ANNEX A6
THE PISA 2015 FIELD TRIAL MODE-EFFECT STUDY

This document presents the analyses that were undertaken to ensure the comparability of test results across paper and computer modes. While this "mode-effect analysis" is important, in light of the significant change introduced in the administration of the PISA test, it was not the only goal for the PISA field trial. The PISA 2015 field trial had, in fact, four major goals:

1. to elicit information about the quantity of data obtained and about survey operations, prior to the main survey

2. to test the computer-delivery platform and the related operations

3. to assess the quality of the items that were either newly developed for computer-based delivery or adapted from earlier cycles for both paper- and computer-based delivery

4. to select the most appropriate item response theory (IRT) models to establish reliable valid, and comparable scales.

The design of the field trial thus had to balance the need to gather sufficient information about the newly developed items, in order to validate their quality, at the international level and across all languages used in PISA, with the need to ensure the comparability of scales based on existing items, across paper and computer modes.

Overall, 58 national centres that were planning to assess students on computers in the main survey (CBA countries) conducted the field trial using both paper and computers, while 16 countries/economies conducted the field trial as a paper-based assessment (PBA).[1]

The field trial design for CBA countries depended on a random assignment of students to three groups:

- Group 1: paper-based administration (PBA) of trend items

- Group 2: computer-based administration (CBA) of trend items

- Group 3: computer-based administration of new items for science and collaborative problem solving (CPS).

The design was based on a targeted national sample size of 1950 students from 25 to 30 schools (a larger sample was required for countries that participated in the financial literacy option and for countries with multiple testing languages; more schools were required if a within-school sample of 78 students could not be achieved). Students were split into the three groups within each participating school. All but three countries/economies met the sample size requirement (achieving at least 95% of the target).

The random assignment of students was managed by KeyQuest software and required a distribution of 23%, 35% and 42% for Groups 1, 2 and 3, respectively. The design assigned a smaller

percentage of students to Group 1, because the items administered to that group were unchanged from those used in past PISA rounds and did not, therefore, require thorough quality checks. It was expected that students' answers from Group 1 would reflect the typical patterns observed in past PISA rounds and could be combined with the historical PISA database for the purpose of mode-effect analyses. Across all countries, random assignment worked as designed with the following percentages observed: 23%, 33% and 43% for Groups 1, 2 and 3, respectively. Individual countries showed little deviation from these design requirements, but occasionally, some schools did not administer the test in both modes.

Comparability of PBA and CBA groups in the field trial

Comparability of groups 1, 2 and 3, created by random assignment, was established based on a number of background variables. A chi-squared test for independence to detect systematic differences for groups of students was conducted on a small subset of variables that were available in the context questionnaires administered on the paper and computer. Due to the rotation design used for field-testing a large number of background variables in the computer-based context questionnaire, however, a number of questions that were given in both modes have data on only one in four of the CBA cases. In addition, another set of questions was based on self-reports that relate to computer and Internet use. They asked about how often technology is used by the student, and hence are based on subjective judgement. The context questionnaire was distributed after students answered two hours' worth of computer-based tasks, which, in light of the result below, may have had a priming effect. This leaves a relatively small number of questions that are suitable for the comparison of CBA and PBA samples that were used to calculate tests of independence of these variables from administration mode: grade, gender, and country of birth for student and parents.

Table A6.1 shows the results from the Chi-square tests, which provide an indication of how often the null hypothesis of independence of these variables from the administration mode was rejected. The analysis is based on a total of 42 national centres that conducted the test in both modes.

Table A6.1. Comparability of groups based on random assignment of students to forms

Sample size for Groups 2 and 3 (CBA), as a proportion of group size	Variable	Samples (language groups within countries/economies) with significant Chi-square tests
1	Gender	1/57
1	Grade	1/57
¼	Frequency of Internet use	20/57
¼	Frequency of computer use	9/57
¼	Frequency of smartphone use	20/57
¼	Immigrant-Self	3/57
¼	Immigrant-Mother	3/57
¼	Immigrant-Father	4/57
¼	Language spoken at home	8/57

Note: Each row contains 57 (language groups-based) comparisons, resulting in a sum of 9*57=513 comparisons, which would lead to roughly 26 randomly significant comparisons at the 5% error level. Note that only grade and gender involved all cases, while all other comparisons of context questionnaire variables are based on 25% of the CBA cases only.

Source: OECD, PISA 2015 field-trial database.

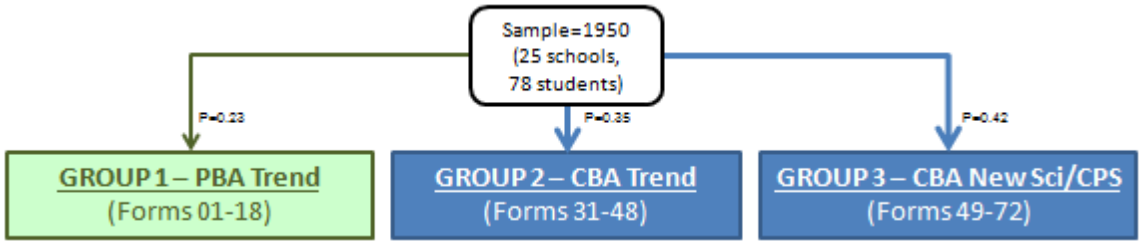
The large number of samples in which it appears that students taking the CBA reported differently on technology use, such as use of the Internet or a smartphone, may be more of a priming effect than truly based on non-random assignment. It appears unlikely that such a large number of national centres assigned students based on computer use or related variables instead of adhering to random assignment.

The PISA test for the mode-effect study: Parallel booklets for PBA and CBA groups

In order to examine the transition to computer delivery, a set of 18 paper-based forms covering the domains of science, reading and mathematics was constructed for Group 1, including only trend items administered in the past PISA surveys (Forms 01-18 in Figure A6.1). A set of tasks "identical" to those contained in the 18 paper-based forms was then adapted and authored for computer administration, yielding 18 equivalent computer-based test forms (these booklets are not identical to those used for the main study) (Forms 31-48). In addition, there were 12 computer-based test forms consisting of the new science tasks developed for administration in the 2015 survey and 12 new test forms combining those 2015 science tasks with the new collaborative problem-solving tasks (Forms 49-72).

