Harrison *et al.*
Large-scale Assessments in Education (2023) 11:8
https://doi.org/10.1186/s40536-023-00157-9

Large-scale Assessments in Education

RESEARCH Open Access

Comparing the score interpretation across modes in PISA: an investigation of how item facets affect difficulty

Scott Harrison[1,2*], Ulf Kroehne[1], Frank Goldhammer[1,2], Oliver Lüdtke[2,3] and Alexander Robitzsch[2,3]

*Correspondence: s.harrison@dipf.de

[1] DIPF | Leibniz Institute for Research and Information in Education, Rostocker Straße 6, 60323 Frankfurt Am Main, Germany
[2] Centre for International Student Assessment – ZIB, Marsstr. 20-22, 80335 Munich, Germany
[3] IPN — Leibniz Institute for Science and Mathematics Education, Olshausenstraße 62, 24118 Kiel, Germany

Abstract

Background: Mode effects, the variations in item and scale properties attributed to the mode of test administration (paper vs. computer), have stimulated research around test equivalence and trend estimation in PISA. The PISA assessment framework provides the backbone to the interpretation of the results of the PISA test scores. However, an identified gap in the current literature is whether mode effects have affected test score interpretation as defined by the assessment framework, and whether the interpretations of the PBA and CBA test scores are comparable.

Methods: This study uses the 2015 PISA field trial data from thirteen countries to compare test modes through a construct representation approach. It is investigated whether item facets defined by the assessment framework (e.g., different cognitive demands) affect item difficulty comparably across modes using a unidimensional two-group generalized partial credit model (GPCM).

Results: Linking the assessment framework to item difficulty using linear regression showed that for both maths and science domains, item categorisation relates to item difficulty, however for the reading domain no such conclusion was possible. In comparing PBA to CBA in representations across the three domains, maths had one facet with a significant difference in representation, reading had all three facets significantly different, and for science, four out of six facets had significant differences. Modelling items labelled "mode invariant" in PISA 2015, the results indicated that in every domain, two facets showed significant differences between the test modes. The graphical inspection of difficulty patterns confirmed that reading shows stronger differences while the patterns of the other domains were quite consistent between modes.

Conclusions: The present study shows that the mode effects on difficulty vary within the task facets proposed by the PISA assessment framework, in particular for reading. These findings shed light on whether the comparability of score interpretation between modes is compromised. Given the limitations of the link between the reading domain and item difficulty, any conclusions in this domain are limited. Importantly, the present study adds a new approach and empirical findings to the investigation of the cross-mode equivalence in PISA domains.

Keywords: PISA, Assessment framework, Mode effects, Item facets, Construct representation, Generalized partial credit model

Background

When altering how a test is administered, referred to as test mode, an important step is ensuring the two versions of the test are equivalent, minimising the impact of the change. For the Programme for International Student Assessment (PISA), a major change occurred in 2015, when the main domains of mathematics, reading, and science, were digitised and assessed using computers in the majority of participating countries. This change in mode gives rise to questions about test score interpretation, particularly with respects to the underlying framework which is used to organise and operationalise the test items. To frame this study, three key areas need to be considered: (1) previous research on mode effects in PISA; (2) cross-mode equivalence in terms of the test score interpretation; and (3) the PISA assessment framework defining item facets, that may determine item difficulty.

Mode effects

The term mode effect refers to non-equivalence in psychometric item and scale properties arising from changing the mode of test administration (Kroehne & Martens, 2011). Cross-mode equivalence has formed an important part of the discussion around the psychometric equivalence of test versions, and refers among other criteria to the comparability of the test score interpretation (Buerger et al., 2016; Huff & Sireci, 2001; Kingston, 2008; Wang et al., 2008). Importantly, "research does generally seem to indicate, however, that the more complicated it is to present or take the test on computer, the greater the possibility of mode effects" (Pommerich, 2004, pp.3–4). A number of large scale assessments, such as the National Assessment of Educational Progress (NAEP) (Bennett et al., 2008), the Programme for International Assessment of Adult Competencies (PIAAC) (OECD, 2013), the Programme for International Student Assessment (PISA) (OECD, 2016), Trends In International Mathematics And Science Study (TIMSS) (Fishbein et al., 2018) and Progress in International Reading Literacy Study (PIRLS) (Mullis & Martin, 2019), have made or are making the transition from paper based assessment (PBA) to computer based assessment (CBA).

For PISA, the main transition from PBA to CBA was part of the 2015 main study, after offering CBA options in other areas in previous cycles. Each cycle of the PISA study is proceeded with a field trial, which is used to evaluate newly developed items and to try out field operations within PISA. The field trials for the 2015 PISA cycle, conducted in 2014, used both CBA and PBA assessment modes across 58 countries to assess equivalence across modes. Within schools, tests were randomly assigned to students as either PBA or CBA using a rotated booklet design. As such, the field trial was a form of bridge study (Mazzeo & von Davier, 2008), where the collected data was used for investigating the linking of the two assessment modes, by identifying PISA test items with no significant mode effect. The 2015 OECD PISA technical report showed mode effects were present for some items across countries (non-invariant items), with CBA being on average harder than PBA (OECD, 2016).

Following the field trial, the OECD report on the PISA 2015 results dedicated Annex 6 (OECD, 2016) to mode effects, explaining in detail the model selection process and subsequent domains, country, and gender analysis. There are a number of important conclusions drawn from the mode effect analysis that help establish the motivation for this study. First, it was concluded that "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" (OECD, 2016, p. 9). Those items for which no significant mode effect can be detected form the basis for linking the CBA assessment to past PISA cycles, while all trend items can be used, if retained in future studies, to measure the construct, due to the invariance properties. It was concluded from this study that "the effects seen do not imply that the validity of performance assessment on the computer test is influenced by an additional latent variable" (OECD, 2016, p. 9). Important for the present study, the technical report did *not* address how the identified mode effects at an item level (e.g., item difficulties and item discriminations) may affect the construct interpretation as defined by the PISA assessment framework (OECD, 2017b), leaving an obvious gap in the literature.

Following on from the initial PISA technical report, several independent studies (Feskens et al., 2019; Jerrim et al., 2018; Robitzsch et al., 2020) have since added to the body of literature on mode effects in PISA. For Germany, an analysis was undertaken by Robitzsch et al. (2020), which focused on marginal trend estimation and mode effects using the PISA 2015 field trial data. There was a decline in mathematics and science in the PISA 2015 main study. A key finding of the work by Robitzsch et al. was that in the presence of mode effects, trend estimation is still possible using the 2015 field trial items as a bridge study, which enables the linking of PBA (until PISA 2012) and CBA test (since 2015). Using linking procedures for estimating marginal trends that account for mode effects, the German average PISA scores for mathematics and science were estimated to have increased over this time.

Reanalysing the PISA 2015 field trial data, Jerrim et al. (2018) used data from Germany, Ireland, and Sweden to identify the presence of mode effects. One of the essential goals of this research was to test for mode effects within the data only on items deemed mode invariant by the OECD report. It was expected that in removing these affected items, the mode effect should disappear. However, there are still negative effects for all countries that are not statistically significant for mathematics and reading, and in Germany, a significant negative effect for science, that are of an important magnitude (3–9 points) on the PISA metric. It can be argued that the previous research on mode effects in PISA converges on the idea that item difficulty differs between modes. In contrast, the question of whether the assessed constructs are equivalent across modes has attracted less attention, although it represents an important prerequisite of test equating (Holland & Dorans, 2006).

