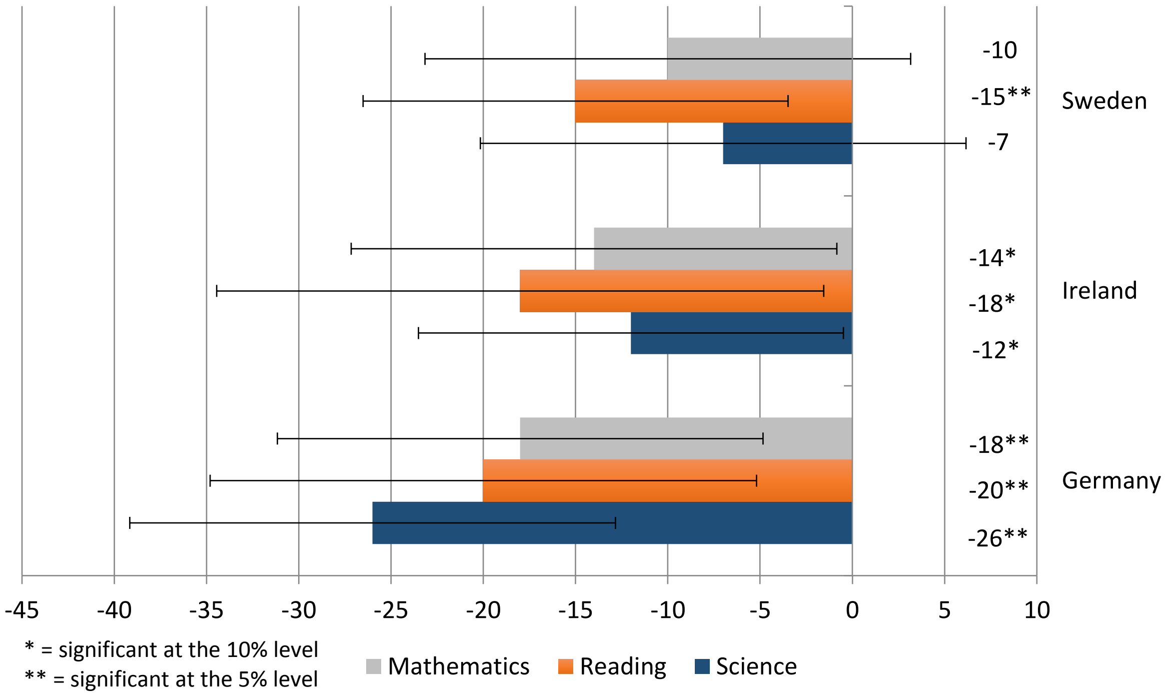
effect present in a small sub-set of questions, or if there are smaller mode differences across a number of items. Due to our limited sample size at the test question level, this part of our analysis is carried out using data pooled across the three countries. We therefore have around 700 observations per item, with 300 pupils who took the paper-based assessment and 400 pupils who took the computer-based assessment.

4. Results

Figure 1 documents the impact of administration mode on pupils' average mathematics, reading and science scores. All differences are reported as PISA scores.[7] The line across the columns represents a 90% confidence interval; if the line crosses 0, the difference is not statistically significant at the 10% level.

The results show that computer-based assessment has a substantial negative 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, although the estimates are overall more precise in Ireland and Germany than in Sweden. For instance, in mathematics, pupils who sat the computer version scored, on average, 10-18 PISA points lower than their peers who took the paper test. Similar findings hold for reading, where computer-assessed pupils scored 15-20 PISA points lower than the paper assessed pupils in each of the three countries. In terms of cross-national variation, the most pronounced difference is in science, where the negative effect of taking the test on a computer is more than three times larger in Germany than in Sweden (-26 PISA points versus -7 PISA points).

Figure 1. The impact of computer-based assessment on pupil scores (all items in the field trial)



[7] Technically, the results are reported in z-scores, which we multiply by 100. This makes the estimates roughly equivalent to PISA points (since 100 points are equivalent to one international standard deviation).