Figure A6.1 summarises the main elements of the field-trial design.

Figure A6.1. The random-assignment design for the field trial



Functioning of the computer-based platform

During the field trial, the computer-based platform successfully delivered, captured and exported information for the vast majority of items. Less than 2% of the items had errors; some of these could be fixed for the main study, otherwise, the items were dropped. These items were excluded from subsequent field trial analyses.

Comparability of the computer- and paper-based items: Classical test theory analyses

The comparability of computer- and paper-based items was established using both analyses based on classical test theory (CTT) and item-response theory (IRT).

Item difficulties (proportion correct, P+), frequencies of scores (number of responses attempted, correct and incorrect responses, omitted items, not-reached items), cluster score, point biserial correlations, and (for CBA groups only) response time information within each domain per item and item cluster were examined in the PISA 2015 field trial. Proportion correct and missing rates were compared to results from all prior PISA cycles. This was done separately for data from the PBA and the CBA and examined at an aggregate level across countries. These analyses were also performed separately for each country in order to identify outliers (single items that seem to work differently across assessment cycles and countries).

These analyses highlighted that there were, in general, fewer omitted responses in CBA mode than in PBA mode, and reduced position effects in CBA (i.e. percent-correct measures tended to be more similar across all four cluster positions, compared to PBA). The analyses also indicated that the proportion of correct responses was, in general, similar across PBA and CBA groups. More rigorous analyses of item invariance were conducted using IRT tools.

Comparability of the computer- and paper-based items: Item-response theory analyses

Selecting an IRT model for PISA data

PISA has collected data in representative samples of 15-year-old students around the world every three years since 2000. In each of these five cycles (2000, 2003, 2006, 2009, 2012), both OECD and partner countries/economies participated, resulting in almost 300 cohorts defined by assessment year and country. Many of the OECD countries as well as a substantial number of partner countries/economies participated in each of the five PISA cycles so far.

In an effort to use the complete evidence on item functioning and scale coverage of the task material used in PISA, a database that merged all five cycles and all countries/economies was produced. This yielded a file that contains roughly two million student records assigned to each of the cycles by country/economy combinations.

Several analytical steps were performed to select the best fitting model. In particular, different and increasingly complex IRT models were specified and estimated, and model-data fit was compared using both AIC and BIC as well as measures of item fit. The analyses were carried out separately for each of the three PISA core domains of science, reading and mathematics.

This process resulted in the selection of a hybrid combination of item functions from either the Rasch model or the 2PL/GPCM. This model allowed fitting a wider range of items compared to using the Rasch model alone, which tends to lead to misfit in tests that contain a variety of response formats. In contrast to the 2PL/GPCM model being applied to all items, however, a number of slope parameters were fixed across items, so that the resulting model makes the same assumption as past models used in PISA on a subset of items (see Annex A5 and the *PISA 2015 Technical Report* [OECD, forthcoming]).

Modelling mode effects in computer- and paper-based administration

The hybrid model selected, based on analyses conducted on the historical database, was then applied to field-trial data for Groups 1 and 2 defined above.

The term “mode effect” refers to the observation that tasks presented in one mode, say, in paper-based assessments, may function differently when presented in another mode, say, when delivered as computer-based tasks.

To identify possible mode effects, the data collected in each of the two modes (PBA and CBA) were, first, validated separately against the hybrid item-response model, using item fit statistics (MD and RMSD). Model parameters (representing item difficulties and slopes) were then compared across modes. These analyses indicated a similarly high level of consistency for both the paper- and computer-based trend items when compared with the historical PISA database.

The nature of “mode effects” in PISA was then assessed by evaluating alternative statistical models on the pooled responses of students who took the test in paper and computer mode; in these models, mode effects are quantified by additional parameters or interaction terms.

Initial mode-effect explorations using an equivalent groups design

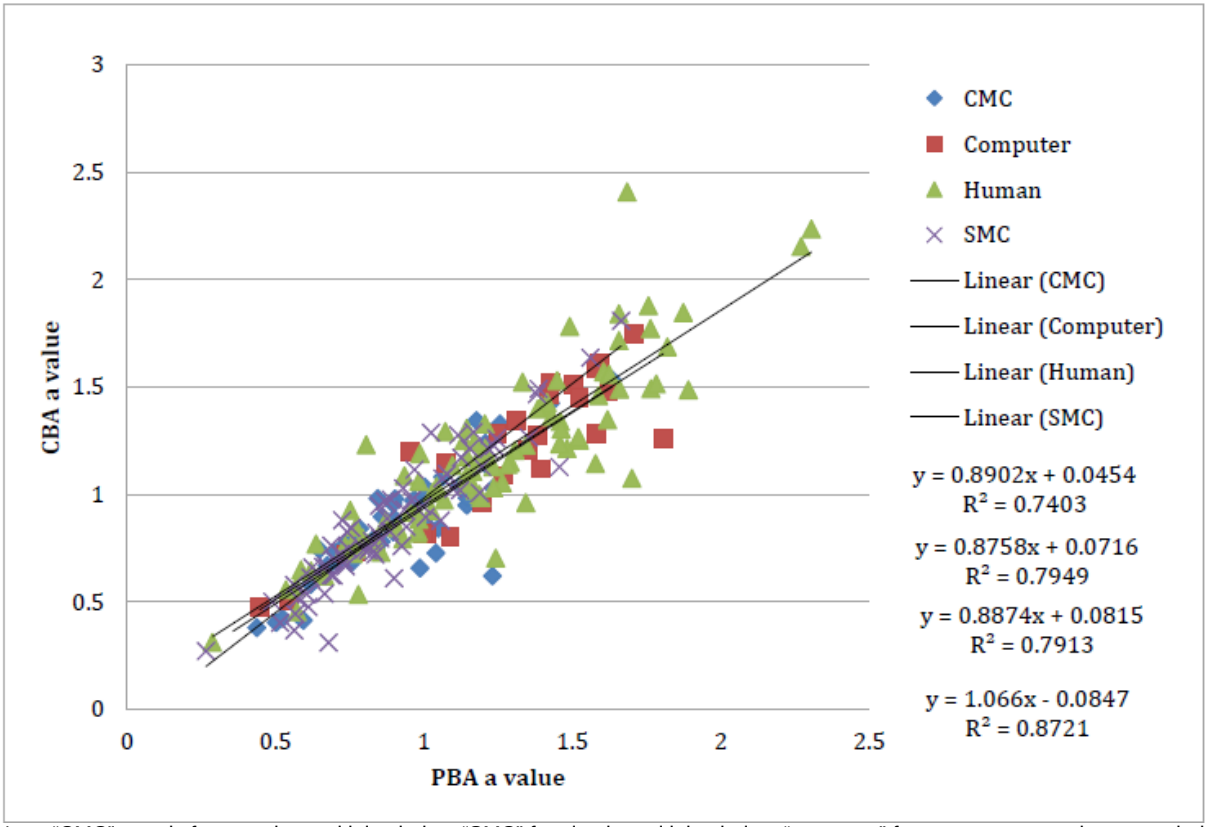
The initial explorations of mode differences followed an approach that goes back to Rasch (1960). Parameter invariance across groups (such as students who sat the test using paper and computer) can be examined by estimating the parameters of a model in these groups separately, and then by looking at the level of agreement among the two sets of parameters. This “graphical model test” is useful to spot systematic differences between modes of administration, but provides less statistical rigour than the models used in the following sections.