Test score interpretation and construct representation

An important criterion of test equivalence means that the test score from each mode can be interpreted to be determined by the same constructs. The International Test Commission (ITC) developed best practice guidelines for test developers, publishers and users on how to ensure equivalence (International Test Commission, 2005, pp.24–25). Specifically, Sect. 2c outlines that developers should "provide clear documented evidence of the equivalence between the CBT/Internet test and non-computer versions". Investigating construct equivalence across modes can be empirically done through a number of approaches, including: (1) cross mode correlation, which requires a within-subject design (Kroehne et al., 2019); (2) comparing construct representations as reflected by the effect of construct-relevant item characteristics (item facets) on difficulty; (3) comparing the nomothetic span using theoretically relevant covariates (Buerger et al., 2019); and (4) analyzing dimensionality, such as a random mode effect component across persons (Annex 6, OECD, 2016).

Recent empirical studies investigating mode effects in terms of the construct interpretation have used approach (3) for the German National Educational Panel Study (NEPS) reading items (Buerger et al., 2019) and approach (1) for PISA reading items (Kroehne et al., 2019). In PISA 2015, approach (4) was used to evaluate construct equivalence by assessing whether another latent variable is required to model the data (see Annex 6, OECD, 2016). However, a gap in the literature to date is to investigate whether item facets, as defined by the PISA assessment framework, determine item difficulty comparably between modes.

The theoretical work by Embretson (1983) lays the foundation for an analysis of approach (2), which underpins the item construction and interpretation of test scores. The construct representation approach described by Embretson (1983) is "concerned with identifying the theoretical mechanisms that underlie task performance" (Embretson, 1983, p.180). It thus can be used to investigate how mode effects interact with facets defined by the PISA assessment framework and, in turn, affect the comparability of test score interpretation across modes. Following the construct representation approach, evidence for a valid construct interpretation is provided if (construct-relevant) item facets determine item difficulty as hypothesized. It is assumed that theory-based item facets determine required components of information processing and thus account for the difficulty of an item. For instance, one facet may represent the type of cognitive processing required in an item and there may be a hypothesis about which type is the most/least challenging one. To illustrate this approach, Fig. 1 shows an example construct representation based on a single item facet with three different categories of items.

In Fig. 1a, the construct for PBA is represented by the line linking P1, P2 and P3, and for the CBA mode, C1, C2 and C3. While there is a overall difference in mean item difficulty between modes, this difference is consistent, and as such, the overall construct representation would be considered to be the same. This would mean in turn that the test scores for both the PBA and CBA versions of the test are equivalent in their

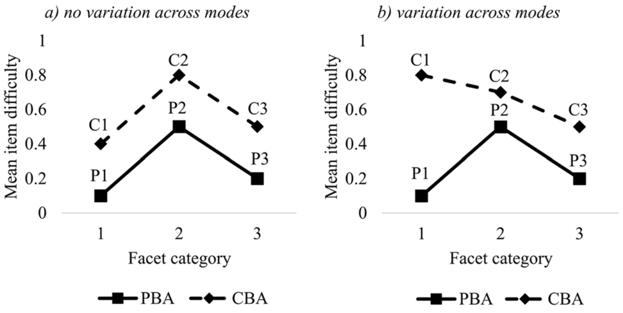

Fig. 1 Example Construct Representation

interpretation. In Fig. 1b, the construct representation for the PBA mode remains the same, however the CBA construct representation shows C1 with the highest mean item difficulty, with C2 slightly less, and C3 having the lowest mean item difficulty. As such, the variation in how the item facet determine difficulty may indicate that the test score interpretations for each mode is different. To obtain a high test score in CBA the test taker needs to meet the challenge represented by category 1 while in PBA test takers are most awarded when they can solve category 2 items.

The PISA assessment framework

The PISA assessment framework was designed by subject matter experts to operationalise the overall assessment objectives of PISA. It is comprised of a number of item facets within each domain, three in mathematics, three in reading, and six within science (OECD, 2017c). The facets themselves are conceptual categorizations that help operationalise the domain specific assessment objectives. A central concept for the assessment development was that the collection of items assembled to a test should be 'balanced' across the facets defined by the assessment framework to ensure a complete construct representation (Stacey & Turner, 2015).

Each item reflects the underlying facets of this assessment framework and correlates to various aspects as to what the student is required to undertake in answering the question. The facets and their corresponding facet categories are presented in Table 1, along with the number of items associated with each facet and the categories within the facets (OECD, 2017c). The percentage of items also identified by the OECD as mode-effect invariant by category has been added to highlight mode effect variation between the facet categories.

One of the challenges with PISA is that the assessment framework is developed by contractors with consultation of the subject matter expert groups in each cycle. The development of the framework can be "characterized by continuous revision of the same framework over many years and the involvement of academic experts across science education, science, learning psychology, assessment, and policy makers" (Kind, 2013, p. 672). The framework for 2015 was applied equally to the PBA and CBA implementations of the test. As such, it is a reasonable expectation that the two tests are equivalent; that the PBA and CBA versions of the test measure the construct in the same manner.

Construct-relevant item characteristics (item facets), as defined by the PISA assessment framework, can be expected to determine item difficulty, for example, the item difficulty in reading items can be assumed to depend on the required cognitive processing (i.e., the cognitive aspect of reading). While the coverage of the assessment framework is achieved by many items, the composition of the measurement is constant at the level of interest (countries), because of the rotated booklet design. If individual facets of the assessment framework are affected differently by the mode effect, construct representation and test score interpretation, respectively, might change.

The PISA assessment framework defines facets that can be regarded to be more (e.g., type of cognitive processing) or less (e.g., situation) related to the targeted construct. Nevertheless, we decided to consider all facets, as all facets contribute to the composition of measurement, and in turn may affect the test score interpretation. Thus, for

Table 1 PISA facets and categories

Maths—Facet	**Category**	**N items = 83 (% invariant items)**
Content	Change and relationships	22 (77)
	Quantity	21 (52)
	Space and shape	19 (58)
	Uncertainty and data	21 (52)
Situation and context	Occupational	20 (65)
	Personal	13 (85)
	Scientific	22 (59)
	Societal	28 (46)
Process	Employing mathematical concepts, facts and procedures	36 (61)
	Formulating situations mathematically	24 (58)
	Interpreting, applying and evaluating mathematical outcomes	23 (61)
Reading—Facet	**Category**	**N items = 103 (% invariant items)**
Situation	Educational	30 (63)
	Occupational	20 (65)
	Personal	29 (55)
	Public	24 (71)
Text format	Continuous	62 (66)
	Mixed	7 (57)
	Multiple	3 (33)
	Non-continuous	31 (61)
Aspect	Access and retrieve	26 (54)
	Integrate and interpret	53 (60)
	Reflect and evaluate	24 (79)
Science—Facet	**Category**	**N items = 85 (% invariant items)**
Context 1	Global	17 (59)
	Local/National	58 (74)
	Personal	10 (80)
Context 2	Environmental quality	11 (64)
	Frontiers	31 (71)
	Hazards	9 (67)
	Health and disease	15 (87)
	Natural resources	19 (68)
System	Earth and space	18 (67)
	Living	39 (74)
	Physical	28 (71)
Competency	Evaluate and design scientific enquiry	16 (75)
	Explain phenomena scientifically	41 (73)
	Interpret data and evidence scientifically	28 (68)
Knowledge	Content	51 (71)
	Epistemic	10 (80%)
	Procedural	24 (71%)
Depth of knowledge	Low	30 (63%)
	Medium	48 (73%)
	High	7 (100%)

investigating the comparability of the test score interpretation all facets defined by the respective assessment framework are taken into account.

Although the theoretical mapping of items to facet categories based on expert opinion is justifiable (American Educational Research Association et al., 2014), the way this was done in PISA is not necessarily based on a priori assumptions that guided the development of the items. However, this would be a most rigorous approach to demonstrate validity evidence based on test content. Moreover, these is not a strong body of literature about the specific categorisations in PISA based on theoretical arguments and empirical evidence, especially in relation to item difficulty. Given these limitations (i.e., justification of each facet, completeness of construct-relevant facets, and assignment of items to facets) we do not claim to perform construct validation because this would require a strong(er) theoretical justification of facets and their relation to the respective construct. Instead we investigate the comparability of score interpretations across modes using the available facets as defined by the PISA assessment framework. If the facets determine item difficulty, they affect the score interpretation no matter how strong the facets are theoretically justified.