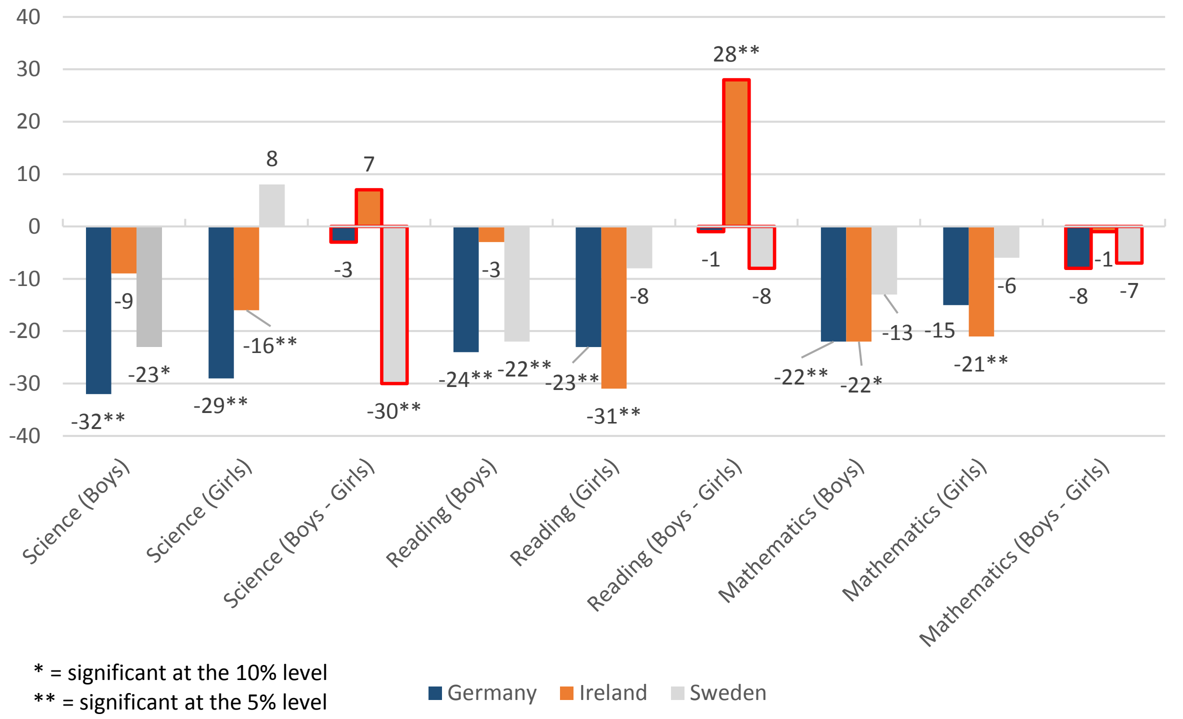
Since 30 PISA points are roughly equivalent to one school year of learning, the estimates indicate that sitting the computer-based rather than the paper-based assessment decreases performance by the equivalent of 23-87% of one school year, depending on country and subject. This is a considerable impact indeed.

Overall, therefore, there is consistent evidence that pupils perform considerably less well on average on the computer versions of the PISA test, as administered in the field trial. Moreover, although the mode effect is broadly speaking of a similar magnitude across the three countries in reading and mathematics, it differs more markedly in science.

4.1. Does the effect differ by gender?

Figure 2 investigates whether the impact of test mode differs by gender. Each bar represent the effect of taking the test on a computer, and bars with red borders indicate the difference between boys and girls in each domain. Overall, the mode effect tends to be similar for boys and girls. For instance, in science, both German boys and girls who took the computer version of the test scored significantly lower than their peers who took the paper version (-32 points for boys versus -29 points for girls). The same is true in Ireland, though with somewhat smaller effects (-9 points versus -16 points).