The equivalent groups' design of the field trial allowed for testing the hypothesis of "no mode effect" by comparing parameters estimated for the CBA mode to those that could be estimated by using the combined evidence from the PBA field trial and the responses of students who were given the same items in paper-based form in prior PISA rounds (2000-12).

The underlying ability distribution of PBA and CBA field trial samples is assumed to be identical (due to the random assignment of test-takers to modes). In the absence of mode effects, the CBA-based parameters should not differ significantly, or systematically, from the parameters obtained from a reanalysis of PISA 2000-12 datasets and verified using the PBA field trial sample. (PBA item parameters were derived under the guiding principle to retain as many Rasch model-based parameters as possible. For this reason, while the difficulty parameters can be compared for all items that were administered in paper and computer modes, the PBA-based set contains a number of slope values that are not estimated but fixed to be equal to 1, which produces fewer pairs of freely estimated parameters.)

Figures A6.2 and A6.3 show parameter comparisons between the mode-based samples (estimates of item parameters are based on 68 national centres that submitted their data through November 2014, i.e. about 3 000 responses per item, or an overall sample size of 150 983 for science, reading and mathematics, and 34 443 for financial literacy).

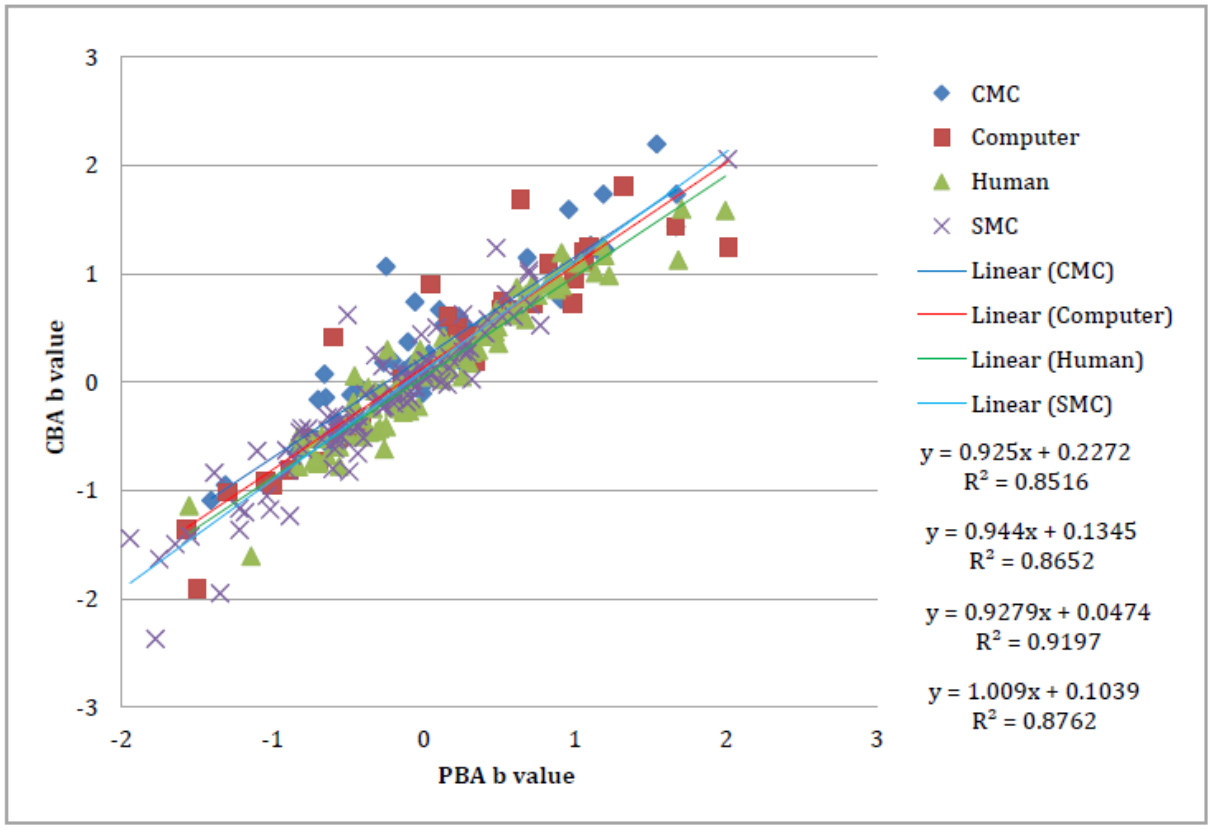
Figure A6.2. Comparison of slope parameter estimates across paper- and computer-based assessment modes for the PISA 2015 field-trial data



Note: "CMC" stands for complex multiple choice, "SMC" for simple multiple choice, "computer" for computer-scored open-ended and "human" for human-scored open-ended items. Each point represents one item from the domains of reading, mathematics, science or financial literacy.