Research questions

The main objective of this study is to extend the evidence regarding the equivalence of score interpretation between modes in the main PISA domains. To this end, we analyse mode effects in relation to item facets defined by the PISA assessment framework. More specifically, we investigate whether the considered facets determine item difficulty and in turn score differences comparably across modes. For this, the PISA 2015 field trial data for 13 countries is used.

Three research questions frame our study. The first question focuses on whether the item facets defined by the assessment framework relate to item difficulty. This is required to establish that a link between the assessment framework and item difficulty exists, and lays the foundation for the construct representation approach used for the other research questions. The second question focuses on whether there is a significant difference between modes in how the difficulty varies across categories of the item facets defined the PISA assessment framework. The third question focuses on the items flagged as mode invariant after the PISA 2015 field trial. It is investigated whether the link between mode effect and item facet categories will change when only using mode invariant items. As such, the third research question is whether any differences between modes persist when using only mode invariant items?

Methods

Sample

For the study, all countries that participated in the 2015 PISA field trials with both PBA and CBA modes were approached to provide data. Overall, we attained the support of 13 countries to provide their data. The sample size for each domain varies slightly due to the PISA test rotation design. Importantly, the 2015 field test resulted in more students taking the computer-based version of the test than the paper-based version. The average number of responses per item are presented in Table 2 by domain and mode. For one item in the reading domain, only three countries had CBA responses, resulting in

Table 2 Summary of pooled data for 13 Countries

	Mathematics		Reading		Science	
	PBA	CBA	PBA	CBA	PBA	CBA
N	4760	7200	4701	7094	4703	7158
Avg. N per item	1555.6	2300.2	1546.9	2141.5	1553.4	2282.7
(SD)	(60.9)	(95.8)	(23.9)	(353.2)	(23.4)	(96.1)
Avg. Age (Yrs)	15.53	15.53	15.53	15.53	15.53	15.53
(SD)	(0.29)	(0.29)	(0.28)	(0.29)	(0.29)	(0.29)
Female	48.45%	49.03%	48.22%	48.98%	48.71%	48.70%

the standard deviation for average number of CBA responses to be higher when compared to the mathematics or science domains.Table 2 also shows the average number of responses per item for the two modes of test administration, along with the average age and gender composition.

The rotation design of the PISA study means that not all items were administered to every student, resulting in the variation between the sample sizes in the three assessed domains. A key consideration for this study (as with the OECD's original study) (OECD, 2016) is that country specific model-based analyses would be limited due to the number of responses elicited at the country level. Given the data were obtained from a field trial with a relatively small number of students per item, the average number of responses per item at a country level would typically be insufficient to facilitate two parameter logistic (2PL) modelling being used in PISA since 2015. For example, the average number of responses for Germany was between 100 and 200 responses per item within each mode. Country level analysis is outside the scope of this research, so individual countries are not used as a covariate here, however using a pooled approach with countries as strata, the average number of responses per item is over 1500, which means that a 2PL model is expected to provide stable item parameter estimates.

Statistical modelling

The statistical approach to answer the three research questions can be understood as a multi-step process. The first step is to estimate item difficulties on the IRT scale for PBA and CBA separately using a two-group model. Here, the PBA item difficulties serve as a benchmark for assessing whether there is a relationship between the item difficulties and the facets used in the PISA assessment framework, as asked in the first research question. The second step is to estimate the mean difficulty for PBA and CBA for each item facet category included in the PISA 2015 assessment framework. This provides the basis for the third step. Here, the relationship between facet categories and difficulty within each domain facet is compared across modes. Thus, the aim is to falsify the null hypothesis that there is no difference across modes in how the average difficulty varies across facet categories (see Fig. 1a). This will answer research question two, and when repeated with only mode-invariant items, will also answer question research question three.

For step one, a statistical approach to estimating item difficulties was undertaken that is similar to the approach undertaken by the OECD, as described in Annex 6 (OECD, 2016, pp. 7–8). The OECD approach used a hybrid combination of item functions drawn from the Rasch model, two parameter logistic model (2PL), and Generalized Partial Credit Model (GPMC) model. For this analysis, the measurement model chosen is the GPCM, as it can most closely approximate the OECD approach in a single model. The GPCM proposed by Muraki (1992) is shown in Eq. 1, with an additional subscript indicating mode.

$$P(X_{\Im}=k|\theta)=\frac{exp\left[\sum_{h=0}^{k}a_{\Im}(\theta-b_{\Im}+d_{ihm})\right]}{\sum_{j=0}^{K_i}exp\left[\sum_{v=0}^{j}a_{\Im}(\theta-b_{\Im}+d_{ivm})\right]},k=0,\ldots,K_i \tag{1}$$

where X_{im} denotes the item response of item i in mode m (m = pba, cba; for paper-based and computer-based administration), across categories k. Note that the item discrimination a_{im} and item difficulty b_{im} are mode-specific parameters. Given m, item step parameters d_{ihm} are estimated using the constraints $d_{i0m}=0$ and $\sum_{h=1}^{K_i}d_{ihm}=0$.

To address the three research questions a multi-group modelling approach was used. More specifically mixture models with two known classes representing the administration modes CBA and PBA were tested. Assuming random equivalence of the two mode groups, the latent variable for ability in each group was constrained to a mean of 0 and variance of 1, whereas the item parameters (thresholds, loadings) were estimated freely between groups to capture potential item-level mode effects. For each facet such a model was estimated.

To transform the estimated model parameters to the IRT scale, item thresholds were converted to item difficulties using Eq. (2), taken from Asparouhov and Muthen (2016):

$$b_{ik}=\frac{\tau_{ik}-\lambda_{ik}\alpha}{\lambda_{ik}\sqrt{\psi}} \tag{2}$$

where b is the estimated difficulty for item i in category k; τ is the threshold, λ is the factor loading, and α and ψ are the mean and variance of the factor f, respectively (Muthen, 2017). Given the model constraints, Eq. (2) simplifies to the fraction of threshold and factor loading.

To address the first research question, item difficulties b_{ik} from the PBA group are used as the criterion variable in a multiple regression model, with the facets included in the PISA assessment framework forming the predictor variables (see e.g., Hartig et al., 2012). The regression model is shown in Eq. (3):

$$b_{i,PBA}=\sum_{p=0}^{P}\beta_p x_{ip}+e_i \tag{3}$$

where $b_{i,PBA}$ is the item difficulty from Eq. (2) for the PBA items only, as the PBA item difficulties are used as the benchmark (note that in the case of a partial credit item we simply used the average of the item's category difficulties b_{ik}). β_p is the regression coefficient for item facet p, where P is equal to the total number of facets in the domain. x_{ip} indicates which category of facet p applies to item i.

For the second step in the analysis, additional parameters are derived for the average item difficulty for each facet category and for each of the two test modes. These average facet category difficulties are shown in Fig. 1. To do this, the mean item difficulty was added as a new mode-specific parameter for each facet category by Eq. (4):

$$\overline{b}_{fm} = \frac{\sum_{i=1}^{n_f} b_{\Im}}{n_f} \tag{4}$$

where $\overline{b}_{fm}$ is the mean item difficulty for facet category *f*, and $b_{\Im}$ is the difficulty of item *i* administered by mode *m*. For example, the Maths facet for *process* in Table (1) has three categories, so the mean item difficulty is calculated for each facet category, by mode, to create six mean values in total.