Figure 2. The impact of computer-based assessment on pupil scores (by gender)



* = significant at the 10% level
** = significant at the 5% level

In fact, there are only instances when the results indicate statistically significant differences between how boys and girls are affected by assessment mode. The first is in the reading domain in Ireland, where there is essentially no effect for boys (-3 points), but a large negative effect for girls (-31 points).[8] The second is in the science domain in Sweden, where boys who took the computer version of the test scored lower than those who took the paper version (-23 points), though the same does not hold true for girls (+8 points). Nevertheless, overall, there is little evidence that assessment mode has a differential impact by gender on the PISA trend items.[9]

4.2. Item-level analysis

Having established that there are substantial mode effects, we turn to the analysis of how answers to individual items are affected. The top half Table 1 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 10% 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 1: Number of questions where there is a difference in performance across modes

	Maths	Reading	Science
Negative impact of computers	46 (69%)	56 (67%)	67 (74%)
Significant at 5% level	18 (27%)	22 (27%)	19 (21%)
Significant at 10% level	23 (34%)	26 (31%)	22 (24%)
Difference in per cent 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**

Note: Authors' calculations using the pooled PISA 2015 field trial data from Germany, Ireland, and Sweden.

[8] 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.

[9] A similar finding holds with respect to differences for high and low achievers. 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. These results also tended to be similar across the three countries, with the exception of science where the mode effect is always estimated to be of greater magnitude in Germany than Sweden.

The lower half of Table 1 in turn provides a summary of item-level mode effects. Specifically, each item has been placed into one of five categories, based on the difference in probability of the question being answered correctly using the different modes. The mode effect for 7 out of 67 maths items (11%), 6 out of 90 science items (7%) and 7 out of 83 reading items (8%) fall into either the 'large' or 'substantial' category. In contrast, 'small' or 'negligible' differences can be observed for around 60-70% of questions within each of the three domains. This indicates that relatively small mode effects occur across a large number of trend items – which accumulate and generate the large effect sizes displayed in the figures above.

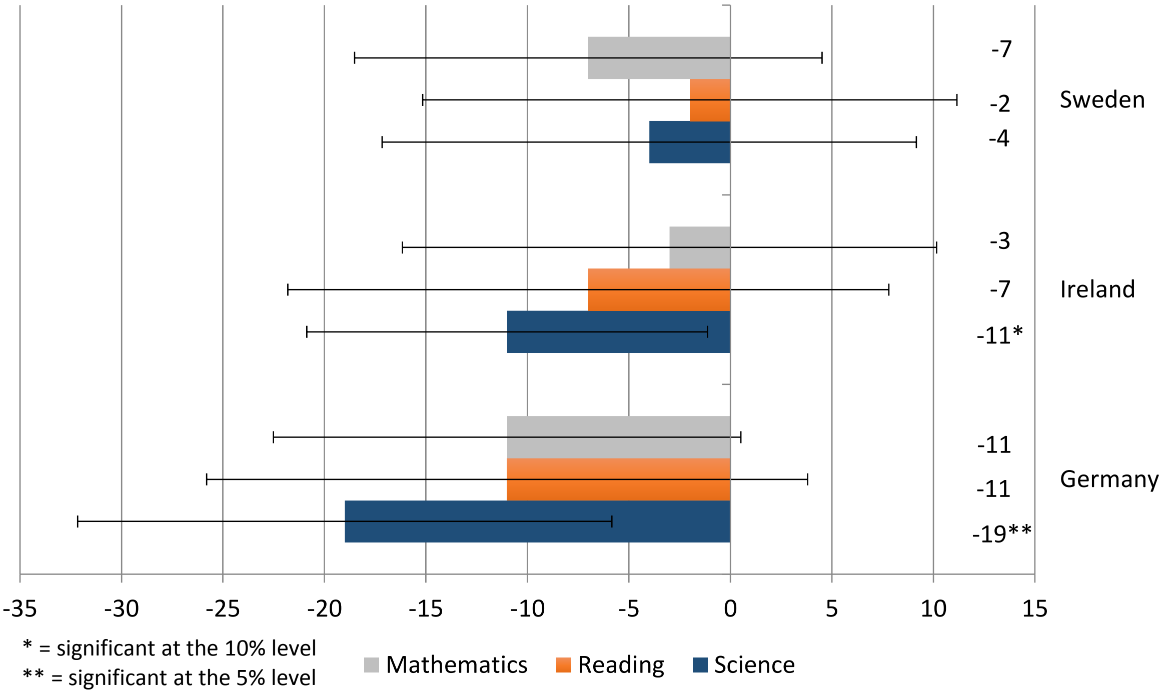
4.3. Effects when we only study items deemed equivalent across modes