Source: OECD, PISA 2015 field-trial database.

Figure A6.3. Comparison of difficulty parameter estimates across paper- and computer-based assessment modes for the PISA 2015 field trial data.



Note: "CMC" stands for complex multiple choice, "SMC" for simple multiple choice, "computer" for computer-scored open-ended and "human" for human-scored open-ended items. Each point represents one item from the domains of reading, mathematics, science or financial literacy.

Source: OECD, PISA 2015 field-trial database.

This visual test provided preliminary evidence of very good agreement between the two sets of parameters obtained from separate calibrations of responses in PBA and CBA modes. While there were differences, the level of difficulty of an item remains largely the same between PBA parameters based on historical data and CBA-based estimates. The same holds for the freely estimated slope parameters. Correlations between the difficulty parameters of PBA and CBA trend items were high within each domain, ranging from 0.92 to 0.95, as were correlations between the discrimination parameters (slopes), ranging from 0.90 to 0.94 (note that only the 2PL model based slopes were used to calculate correlations). The correlation of item-difficulty parameters across modes and domains is 0.94, and the correlation of item slope parameters is 0.91. Table A6.2 presents an overview of these correlations.

Table A6.2. Correlations of item-difficulty and item-slope parameters between PBA and CBA trend items within and across domains

Domain	Correlation of difficulty parameters (PBA,CBA)	Correlation of slope parameters (PBA,CBA)
Math	0.95	0.91
Reading	0.95	0.90
Science	0.92	0.94
Financial Literacy	0.94	0.92
All Domains	0.94	0.91

Source: OECD, PISA 2015 field-trial database.

The high correlations between item parameters across modes and domains (0.94) imply that the two modes measure the same constructs.

Item-response-theory models for mode effects

The next step was to more formally establish the invariance of item parameters across both modes. Strong measurement invariance holds if the same item parameters fit the items independent of the mode of administration.

In each domain, this formal investigation resulted in identifying a set of items for which invariance among the item parameters did not hold. Nevertheless, the results indicated strong invariance for most of the items.

This formal investigation focused on testing alternative conceptualisations of a "mode effect" by evaluating alternative statistical models that contain parameters that help to quantify and compare potential differences between PBA and CBA in an objective manner. The base model for these investigations was the 2PL model (Birnbaum, 1968):

$$P(x = 1|\theta, \alpha_i, \beta_i) = \frac{\exp(\alpha_i\theta + \beta_i)}{1 + \exp(\alpha_i\theta + \beta_i)} \qquad (1)$$

The alternative models tested provide information about whether the mode effect is best described in terms of an overall difference between assessment modes, whether this effect is a person-specific effect that may have a different impact on different groups, or whether it is an effect that has an impact on some subset of tasks. The comparisons of these models provide the basis for selecting the best-fitting model; this model was then applied in a way that anticipated the analysis of main survey data.

The first model tested on the data included a mode-effect parameter that changed the difficulty of all items in a test in the same direction and to the same extent. This was modelled by adding the same constant to all difficulty parameters in the case of the affected mode. This model describes a situation in which reading or, more generally, processing the item stem or stimulus is generally harder (by the same amount for all stimuli) on the computer, or entering a response is more tedious than filling in a bubble on an answer sheet of a paper-based instrument. This model can be written, using the above notation, as:

$$P(X = 1|\theta, a, \beta_i, \delta_m) = \frac{\exp(\alpha_i\theta + \beta_i - \delta_m)}{1 + \exp(\alpha_i\theta + \beta_i - \delta_m)}. \qquad (2)$$

where $-\delta_m$ represents how much more difficult (or easy) solving an item is when presented in a different mode relative to a reference.

In contrast to the assumptions made in model (2), one could argue that not all items become more difficult after moving them to the computer. Some could be more difficult, some could be at the same difficulty level, and some could become easier. This leads to a model that adds an item-specific effect to the difficulty parameter. In model (3) this is written as a differential item functioning (DIF) parameter, which quantifies the difference from the paper-based assessment, namely

$$P(X = 1|\theta, a, \beta_i, \delta_m) = \frac{\exp(\alpha_i\theta + \beta_i - \delta_{mi})}{1 + \exp(\alpha_i\theta + \beta_i - \delta_{mi})}. \qquad (3)$$

This decomposition indicates that the difficulties are shifted by some (item or item feature) dependent amount. For some items for which the response mode has no significant effect, constraint of the form $\delta_{mi} = 0$ may be added to the model.

The model given in equation (3) with constraints across both modes on slope parameters, as well as potential constraints on the DIF parameters, establishes a measurement invariance (e.g. Meredith, 1993) IRT model that can be viewed as representing weak factorial invariance. The more constraints of the type that $\delta_{mi} = 0$ there are, the more this model approaches a model with strong factorial invariance. The equality of means and variances of the latent variable in both modes is assumed because it is also assumed that respondents receiving the test in computer or paper mode are randomly selected from a single population.

Finally, if it cannot be assumed that the mode effect is a constant shift in difficulty for all respondents, one may assume that an additional ability ϑ is required to predict the response probabilities in the new mode accurately. This leads to model (4) in which a second latent variable is assumed, that is, another random effect is added to the item function for items administered in the new mode. The expression $\alpha_{mi}\vartheta$ in model (4) below indicates that there is a second slope parameter α_{mi} for items administered in the new mode ($i = I, \ldots, 2I$) and that the effect of the mode is person-dependent, quantified in the second latent variable ϑ:

$$P(X = 1|\theta, a, \beta_i, \delta_m) = \frac{\exp(\alpha_i\theta + \beta_i - \alpha_{mi}\vartheta)}{1 + \exp(\alpha_i\theta + \beta_i - \alpha_{mi}\vartheta)}. \qquad (4)$$

Note that the slope parameters and item difficulties, α_i, β_i, are as before in models (2) and (3) equal across modes; only the additional "mode slope" parameter α_{mi} needs to be estimated, plus the joint distribution $f(\theta, \vartheta)$ for which the variables are assumed to be uncorrelated: $\mathrm{cov}(\theta, \vartheta) = 0$.

Models (2), (3) and (4) can be applied to multiple populations, that is, by assuming one population per participating country or language group in PISA.