Once the mean values are obtained for each facet category, the representation of differences in mean difficulties for each mode is then formulated. This is done by differencing the mean values within each mode to its adjacent category. Adapting Fig. 1a as an example, the three facet categories shown requires two additional parameters to be derived. This is done for both modes, shown in Eqs. (5–8):

$$D_{1,PBA} = \left(\overline{b}_{1,PBA} - \overline{b}_{2,PBA}\right) \tag{5}$$

$$D_{2,PBA} = \left(\overline{b}_{2,PBA} - \overline{b}_{3,PBA}\right) \tag{6}$$

$$D_{1,CBA} = \left(\overline{b}_{1,CBA} - \overline{b}_{2,CBA}\right) \tag{7}$$

$$D_{2,CBA} = \left(\overline{b}_{2,CBA} - \overline{b}_{3,CBA}\right) \tag{8}$$

where $D_{1,PBA}$ and $D_{2,PBA}$ are the differences between P1 and P2, and P2 and P3 in Fig. 1 respectively. The same applies to $D_{1,CBA}$ and $D_{2,CBA}$, estimating the differences from C1 to C2, and C2 to C3 respectively.

For the final step in the analysis, a statistical test is required to measure if the estimated differences between adjacent facet categories are significantly different across modes. If there are cross-mode differences, the facet determines item difficulty differently and in turn cross-mode differences in score interpretation are suggested. To test cross-mode differences, either a Wald test or likelihood ratio test (LRT) can be used. In comparing the two tests, "they have similar behaviour when the sample size *n* is large and H_0 is true" (Agresti, 2007, p.11). Given the number of observations attained through pooling the data, the Wald test is expected to provide comparable results to an LRT approach. Given the complexity and number of the models estimated, the Wald test was also selected for its computational simplicity in that each model only needs to be estimated once.

The Wald test is applied across all categories of a facet to test the null hypothesis (H_0) that there is no significant difference between modes in how average difficulty varies across facet categories. For the example from Fig. 1 with three facet categories for both PBA and CBA we constrain the difference in means from Eqs. (5) and (7) to zero which gives Eq. (9), and the difference in means from Eqs. (6) and (8) to zero which gives Eq. (10). As such, the Wald test statistic used to test the null hypothesis combines Eqs. (9) and (10) to test if there is a significant cross-mode difference in how facet categories determine difficulty:

$$D_{1,CBA} - D_{1,PBA} = 0 \tag{9}$$

$$D_{2,CBA} - D_{2,PBA} = 0 \tag{10}$$

Equation (9) is equivalent to saying there is no significant difference across modes, of the differences between facet categories 1 and 2 within each mode. Equation (10) repeats the procedure, but instead compares facet categories 2 and 3. Combining both equations into one Wald test statistic allows testing of the null hypothesis, that is, there is no difference in score interpretation between modes.

Using the Mplus software (Muthen & Muthen, 2017), the parameters of the two-group item response model were estimated using robust maximum likelihood estimation (MLR). Missing responses for items not reached or flagged as not applicable, are incorporated into estimation by being scored "NA", in accordance with the PISA scoring guide for 2015 data (OECD, 2017a, pp. 198). Furthermore, stratification for countries, and clustering of students within schools to model the PISA sampling process was incorporated for obtaining adjusted standard errors.

Statistical inference

For determining if there is a significant difference between modes, we need to select a suitable type 1 error rate (alpha). This is a more nuanced matter, especially when multiple tests are being conducted for hypothesis testing. A common level of alpha is 0.05, meaning there is a 5% chance of falsely rejecting the null hypothesis, concluding that there is evidence of inequality between test modes. When multiple tests are conducted, as is the case with this analysis (three for maths, three for reading, six for science), it is common to correct the alpha level according to the number of tests undertaken to avoid alpha accumulation. For example, the Bonferroni technique, would reduce the alpha level to reduce the chances of a type 1 error across all the tests. However, doing so in the context of this study would also help make more inferences supporting score interpretation equality between modes. As such, methods for correcting for alpha accumulation (such as the Bonferroni technique) will only further support the null hypothesis. As a consequence, alpha could be increased to 0.1 meaning the chance of a type 1 error increases, but for this study, results in a more conservative approach to inferring test score interpretation and its equivalence between the modes. As such, these competing priorities for both reducing alpha to avoid alpha accumulation, and increasing alpha to have a more conservative approach to equivalence testing, counteract one another. Therefore, we decided to use an alpha level of 0.1 in this study.

Results

Explaining item difficulties by facet categories

The first research question focuses on establishing that there is a relationship between the facets proposed by the assessment framework and the estimated item difficulties using the GPCM model outlined in Eq. (1) and scaled to the IRT scales using Eq. (2). For mathematics, the facets explained item difficulty with an $R^2 = 0.30$, $F(8,59) = 3.19$, $p = 0.004$. This indicates that 30% of the variation in item difficulty can be explained by the facets of the assessment framework for mathematics. Furthermore, the p-value indicates that the overall the explanation of variance is statistically significant. For reading, the facets predicted the item difficulty with an $R^2 = 0.14$, $F(8,60) = 1.20$, $p = 0.310$. The *p*-value indicates that there is no significant relationship between the reading framework and item difficulty. While the analysis for the reading framework is undertaken in subsequent steps for completeness, it needs to be prefaced that no strong conclusions should be drawn about test score interpretation in the reading domain, given the regression results. For the final domain, science, the framework predicted the item difficulty with an $R^2 = 0.31$, $F(14,59) = 1.92$, $p = 0.041$. This indicates a link between the assessment framework's facets and item difficulty, where 31% of the variation in item difficulties can be explained by the variation in the assessment framework. As such, there is reasonable evidence for a link between item difficulties and facet categories for mathematics and science, but not for reading.

Mathematics facets

The mathematics domain contains three facets relating to content, situation and context, and the cognitive processes expected to be used by students in responding to the items. The results presented in Table 3 show the estimated mean item difficulty for each

Table 3 Mean facet level item difficulty across mathematics facets and levels

Facet	Levels	All Items			Mode invariant items		
		$\bar{b}_{PBA}$	$\bar{b}_{CBA}$	Wald test χ^2 (N = 11960)	$\bar{b}_{PBA}$	$\bar{b}_{CBA}$	Wald test χ^2 (N = 11950)[a]
Content	Space and shape	0.52 (0.04)	0.90 (0.04)	χ^2 (3) = 7.19*	0.85 (0.04)	1.17 (0.05)	χ^2 (3) = 11.94*
	Quantity	−0.68 (0.07)	−0.42 (0.07)	p = .07	−0.61 (0.18)	−0.40 (0.12)	p = .01
	Change and relationships	0.13 (0.04)	0.39 (0.07)		0.07 (0.04)	0.37 (0.07)	
	Uncertainty and data	0.05 (0.04)	0.33 (0.03)		−0.11 (0.05)	0.00 (0.05)	
Situation and context	Personal	−0.78 (0.12)	−0.42 (0.14)	χ^2 (3) = 5.17	−0.64 (0.18)	−0.22 (0.16)	χ^2 (3) = 7.21*
	Scientific	0.30 (0.03)	0.55 (0.04)	p = .13	0.30 (0.04)	0.57 (0.04)	p = .07
	Societal	−0.08 (0.04)	0.24 (0.04)		0.06 (0.04)	0.21 (0.04)	
	Occupational	0.29 (0.04)	0.54 (0.04)		0.66 (0.04)	0.81 (0.05)	
Process	Interpreting, applying and evaluating mathematical outcomes	−0.56 (0.04)	−0.29 (0.04)	χ^2 (2) = 3.83	−0.70 (0.04)	−0.56 (0.05)	χ^2 (2) = 2.99
	Employing mathematical concepts, facts and procedures	−0.10 (0.05)	0.25 (0.06)	p = .15	0.14 (0.09)	0.48 (0.09)	p = .22
	Formulating situations mathematically	0.63 (0.04)	0.86 (0.03)		0.76 (0.04)	0.95 (0.04)	

[*] Significant difference between PBA and CBA variation in estimated mean item difficulty within facet

[a] Sample sizes between all items and mode invariant items vary due to limited cases where students recorded no responses across the mode invariant items

item facet category and by the two assessment modes. For each facet, Fig. 2 is added to assist in the visual inspection of the pattern for each facet by mode, which depicts the comparability of score interpretation. A Wald test statistic and *p* value are presented for each facet, indicating whether there is a significant variation between the PBA and CBA assessment modes in how the estimated mean item difficulties differ across facet categories. The table and the figure also present the results for those items deemed as 'mode invariant' as classified by the OECD mode effects analysis (OECD, 2017a, Annex A).