Our results lead to an obvious question: are the PISA 2015 results comparable with previous rounds? To take into account possible mode effects when studying changes in performance over time, ETS made various adjustments to the scores based on the results from the field trial. We have described in detail the technical procedure elsewhere (Jerrim et al. forthcoming). Here, we only describe the intuition of their approach: the idea behind the adjustments is that a subset of items is not affected by assessment mode – and these items are the basis for linking computer-based PISA 2015 scores with those from previous cycles. In other words, questions deemed to be equally difficult across paper and computer tests – based on evidence from the field trial – form the key link between the two assessment modes. Overall, ETS 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 (2016a). Consequently, only this subset of questions has been used to create countries' average scores in PISA 2015.

How well, then, is the ETS's solution likely to work? First, we note that the adjustment assumes that the mode effect applies equally to pupils across all countries, which our findings indicate does not hold true. To study whether or not this is likely to pose a problem, we replicate our analysis in the previous sections but make one key change: we derive pupils' total test score using only questions that were deemed by the ETS to be equivalent on the computer-based and paper-based assessments. Doing so enables us to study whether or not the results change when we only study the sub-set of questions that were deemed to be equal across modes.

Figure 3 reports the results after removing questions deemed to be influenced by the change of mode. Compared to Figure 1, the differences are in fact lower. Indeed, pupils sitting the test on a computer now score 2-11 PISA points lower than the paper-based group in reading and mathematics – and these differences are not statistically significant. In mathematics and reading, the evidence therefore indicates that the adjustment applied in PISA 2015 worked reasonably well: countries' average scores are at least not likely to be strongly affected by mode effects.[10] Still, while not statistically significant, the maximum observed difference in these subjects in each country (7-11 points) still represents 23-33% of one school year worth of learning. This is not huge – but certainly not small either. In other words, while the mode effect in mathematics and reading is unlikely to be very strong, it may still be meaningful.

Figure 3. The impact of computer-based assessment on pupil scores (items deemed equivalent across modes in PISA 2015)



Furthermore, even after restricting the science scale to questions that are meant to be equivalent across modes, pupils taking the computer-based test continue to score below their peers who sat the paper-based test, at least in Ireland and Germany: the impact of taking the computer-based test in Ireland is -11 points and in Germany -15 points. The difference in Sweden is only -4 points and not statistically significant. Even after the adjustments made in PISA, there is therefore a moderate-to-large mode effect present in Germany, a small-to-moderate effect in Ireland, and a negligent-to-small effect in Sweden.

[10] 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.

Overall, the results therefore indicate that the ETS's method to account for mode effects in PISA 2015 has been beneficial: paper and computer scores are more comparable once we use the same method to adjust for mode effects in the field trial. Yet the differences do not disappear entirely and are still likely to be meaningful. This is especially true in science, where we find strong mode effects in Germany, which are about twice as big as the ones found for Ireland – which in turn is almost three times as big as the statistically insignificant impact found for Sweden. Extra caution is therefore needed when interpreting how average science scores changed in PISA 2015 compared with previous rounds.

5. Conclusions

Since its inception, PISA has grown to become the most important international survey of pupils' knowledge and skills, with policymakers and the media giving considerable attention to how their countries perform as well as trends in this respect. However, comparing changes over time in international surveys could be problematic, if changes to the administration of the test affect performance by themselves. If so, it is impossible to compare 'like for like'.

In this paper, we have studied a key change made to the PISA test: most countries moved from paper-based to computer-based assessment. We analysed field-trial data for Germany, Sweden and Ireland to investigate the impact of this change on pupil performance. Since pupils taking part in the field trial were randomly assigned to complete the same PISA questions using either paper-based on computer-based assessment, we are able to draw causal inferences in this respect.

We found that pupils completing the computer-based assessment perform substantially worse than pupils completing the paper-based assessment in all three countries. The difference is most pronounced in Germany (up to 26 PISA points), followed by Ireland (up to 18 PISA points) and Sweden (up to 15 PISA points). However, we found little evidence of systematic gender differences in the impact of computers.