Table A6.3 below shows the results of models (2), (3) and (4) estimated on multiple populations using the PISA field trial data. All analyses were conducted with the software mdltm (von Davier, 2005).

Table A6.3. Measurement invariance assessment using mode-effect models

Domain	Model	Penalty AIC	AIC	Penalty BIC	BIC	Penalty CAIC	CAIC	Log Penalty	Akaike
Science	(2)	1694	5378045	10100	5386451	10947	5387298	0.586249	0.586433
	(3)	1984	5361306	11830	5371152	12822	5372144	0.584392	0.584608
	(4)	2180	5356556	12998	5367374	14088	5368464	0.583852	0.58409
Mathematics	(2)	620	1416987	3697	1420064	4007	1420374	0.526304	0.526534
	(3)	674	1409948	4019	1413293	4356	1413630	0.523668	0.523919
	(4)	714	1409235	4257	1412778	4614	1413135	0.523388	0.523654
Reading	(2)	818	1770885	4877	1774944	5286	1775353	0.534144	0.534391
	(3)	990	1760709	5903	1765622	6398	1766117	0.531022	0.53132
	(4)	1104	1758594	6583	1764073	7135	1764625	0.530349	0.530682
Financial literacy	(2)	192	253996	1003	254807	1099	254903	0.564498	0.564925
	(3)	236	251899	1233	252890	1351	253013	0.559736	0.56026
	(4)	248	251744	1295	252792	1419	252916	0.559365	0.559917

Source: OECD, PISA 2015 field-trial database.

As a general rule, lower values for the statistics (AIC, BIC, CAIC, log-penalty and Akaike) indicate better fit. However, when the magnitude of the statistics is similar, the more parsimonious model should be preferred. In all cases, Model (4) has the lowest values for these statistics, yet they do not differ appreciably from the fit for Model (3). To provide additional evidence for this interpretation, the marginal reliability of scores under each model as well as the correlation between estimates of the examinee's ability was computed. The median reliability for scores in all domains for each of the models was similar across groups, with median values ranging from 0.8 to 0.85. There are a few groups where the reliabilities are notably weaker (less than 0.6). The inclusion of these data has some influence on the model fit, but there is insufficient evidence based on reliability to suggest that Model (4) should be preferred over Model (3). Additionally, the correlation between estimated scores for

Models (3) and (4) in each domain is r = 0.999, which suggests that there is little added value in using Model (4). Based on these results, model (3) appears to describe the data sufficiently well. This means that there is a need to specify an item-specific, but not a person- (or country-) specific, mode effect parameter.

While these results imply the existence of mode effects (and the need to model these), the effects seen do not imply that performance on the computer test is influenced by an additional latent variable; so it can be assumed that weak factorial invariance holds, and that the CBA version of the test measures the same construct as the PBA. Also, the mode effect is not a homogeneous shift of difficulties, but rather one that affects some items more than others. A large percentage of items shows strong invariance and is not significantly affected by mode differences. While this makes common linear adjustment-based equating methods pointless, it opens opportunities to optimise the linking between paper- and computer-based assessments by means of item selection, and parameter-equality constraints for those items that are least affected by changes in presentation mode.

Results of estimating model (3) for each domain show that most mode effects on individual tasks are positive and some are negative, while in many cases the magnitude of these effects does not appear to be significant. Positive mode-effect parameters indicate that the computer-based version presents a greater difficulty; negative mode-effect parameters indicate that the computer-based version is, on the contrary, easier than the paper-based version. The inclusion of mode-effect parameters allows for fair comparison of performance across the two modes.

The existence of both positive and negative mode-effect parameters further implies that we can identify a set of items for which strong measurement invariance holds. Those items for which no significant effect can be detected form the basis for linking the CBA assessment to past PISA rounds, while all trend items can be used, if retained in future studies, to measure the construct, due to the invariance properties established in this section.

In summary, the relatively best-fitting model for evaluating and accounting for item mode effects among those considered here is the model that assumes the same parameters for the PBA as for the CBA and that adjusts the PBA-based item difficulty parameters by a DIF parameter without the introduction of an additional mode-specific skill. This indicates that weak factorial invariance can be assumed *for all trend items* administered in the CBA PISA field trial, while the size and range of the effects found indicates that *strong measurement invariance can be established for a significant subset of items.*

The impact of mode effects on country means

Having selected a model for students' responses to the paper- and computer-based tests, and identified (at the international level, using field-trial data only) which items required a parameter that quantifies mode differences, it is possible to evaluate the impact on country means of any mode effects that are not accounted for by the model.

Initial exploration of country-by-mode interactions

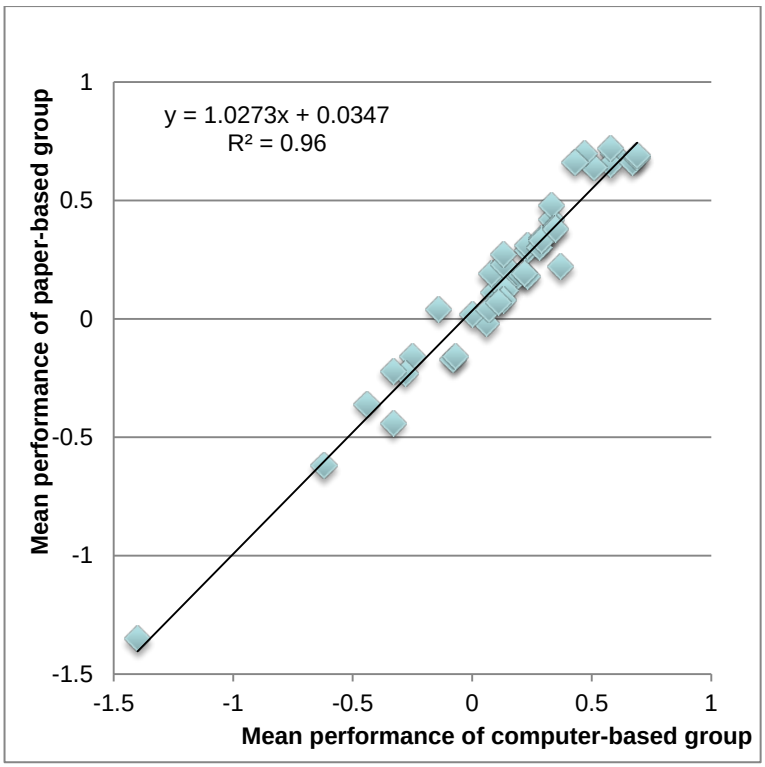
To evaluate the impact of the mode of administration on country means (after accounting for overall effects at the international level), the mean proficiency of students based on Model 3 was initially compared within each country/economy across mode and across a random split of schools (schools with an even number and schools with an odd number in the sample).