Visually inspecting Fig. 2, the first important observation is that in all instances, the estimated mean facet difficulties for CBA are consistently larger than the PBA means. This result aligns with previous research on mode effects, where CBA was found to be more difficult than PBA. The results for the *content* facet indicate that, on average, items relating to *space and shape* are the most difficult for test takers, while items on quantity are the least difficult. This applies to both the PBA and CBA modes. Figure 2 shows no substantial difference in the pattern of the PBA and CBA items' difficulty. However, the magnitude of the differences between the facet categories resulted in a significant Wald test statistic, indicating that there is evidence of a difference between modes in how difficulty varied across facet categories.

For the *situation and context* facet, the results show that items with a personal context are the least difficult, compared to the other facet categories. The resulting Wald test statistic indicates there is insufficient evidence for variation between modes in terms of differences between facet categories' difficulty. The final item facet, *process*, shows that items requiring test takers to undertake *formulating situations mathematically* are the

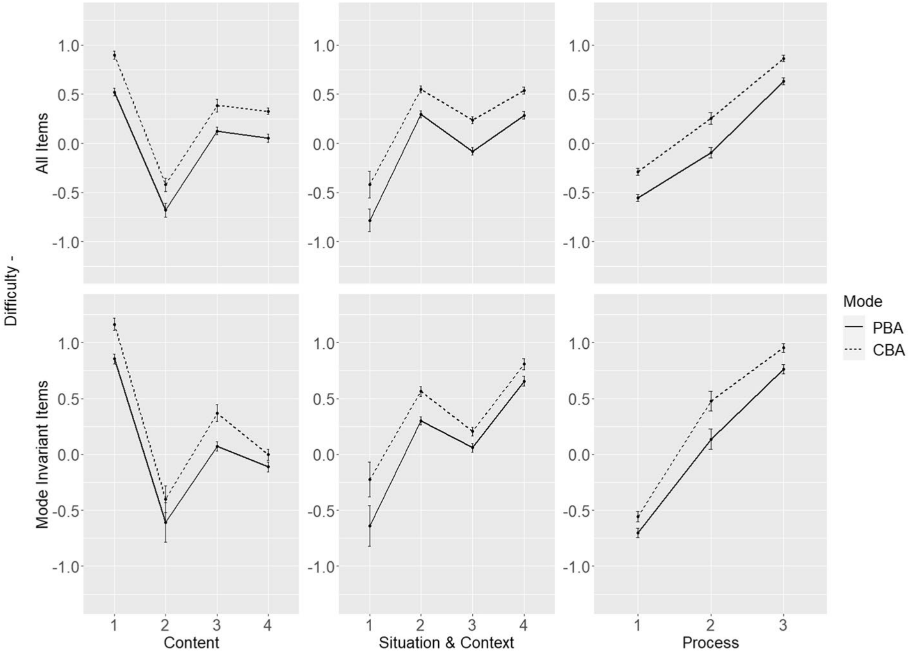

Fig. 2 Mean item difficulty for each item facet category by mode for all items and for mode invariant items in mathematics (Content Categories: 1 = Space and shape; 2 = Quantity; 3 = Change and relationships; 4 = Uncertainty and data; Situation and Context Categories: 1 = Personal; 2 = Scientific; 3 = Societal; 4 = Occupational; Process Categories: 1 = Interpreting, applying and evaluating mathematical outcomes; 2 = Employing mathematical concepts, facts and procedures; 3 = Formulating situations mathematically)

most difficult, relative to the other two facet categories. The Wald test indicates there was no statistically significant difference between modes in how difficulty varies across facet categories.

When analysing only the mode invariant items, the Wald test indicated that both the *content* facet and the *situation and context* facets had a significant difference between modes in the variation of estimated mean difficulties. The difference in the representations is shown in Fig. 2, with an obvious change in estimated item difficulty for facet category 4. For the *situation and context* facet, the difference occurs in the slope between category 2 and category 3, with CBA having a steeper line than the PBA mode. f. The Wald test results indicates a significant difference between the PBA and CBA test modes. Reading Facets.

The reading domain consisted of three facets in 2015, classifying items by *situation*, *text format,* and *aspect.* Aspect relates to the underlying cognitive processes that test takers are expected to utilise in answering items. The results are presented in Table 4 with the mean item difficulty by facet category shown in Fig. 3. Initial inspection confirms that across all instances, the CBA mode of the test is more difficult than the PBA mode.

For the *situation* facet, the key feature associated with the pattern in Fig. 3 is that for the PBA test, category 4 is on average more difficult than the category 3. However, for the CBA items, this relationship is reversed, with the average difficulty of category 3 being more difficult than category 4. The Wald test indicates that there is a significant difference in the variation of difficulties between modes.

For the second facet, *text format*, items with a mixed text format are the most difficult for test takers in both modes, with a mean estimated difficulty of -0.07 in the PBA mode, and 0.43 in the CBA mode. Figure 3 shows that in the CBA facet

Table 4 Mean facet level item Difficulty across reading facets and levels

Facet	Level	All items			Mode invariant items		
		$\bar{b}_{PBA}$	$\bar{b}_{CBA}$	Wald test χ^2 (N = 11795)	$\bar{b}_{PBA}$	$\bar{b}_{CBA}$	Wald test χ^2 (N = 11795)
Situation	Public	−0.16 (0.04)	0.10 (0.04)	χ^2 (3) = 34.34*	−0.44 (0.05)	−0.35 (0.04)	χ^2 (3) = 9.16*
	Personal	−0.55 (0.04)	−0.24 (0.04)	p < 0.01	−0.69 (0.04)	−0.43 (0.04)	p = 0.03
	Educational	−0.72 (0.05)	−0.31 (0.06)		−0.39 (0.07)	−0.14 (0.12)	
	Occupational	−0.65 (0.04)	−0.57 (0.05)		−0.78 (0.05)	−0.62 (0.07)	
Text format	Continuous	−0.55 (0.04)	−0.30 (0.04)	χ^2 (3) = 26.15*	−0.59 (0.04)	−0.36 (0.06)	χ^2 (3) = 12.93*
	Mixed	−0.07 (0.04)	0.43 (0.05)	p < 0.01	0.00 (0.05)	0.36 (0.10)	p < 0.01
	Non-continuous	−0.63 (0.04)	−0.31 (0.04)		−0.69 (0.05)	−0.59 (0.05)	
	Multiple	−0.99 (0.05)	−0.56 (0.04)		−1.07 (0.06)	−0.73 (0.06)	
Aspect	Integrate and interpret	−0.47 (0.04)	−0.16 (0.05)	χ^2 (2) = 15.86*	−0.42 (0.05)	−0.15 (0.09)	χ^2 (2) = 1.76
	Reflect and evaluate	−0.45 (0.04)	−0.26 (0.04)	p < 0.01	−0.39 (0.04)	−0.22 (0.04)	p = 0.41
	Access and retrieve	−0.81 (0.05)	−0.44 (0.04)		−1.17 (0.07)	−1.04 (0.05)	

* Significant difference between PBA and CBA variation in estimated mean item difficulty within facet