Once we applied the method that was used to account for mode effects in PISA 2015, the differences decrease in all three countries. However, there is heterogeneity in this respect. Whereas no statistically significant differences in performance remain in Sweden, pupils sitting the computer-based test in Ireland and Germany still perform 11 and 15 points lower in science respectively.

Our key conclusion is that the adjustment made in PISA 2015 does not overcome all potential challenges of switching to computer-based tests, but that it nevertheless represents an improvement compared with not making any adjustment at all. While more research is needed to establish to what extent our results are relevant for other countries than the three we have analysed here, the results indicate that policymakers should take great care when comparing the results across countries and within countries obtained through different modes.

Certainly, unlike the PISA 2015 sampling process, it is important to note that the field trial sample was not recruited in a way that ensures representativeness of the entire pupil populations. Although pupils from a range of different schools and locations participated, it is difficult to make strong inferences about the impact of computer-based assessment the wider pupil populations. Of course, this problem applies as much to the official method to account for mode effects in the PISA 2015 survey as to our study. While it is too late to deal with this issue for the purposes of PISA 2015, it is important to take this into consideration in the design of future rounds.

Furthermore, due to data availability, we could only study heterogeneity in mode effects by gender. Similarly, the data did not permit analyses of potential mechanisms that may cause the mode effects displayed, such as differences between how pupils read on paper versus on screen, differences in computing skills, test-taking strategies, and pupil engagement. It is important that future data collection gathers more detailed information on participants to enable more in-depth analyses of heterogeneity in mode effects as well as mechanisms behind them.

Of course, it is also important to note that our analysis only focused on mode effects for originally paper-based questions that have been converted into a computer-based format. While the change in mode did apply to the background and school questionnaires, the field trial did not seek to establish to what extent the new mode affected responses in this respect. This, in turn, complicates comparisons of trends in important statistics over time, such as the impact of socio-economic background on PISA performance, which hinge on comparable responses to the questionnaires over time.

Finally, our study has not considered the impact of new interactive PISA questions, which were introduced in the science domain in 2015 and will be introduced in reading and mathematics in 2018 and 2021 respectively. This may in fact change the construct underpinning PISA fundamentally, as this then will be reinforced by computer-based skills and technology. While the OECD (2016b) 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 questions designed specifically for computer-based delivery makes it even more difficult to interpret the results from PISA 2015. We therefore suggest that an entire future cycle of PISA should be devoted to a full mode-effect study, in order to provide evidence in this respect.

References

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.

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.

Jerrim, J. 2013. 'The reliability of trends over time in international education test scores: is the performance of England's secondary school pupils really in relative decline?' *Journal of Social Policy* 42(2): 259–79.

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., McKeown, C. Forthcoming. 'PISA 2015: how big is the 'mode effect' and what has been done about it?' *Oxford Review of Education*.

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.

OECD. 2016a. *PISA 2015 Technical Report*. Paris: OECD.

OECD. 2016b. PISA 2015 Assessment and Analytical Framework.

OECD. 2011. *Strong performers and successful reformers in education. Lessons from PISA for the United States*. Paris: OECD.

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

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

Ward, H. 'Pisa data may be incomparable, Schleicher admits.' TES. 2017. 24 March. https://www.tes.com/news/school-news/breaking-news/exclusive-pisa-data-may-be-incomparable-schleicher-admits.

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

Appendix

Table A1. Balancing tests between treatment and control groups

	Germany		Ireland		Sweden	
	Computer	Paper	Computer	Paper	Computer	Paper
Male	51%	53%	47%	50%	49%	48%
Immigrant	5%	4%	15%	12%	8%	10%
Mother immigrant	18%	17%	19%	16%	21%	18%
Father immigrant	18%	18%	17%	23%	**22%**	**16%**
Special educational need	2%	2%	1%	2%	7%	6%
Foreign language	7%	7%	7%	7%	6%	7%

Note: Authors' calculations. Figures reported for all observations where data available. Figures in bold refer to differences that are statistically significant at the 5% level.