The model used for these comparisons incorporates scalar invariance for those items that showed little or no mode difficulty differences and assumes metric invariance for the remaining items. There are no country-specific mode effects applied in these analyses. This ensures comparability across countries while accounting for item-specific differences in difficulty for a subset of items only, with

these differences applied across all countries in the same way. This approach ensures that comparability is maximised, while mode effects that affect different items in different directions are accounted for so that potential effects on scale comparisons are minimised.

These comparisons are illustrated in Figures A6.4 and A6.5 for the domain of science. These figures show that for each domain, a good agreement of country means by assessment mode can be achieved. The differences observed between modes within countries are in the same order of magnitude (or lower) than differences based on a random school split. Thus, differences between modes within countries could, given the limited sample size for the field trial, plausibly result from differences, due to random sampling, between students who were assigned to computer- and paper-based groups, rather than from differences based on the mode of assessment.

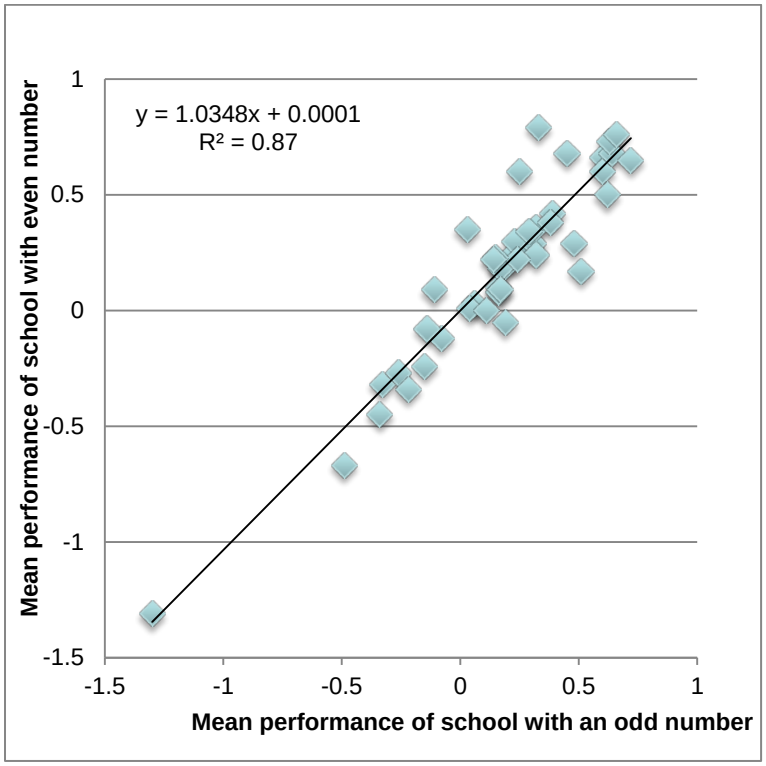
Figure A6.4. Mean science performance of computer-based and paper-based groups in the field trial



Note: All values are in the logit scale. Each point represents one national sample.

Source: OECD, PISA 2015 field-trial database.

Figure A6.5. Mean science performance of even- and odd-numbered schools in the field trial



Note: All values are in the logit scale. Each point represents one national sample.

Source: OECD, PISA 2015 field-trial database.

Formal testing of country-by-mode and gender-by-country-by-mode interactions

To investigate further the statistical significance of country-by-mode interactions, the preliminary proficiency estimate based on model (3) and on the instruments used in the field trial was compared across mode and gender within each national sample.

Data could be analysed for 57 national samples (among CBA countries in the main study, Austria and Brazil could not be included in field trial analyses; Scotland [United Kingdom] was treated as a separate sample; and Belgium contributed two distinct samples). In each of these samples, some examinees took the computer-based assessment (CBA) while others took the paper-based assessment (PBA). The ratio of CBA to PBA data available for analysis in each country was approximately 60:40 for mathematics and reading, and approximately 80:20 for science (due to the addition of new science clusters only available on CBA). The question of interest for this examination was whether there are significant differences in performance between boys and girls (gender effect) and/or CBA and PBA test versions (mode effect), and whether there is an interaction between these two variables with respect to the outcome.

Before examining gender and mode effects, provisional item parameters were estimated for each of the field-test items within each content domain (mathematics, reading, science). The CBA and PBA items were calibrated concurrently via a multiple-group IRT model to place all of the CBA and PBA items on the same scale using model (3) described above. The calibration was run using the field-trial data and item responses from the historical PISA database, which includes all prior assessment data across the 2000-12 PISA cycles. The response data and provisional item parameters were used to generate ten plausible values of performance for each student.

Methods

To examine gender and mode effects, several regression models were specified. For each model (described below), regression coefficients were estimated separately within each country (c), for each content domain (k). The outcome variable in all cases is one of the preliminary proficiency estimates for student *j*, the performance in the given domain. The process of estimating the coefficients and associated standard errors is as follows:

1. All cases were assigned a weight of 1.

2. A set of replicate weights was created for the purpose of computing jack-knife standard errors for the regression coefficients. These replicate weights were created on a country-by-country basis. In short, each country receives as many replicate weights as it has sampled schools in the field trial. The stratification variable used for the creation of the replicate weights was the school membership.

3. Regression coefficients were estimated ten times (once with each plausible value) for each model – again, separately for each country and each content domain. The average of the results from the ten regressions is used in the analysis.

4. The standard errors for significance testing were calculated by combining the corresponding sampling and measurement variance of the coefficient. The sampling variance was calculated using the JK1 replication method, and the measurement variance was calculated using the standard approach of using the variance of the ten estimates multiplied by an expansion factor.