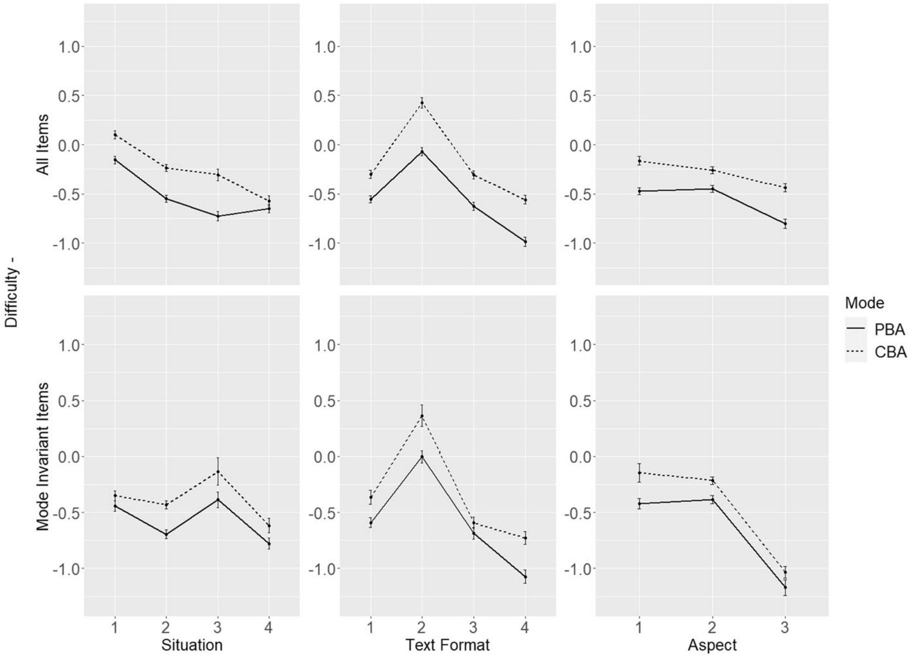

Fig. 3 Mean item difficulty for each item facet category by mode for all items and for mode invariant items in reading (Situation categories: 1 = Public; 2 = Personal; 3 = Educational; 4 = Occupational; Text format categories: 1 = Continuous; 2 = Mixed; 3 = Non-continuous; 4 = Multiple; Aspect categories: 1 = Integrate and interpret; 2 = Reflect and evaluate; 3 = Access and retrieve)

categories, the peak associated with the category 2 is steeper and more pronounced than in the PBA items. Again, the Wald test statistic indicates that there is a significant difference between the two modes as to how the mean estimated difficulties vary across facet categories.

For the final facet, *aspect*, items that require test takers to access and retrieve information are found to be the least difficult to complete in both modes, with a mean estimated difficulty of −0.81 in the PBA mode, and −0.44 in the CBA mode. Figure 3 for the PBA items is slightly increasing from category 1 to 2, while in the CBA items, it is decreasing from category 1 to 2. The Wald test statistic confirms there is a significant difference between the two modes in the variation of mean difficulties. This means all three facets in the reading domain, when using all items, are showing a significant variation between modes in the variation of estimated mean difficulties. For the mode invariant only items, Fig. 3 for the *situation* and *text format* facets show clear differences between the pattern of estimated difficulties. For *situation*, the PBA facet categories have a larger decrease from category 1 to 2 compared to the CBA categories. For the *text format* facet, the mode invariant CBA items have a greater decrease in difficulty from category 3 to 4 when compared to the PBA facet categories. The Wald test statistic indicates that this variation between the two modes is significant with a p value of less than 0.01. The final facet *aspect* however shows that for the mode invariant items, there is now no significant difference in the variation of the estimated means between modes.

Science facets

The science domain was the major domain in 2015. The PISA assessment framework consists of six facets, being: two different dimensions of *context*; *competency*; *knowledge requirements*; *scientific system*; *depth of knowledge* deemed necessary to respond to items. The results for all six facets are presented in Table 5 , and Fig. 4a and b respectively. When analysing all science items by facets, four facets showed significant cross-mode variations in how mean difficulties differ between facet categories. These are the two *context* facets, the *system* facet, and the *competency* facet.

The arrangement of items according to *context 1* (with three categories) indicates a significant difference in how the estimated means are represented between the two modes. Inspecting Fig. 4a, the key differences between the PBA and CBA representation occurs

Table 5 Mean facet level item difficulty across science facets and levels

Facet	Levels	All items			Mode invariant items		
		$\bar{b}_{PBA}$	$\bar{b}_{CBA}$	Wald test χ^2 (N = 11861)	$\bar{b}_{PBA}$	$\bar{b}_{CBA}$	Wald test χ^2 (N = 11859)[a]
Context 1 (2015)	Personal	−0.06 (0.08)	0.10 (0.07)	χ^2 (2) = 5.85*	−0.42 (0.08)	−0.29 (0.07)	χ^2 (2) = 4.4
	Local/National	−0.38 (0.03)	−0.11 (0.03)	p = .05	−0.30 (0.03)	−0.11 (0.03)	p = .11
	Global	0.11 (0.03)	0.43 (0.03)		0.09 (0.04)	0.35 (0.03)	
Context 2 (2015)	Health and disease	−0.37 (0.04)	−0.21 (0.03)	χ^2 (4) = 24.40*	−0.31 (0.04)	−0.16 (0.04)	χ^2 (4) = 3.99
	Natural resources	−0.52 (0.04)	−0.18 (0.03)	p < .01	−0.35 (0.04)	−0.13 (0.03)	p = .41
	Frontiers	0.15 (0.04)	0.40 (0.04)		−0.14 (0.04)	0.06 (0.04)	
	Hazards	−0.76 (0.06)	−0.53 (0.05)		−0.78 (0.07)	−0.63 (0.05)	
	Environmental quality	−0.17 (0.04)	0.14 (0.03)		0.12 (0.04)	0.35 (0.04)	
System	Living	−0.06 (0.04)	0.18 (0.03)	χ^2 (2) = 8.97*	−0.1 (0.03)	0.07 (0.03)	χ^2 (2) = 5.24*
	Earth and space	−0.43 (0.04)	−0.07 (0.03)	p = .01	−0.47 (0.04)	−0.23 (0.03)	p = .07
	Physical	−0.35 (0.04)	−0.12 (0.04)		−0.36 (0.05)	−0.16 (0.05)	
Competency	Interpret data & evidence scientifically	−0.22 (0.04)	0.08 (0.03)	χ^2 (2) = 11.06*	−0.19 (0.03)	0.00 (0.03)	χ2 (2) = 4.02
	Evaluate and design scientific enquiry	−0.30 (0.05)	0.06 (0.04)	p < .01	−0.12 (0.05)	0.17 (0.05)	p=.13
	Explain phenomena scientifically	−0.22 (0.04)	−0.03 (0.03)		−0.36 (0.04)	−0.21 (0.04)	
Knowledge	Procedural	−0.30 (0.04)	−0.04 (0.04)	χ^2 (2) = 4.39	−0.1 (0.04)	0.08 (0.04)	χ2 (2) = 8.37*
	Content	−0.24 (0.04)	0.00 (0.03)	p = .11	−0.33 (0.04)	−0.17 (0.04)	p=.02
	Epistemic	−0.01 (0.05)	0.35 (0.04)		−0.27 (0.06)	0.07 (0.04)	
Depth of Knowledge	Low	−0.37 (0.04)	−0.12 (0.04)	χ^2 (2) = 3.42	−0.49 (0.05)	−0.32 (0.05)	χ2 (2) = 1.35
	Medium	−0.23 (0.03)	0.05 (0.03)	p = .18	−0.26 (0.03)	−0.04 (0.03)	p=.51
	High	0.30 (0.05)	0.47 (0.05)		0.30 (0.05)	0.49 (0.06)	

[*] Significant difference between PBA and CBA variation in estimated mean item difficulty within facet

[a] Sample sizes between all items and mode invariant items vary due to limited cases where students recorded no responses across the mode invariant items

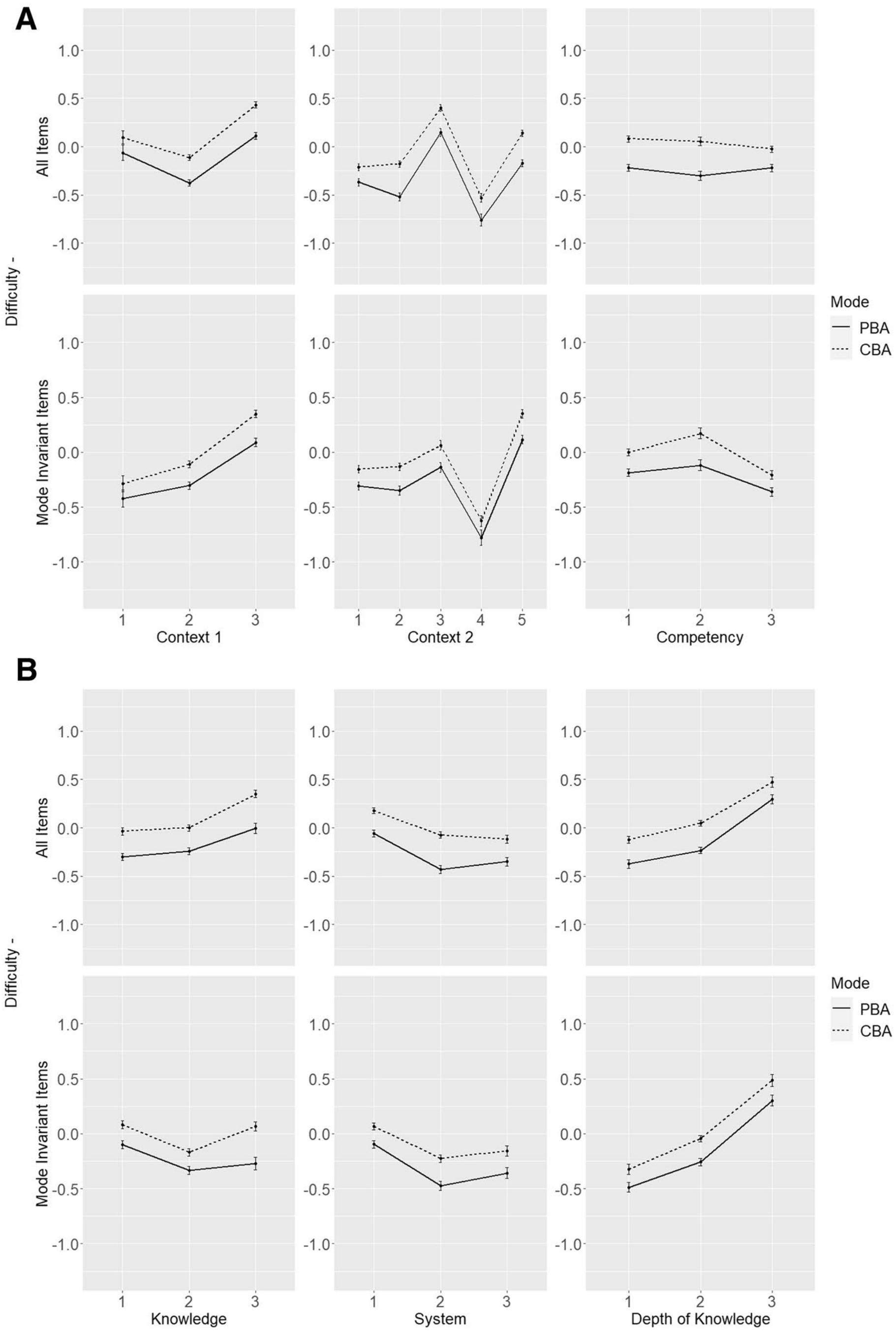

Fig. 4 Mean item difficulty for each item facet category by mode for all items and for mode invariant items in science (Context 1 categories: 1 = Personal; 2 = Local/national; 3 = Global; Context 2 categories: 1 = Health and disease; 2 = Natural resources; 3 = Frontiers; 4 = Hazards; 5 = Environmental quality; Competency categories: 1 = Interpret data and evidence scientifically; 2 = Evaluate and design scientific enquiry; 3 = Explain phenomena scientifically; Knowledge categories: 1 = Procedural; 2 = Content; 3 = Epistemic; System categories: 1 = Living; 2 = Earth and space; 3 = Physical; Depth of knowledge categories: 1 = Low; 2 = Medium; 3 = High)

in the difference between the category 1 and category 2, where the slope of the line for CBA is not as steep between these categories compared to the PBA slope.

In *context 2* (with five categories), there is a significant difference between the PBA and CBA assessment modes in how the estimated mean item difficulties differ across facet categories. For the PBA representation, the category 2 is on average less difficult that the category 1. However, in the CBA representation, this relationship is reversed with the category 1 now more difficult than category 2. The Wald test indicates a significant difference in the representation of the two modes.

For the *competency* (Fig. 4a) and *system* (Fig. 4b) facets the figures descriptively show clear differences in how the facets are represented. For *competency*, the estimated mean difficulties for category 3 is higher than category 2, while the CBA items, category 2 is higher than category 3. For the *system* facet, the PBA line shows category 2 lower than category 3, while for CBA category 2 is higher than category 3.. The Wald test statistics in both facets confirm there is a significant variation in how the estimated mean difficulties relate to the facets by mode. For the *knowledge* and *depth of knowledge* facets, there is no indication that the mean estimated difficulty within the facets varies between the two modes.

When analysing the facets using only the mode invariant items, the Wald test statistic indicates two facets with a significant variation between modes. These are the *system* and *knowledge* facets. In Fig. 4b, the mode invariant items for the *knowledge* facet show that for PBA, the change between category 2 and 3 is relatively small when compared to the same change in the CBA mode. For mode invariant items in the *system* facet, the visual inspection is less clear, however the Wald test results indicate that the pattern representations between the two modes are significantly different.

Discussion

Using data from 13 participating countries in PISA 2015, we compared the score interpretation across modes in the domains of mathematics, reading and science. As a first preparatory step we addressed research question one of whether the facets proposed by the assessment framework actually explain item difficulty. The results showed for the mathematics domain, there was a clear link between the item difficulties and the PISA assessment framework as indicated by the portion of variance explained by facets (substantial effect size). For the reading domain however, there was less evidence to establish a link as indicated by the moderate portion of explained variance. Finally, as for maths, for science there was a clear link between the item facets included in the assessment framework and item difficulty (substantial effect size). The relative weak relation between item facets and item difficulties in reading suggests, that there are other item characteristics not included in the present study determining item difficulty. This in turn limits the conclusiveness of our results on the cross-mode comparability of reading score interpretation.

To address research questions two and three, estimated item difficulties were used to derive item facet category means corresponding to the PISA 2015 assessment framework. Differencing the facet category mean difficulties within each mode represents the score interpretation for each mode, which was tested to see if there was a significant difference between PBA and CBA. The results across all three domains showed some significant difference between the PBA and CBA modes in how the mean difficulty varied across the facet categories.

For the maths domain, the findings suggest that the maths test score interpretation with all items is similar across modes except for the influence of the *content* facet on the test score. When using mode invariant items comparability is reduced in that both *content* and *situation and context* differ in the variation of estimated difficulty means for each of the facet categories between the two modes. Notably, for the *process* facet in mathematics, there is no evidence that the test score interpretation varies by mode when using all items or mode invariant items.

For the reading domain, results showed that when using all items for analysis, all three facets had a significant difference in how difficulty varies across facet categories between modes. This suggests differences in how the test scores are interpreted for PBA and CBA modes. Using mode invariant items only, *situation* and *text format* facets both indicated that there is a significant difference in test score interpretation.