Model (5) was specified as

$$PV_{jck} = b_{0ck} + b_{1ck}Gender_{jck} + e_{jck} \qquad (5)$$

For this model, the variable *Gender* was dummy coded with girls as the reference group (values for *Gender* are 0=girl; 1=boy). Cases where b_{1ck} is significant are indicative of a gender effect, independent of mode. The value of b_{1ck}represents the mean difference between boys and girls in the preliminary proficiency estimate.

Model (6) was specified as

$$PV_{jck} = b_{0ck} + b_{1ck}Mode_{jck} + e_{jck} \qquad (6)$$

For this model, *Mode* was dummy coded with PBA as the reference group (values for *Mode* are 0 = PBA; 1 = CBA). Cases where b_{1ck} is significant are indicative of a mode effect, independent of gender. The value of b_{1ck}represents the mean difference between CBA and PBA in the preliminary proficiency estimate.

Model (7) was specified as

$$PV_{jck} = b_{0ck} + b_{1ck}Gender_{jck} + b_{2ck}Mode_{jck} + b_{3ck}GenderMode_{jck} + e_{jck} \qquad (7)$$

As in models (5) and (6), *Gender* and *Mode* were dummy coded with girls and PBA as the reference groups respectively. In model 3, an interaction term is added where girls taking the CBA version were treated as the focal case (values for *GenderMode* are 1 = "girl taking a CBA test"; 0 = Otherwise). Cases where b_{1ck}, b_{2ck}, and/or b_{3ck} are significant are indicative of a main effect gender, main effect mode, and/or an interaction effect, respectively.

Results

Table A6.4 presents descriptive statistics for the parameter estimates in all three models across domains. The mean gender, mode, and interaction effect – irrespective of the model – across national samples is small, with one notable exception (a main effect for gender in the reading domain). With respect to mode-effects (coefficient b1 in Model 2) and mode-by-gender interactions (coefficient b3 in Model 7), the mean coefficient is always close to 0 (the magnitude can be assessed, for instance, in relation to main effects for gender, i.e. coefficient b1 in Model 5). This supports the conclusions that mode effects at the international level can be modelled, and accounted for, in order to report results of students taking the test on computers and in pencil-and-paper mode on the same scale. However, the table also highlights that there is some variation in individual samples around these average results.

Table A6.4. Mean and distribution of gender, mode and gender-by-mode interactions

		Coefficient	Min	Max	Mean	SD
Mathematics	Model 5	b1	-0.08	0.21	0.06	0.07
	Model 6	b1	-0.36	0.28	-0.03	0.09
	Model 7	b1	-0.14	0.21	0.05	0.08
		b2	-0.30	0.27	-0.04	0.09
		b3	-0.22	0.33	0.01	0.08
Reading	Model 5	b1	-0.40	0.03	-0.16	0.08
	Model 6	b1	-0.33	0.19	-0.03	0.08
	Model 7	b1	-0.50	0.01	-0.18	0.10
		b2	-0.31	0.30	-0.02	0.10
		b3	-0.21	0.19	-0.01	0.08
Science	Model 5	b1	-0.15	0.13	0.02	0.05
	Model 6	b1	-0.26	0.09	-0.02	0.06
	Model 7	b1	-0.16	0.17	0.02	0.07
		b2	-0.22	0.07	-0.03	0.06
		b3	-0.10	0.18	0.01	0.06

Note: All values are in the logit scale.

Source: OECD, PISA 2015 field-trial database.

When testing for the significance of country-level coefficients, given the multiple samples involved in the estimation, it is important to correct the significance level (alpha) for each test to avoid over-reporting the number of significant effects. A Bonferroni correction is commonly applied to produce an alpha level that controls the family-wise error rate for a family of independent tests; however, this type of correction is only an approximation of the Šidák correction (Šidák, 1967). The Šidák correction is formulated as

$$\alpha[FT] = 1 - (1 - \alpha[PF])^{1/C}$$

where $\alpha[FT]$ is the alpha per test, $\alpha[PF]$ is the alpha per family of tests (familywise alpha), and C is the number of tests. Note that this approach assumes that the tests are independent. The Bonferroni correction, on the other hand is formulated as

$$\alpha[FT] \approx \frac{\alpha[PF])}{C}.$$

The family-wise alpha under the Bonferroni approach is always less than or equal to alpha using the Šidák correction, although $\alpha[FT]$ is generally very similar. For each of the models, both the Šidák and Bonferroni corrections were applied to compare the regression slopes across national samples. Table 1 presents the familywise alpha levels and the corresponding critical values for a two-tailed test for both correction approaches. Note that the critical values under both corrections are similar.

Table A6.5. Familywise alphas and critical values

	α[*PF*]	Number of Tests	Šidák Correction for individual tests		Bonferroni Correction for individual tests	
			α[*FT*]	Critical Value	α[*FT*]	Critical Value
Models 5 & 6	0.05	57	0.00090	3.12156	0.00088	3.12894
Model 7		171	0.00030	3.43169	0.00029	3.43857

Tables A6.6, A6.7 and A6.8 present a summary of the significant effects from the regression models for each of the domains. The tables show the number of significant positive and negative effects (family-wise alpha = 0.05, two-tailed) for each of the slope coefficients under each of the models after adjusting the individual alpha level via the Šidák/Bonferroni correction. There were no differences in the number of identified significant effects using these two methods; hence, only one set of results adjusting for the multiple comparisons is presented.

Table A6.6. Count of significant coefficients for gender- and mode-effects in mathematics

Count of significant effects - Šidák/Bonferroni Correction

		Model 5 - Gender main effect	Model 6 - Mode main effect	Model 7 - Gender/Mode main effect + Interaction
Gender	Male -	0		0 / 0
	Male +	5		1 / 0
Mode	Computer -		1	0 / 0
	Computer +		1	0 / 0
Interaction	-			0 / 0
	+			0 / 0

Notes: alpha = 0.05/57 = 0.00088 (for individual effects). In Model 7, the first value is the count using alpha = 0.00088; the second is for alpha = 0.05/171 = 0.00029.

Source: OECD, PISA 2015 field-trial database.