Finally, the findings for the science domain with all items indicate that mode effects have a significant impact on both the *context* facets, the *system* facet, and the *competency* facet. This means there was a significant variation in how mean item difficulties are distributed between the facet categories and therefore differences in how the test scores are interpreted. For two facets, there was not significant difference. Using mode invariant items only, there was a significant difference between modes only in the *system* and *knowledge* facets. This means that the comparability of score interpretation in science is in particular affected when all items are used.

In summary, in all domains there was at least on item facet showing significant differences between modes in the obtained difficulty pattern suggesting gradual differences in test score interpretation between modes. In particular for reading, almost all facets showed significant differences. This applies to both when all items are used for analysis, and also when only items are used deemed to be mode invariant. There was no clear picture in the way that using only items deemed as mode invariant in a domain increased comparability in terms of more facets showing no significant difference. The visual inspection of the difficulty patterns across facet categories and between modes provides a clearer picture. This descriptive interpretation confirms that there are cross-mode differences, especially for reading, while the difficulty patterns for math and science were fairly consistent across modes. That is the slopes representing differences between adjacent facet categories were mostly parallel and had the same direction, respectively.

Limitations

A limitation of this study is that due to the limited number of responses within each country, a pooled approach to modelling the data was taken. Despite incorporating countries as strata weights, it could be expected that there is some variation in the mode effect between countries. This study was limited in the ability to account for country-specific mode effects.

Another limitation of the study is the strength of the evidence that can be attained from the construct representation approach given the item characteristics (item facets) provided by the PISA assessment framework. In particular, we refrained from examining and comparing construct validity and limited ourselves to comparing score interpretation because construct validation would require more theoretical grounding of the facets of the items. This refers to the justification of each facet, the completeness of construct-relevant facets, and assignment of items to facets based. Therefore, a theory-based task analysis is required identifying information processing factors and relating them to item characteristics. Thus, additional supporting evidence would be required to make stronger assertions that the underlying constructs defined by the PISA assessment framework have (not) changed as a result of the change in test mode.

Another area for future research focuses the modelling approach. In the present study, for each facet a model was estimated to limit model complexity. However, modelling item facets simultaneously would be a valuable extension, as, for instance, it would allow to investigate (theoretically relevant) interactions between item facets.

Conclusions

The present study shows that the mode effects on difficulty vary within some of the item facets proposed by the PISA assessment framework, in particular for reading. The obtained findings are based on the construct representation approach relating item facets to item difficulty, and shed light on whether the comparability of score interpretation between modes is given. Thus, the present study adds a new approach and empirical findings to the investigation of the cross-mode equivalence in PISA domains. In particular, it extends previous research that focused on mode effects on item parameters in terms of the equivalence of interpretation of the test scores, which is crucial for maintaining the trend.

Acknowledgements

Not applicable.

Author contributions

SH, UK, and FG developed the concept of the study. SH wrote the first draft of the manuscript, along with undertaking the analysis. UK, FG and AR refined the analytical methods for the modelling. UK and FG refined the discussion. All authors contributed to the editing of the paper and the development of the final manuscript. All authors have read and approved the final manuscript.

Funding

This research was funded by the Centre for International Student Assessment (ZIB).

Availability of data and materials

The data that support the findings of this study is from multiple third parties and restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Due to an anonymity clause in the data processing agreements used to obtain the data, the sources of the data cannot be publicly disclosed yet.

Declarations

Ethics approval and consent to participate

Ethics approval for this work was not required as it uses secondary data analysis.

Consent for publication

We the authors consent to this original work being published upon acceptance of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 13 May 2020 Accepted: 3 March 2023
Published online: 11 March 2023

References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). John Wiley & Sons.

American Educational Research Association American Psychological Association National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Asparouhov, T., & Muthén, B. (2020). IRT in Mplus (Technical Report Version 4). *www.statmodel.com*. https://www.statmodel.com/download/MplusIRT.pdf

Bennett, R. E., Braswell, J., Oranje, A., Sandene, B., Kaplan, B., & 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), 4–38.

Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes. *Psychological Test and Assessment Modeling, 58*(4), 597–616.

Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation, 62*, 1–9. https://doi.org/10.1016/j.stueduc.2019.04.005

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179–197. https://doi.org/10.1037/0033-2909.93.1.179

Feskens, R., Fox, J.-P., & Zwitser, R. (2019). Differential item functioning in PISA due to mode effects. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 231–247). Springer International Publishing.

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-Scale Assessments in Education, 6*(1), 11. https://doi.org/10.1186/s40536-018-0064-z

Hartig, J., Frey, A., Nold, G., & Klieme, E. (2012). An application of explanatory item response modeling for model-based proficiency scaling. *Educational and Psychological Measurement, 72*(4), 665–686. https://doi.org/10.1177/0013164411430707

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 189–220). Praeger Publ.

Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice, 20*(3), 16–25. https://doi.org/10.1111/j.1745-3992.2001.tb00066.x

International Test Commission. (2005). International guidelines on computer-based and internet delivered testing. *International Test Commission (ITC)*. https://www.intestcom.org/files/guideline_computer_based_testing.pdf

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: How big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476–493. https://doi.org/10.1080/03054985.2018.1430025

Kind, P. M. (2013), Conceptualizing the Science Curriculum: 40 Years of Developing Assessment Frameworks in Three Large-Scale Assessments. Science Education, 97(5), 671-694. https://doi.org/10.1002/sce.21070

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. https://doi.org/10.1080/08957340802558326

Kroehne, U., Buerger, S., Hahnel, C., & Goldhammer, F. (2019). Construct equivalence of PISA reading comprehension measured with paper-based and computer-based assessments. *Educational Measurement: Issues and Practice, 38*, 97–111. https://doi.org/10.1111/emip.12280

Kröhne, U., & Martens, T. (2011). 11 Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift Für Erziehungswissenschaft, 14*(2), 169. https://doi.org/10.1007/s11618-011-0185-4

Mazzeo, J., & von Davier, M. (2008). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. Education Working Papers EDU/PISA/GB (2008), 28, 23-24.

Mullis, I. V., & Martin, M. O. (2019). *PIRLS 2021 assessment frameworks*. ERIC.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *ETS Research Report Series, 1992*(1), i–30. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Muthén, L. K., & Muthén, B. O. (2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

OECD. (2013). *Technical report of the survey of adult skills (PIAAC)*. OECD Publishing.

OECD. (2016). *Annex A6. In PISA 2015 results (volume I)*. OECD Publishing. https://doi.org/10.1787/9789264266490-en

OECD. (2017a). *PISA 2015 Technical Report*. https://www.oecd.org/pisa/data/2015-technical-report/

OECD. (2017b). *PISA 2015 assessment and analytical Framework: Science*. OECD. https://doi.org/10.1787/9789264281820-en

Pommerich, M. (2004). Developing Computerized Versions of Paper-and-Pencil Tests: Mode Effects for Passage-Based Tests. *The Journal of Technology, Learning and Assessment*, *2*(6). https://ejournals.bc.edu/index.php/jtla/article/view/1666

Robitzsch, A., Lüdtke, O., Goldhammer, F., Kroehne, U., & Köller, O. (2020). Reanalysis of the German PISA data: A comparison of different approaches for trend estimation with a particular emphasis on mode effects. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2020.00884

Stacey, K., & Turner, R. (2015). The evolution and key concepts of the PISA mathematics frameworks. In K. Stacey & R. Turner (Eds.), *Assessing mathematical literacy* (pp. 5–33). Springer.

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2008). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments: A meta-analysis of testing mode effects. *Educational and Psychological Measurement, 68*(1), 5–24. https://doi.org/10.1177/0013164407305592

Publisher's Note