Table A6.7. Count of significant coefficients for gender- and mode-effects in reading

Count of significant effects - Šidák/Bonferroni Correction

		Model 5 - Gender main effect	Model 6 - Mode main effect	Model 7 - Gender/Mode main effect + Interaction
Gender	Male -	32		18 / 15
	Male +	0		0 / 0
Mode	Computer -		1	0 / 0
	Computer +		0	0 / 0
Interaction	-			0 / 0
	+			0 / 0

Notes: alpha = 0.05/57 = 0.00088 (for individual effects). In Model 7, the first value is the count using alpha = 0.00088, the second is for alpha = 0.05/171 = 0.00029.

Source: OECD, PISA 2015 field-trial database.

Table A6.1. Count of significant coefficients for gender- and mode-effects in science

Count of significant effects - Šidák/Bonferroni Correction

		Model 5 - Gender main effect	Model 6 - Mode main effect	Model 7 - Gender/Mode main effect + Interaction
Gender	Male -	0		0 / 0
	Male +	2		0 / 0
Mode	Computer -		3	0 / 0
	Computer +		0	0 / 0
Interaction	-			0 / 0
	+			0 / 0

Notes: alpha = 0.05/57 = 0.00088 (for individual effects). In Model 7, the first value is the count using alpha = 0.00088, the second is for alpha = 0.05/171 = 0.00029.

Source: OECD, PISA 2015 field-trial database.

In each of the content domains there are some samples for which the test of significance of gender returns a positive result. In mathematics and science, there are five and two national samples, respectively, where boys perform significantly better than girls (at a familywise error level or 5%) under Model 5. In reading, there are 32 samples where girls perform significantly better than boys; this corresponds to a well-known and stable effect seen in many reading assessments, not only PISA. In Model 7, the number of significant cases decreases, but there are still several samples with gender effects.

With respect to the mode effect, there are very few samples with any significant differences between CBA and PBA performance. In instances where there is a significant effect, performance on the PBA version tends to be higher. When accounting for gender and interaction effects in Model 7, all of the mode differences disappear. In all cases for Model 7, there are no significant interaction effects. That is, there appear to be no significant positive or negative interactions for the variable that identifies the group of girls taking the computer-based assessment.

Limitations of field-trial analyses on country-by-mode interactions

There are important limitations that should be noted when interpreting the results of mode effects in national field trial samples.

First, national field trial samples are small, and do not allow for the precision achieved for subgroup comparisons in the main study. The reduced statistical power means that small and moderate mode effects, or mode-by-gender interactions, could not be detected with sufficient statistical power in the field trial sample, because similar differences could plausibly result from random sampling alone.

The sampling design for the field trial was established to allow for the validation of assessment instruments and various procedures, namely a preliminary scaling of items at the international level and the identification of problematic items and administration issues based on the data aggregated across national samples. Given this intended purpose and resource considerations, in addition to reduced samples, some allowances were made in the field-trial data collection to better accommodate schools and national teams. As such, it cannot be assumed that samples drawn are representative of the student population in every country, so that country-level conclusions cannot be derived from the data.

Next, there are cases where the within-school randomisation was not fully achieved, e.g. because the sample of examinees is small and/or students within a school were only given tests via a single mode (all CBA or PBA) due to a technical failure. Every effort was made to include all cases in the analyses; however, the inclusion of cases outside of the sampling design has the potential to affect the jack-knife results. Whether this is likely to result in greater or fewer significant effects is unclear.[2]

In addition to the sampling limitations, which are due to the fact that countries have limited resources and cannot collect a full representative sample of students, there are limitations on the side of proficiency estimation. The assessment materials contain preliminary versions of the final PISA tests, the background data collection is preliminary, and the computer-based delivery platform is tried for the first time in the field test.

Finally, the field trial served not only as a way to review the quality of test materials, but also, in some cases, to identify problems with the hardware (computers and USB keys) used in schools to administer the tests. These problems, where they appeared, could have affected the performance of students in the field trial, but were addressed before the main study to ensure that all students could perform at their best.

Summary

Across all three domains of science, reading and mathematics, there are very few samples where the main effect of the mode of delivery reaches statistical significance; and there is no significant interaction effect between the gender and mode variables. Identifying gender and/or mode effects in the field-trial data relied on preliminary scaling results and field-trial instruments that do not reflect the main survey test. As such, care should be taken not to over-interpret these findings.

References

Birnbaum, A. (1968), "On the estimation of mental ability", *Series Report 15*, USAF School of Aviation Medicine, Randolph Air Force Base, Texas.

Gonzalez, Eugenio J. (2014), "Calculating standard errors of sample statistics when using international large-scale assessment data", *Educational Policy Evaluation through International Comparative Assessments*, Vol. 59.

Meredith, W. (1993), "Measurement invariance, factor analysis and factorial invariance", *Psychometrika*, Vol. 58/4, pp. 525–43, http://dx.doi.org/10.1007/BF02294825.

OECD (forthcoming), *PISA 2015 Technical Report*, OECD Publishing, Paris.

Rasch, G. (1960), "Probabilistic Models for Some Intelligence and Attainment Tests", Nielsen & Lydiche, Copenhagen.

Šidák, Z. K. (1967), "Rectangular confidence regions for the means of multivariate normal distributions", *Journal of the American Statistical Association*, Vol. 62/318, pp. 626-633.

[1] Not all 58 national centres conducted the field trial and submitted the data in time to be included in the analysis that informed the decisions about main survey instruments. Some analyses presented in this annex are therefore based on fewer than 58 samples.

[2] Some 2 763 schools across countries contributed their sample to these analyses. Fifty schools have fewer than 10 students across administration modes, 100 schools have 10 to 24 students, 1 498 schools have 25 to 49 students, and 1 115 schools have more than 50 students. The schools with fewer than 10 students are distributed across 17 national samples; 11 of these 50 schools are located in one country. Not counting schools with fewer than 5 students, there are 476 schools where only one administration mode was used. Of these schools, 12 administered the paper-based version only; the remaining schools administered the computer-based version only.

For more information on the Programme for International Student Assessment and to access the full set of PISA 2015 results visit:

www.oecd.org.edu/pisa