Psychological Test and Assessment Modeling, Volume 62, 2020 (1), 55-75

PISA reading: Mode effects unveiled in short text responses

Fabian Zehner[1], *Ulf Kroehne*[1], *Carolin Hahnel*[1,2], *& Frank Goldhammer*[1,2]

Abstract

Educational large-scale assessments risk their temporal comparability when shifting from paper- to computer-based assessment. A recent study showed how text responses have altered alongside PISA's mode change, indicating mode effects. Uncertainty remained, however, because it compared students from 2012 and 2015. We aimed at reproducing the findings in an experimental setting, in which $n = 836$ students answered PISA reading questions on computer, paper, or both. Text response features for information quantity and relevance were extracted automatically. Results show a comprehensive recovery of findings. Students incorporated more information into their text responses on computer than on paper, with some items being more affected than others. Regarding information relevance, we found less mode effect variance across items than the original study. Hints for a relationship between mode effect and gender across items could be reproduced. The study demonstrates the stability of linguistic feature extraction from text responses.

Keywords: Computer-based assessment, paper-based assessment, open-ended text responses, mode effect, automatic processing

[1] *Correspondence concerning this article should be addressed to:* DIPF | Leibniz Institute for Research and Information in Education, Rostocker Str. 6, 60323 Frankfurt am Main, Germany;
email: fabian.zehner@dipf.de

[2] Centre for International Student Assessment (ZIB)

Consistent comparability within and across points of assessments is the very essence of international educational large-scale assessments, such as the *Programme for International Student Assessment* (PISA; OECD, 2017a) or the *Progress in International Reading Literacy Study* (PIRLS; Mullis, Martin, Foy, & Drucker, 2012). Only then, researchers and policy-makers can draw legitimate inferences from temporal or subgroup comparisons. In order to check for PISA's continuing temporal consistency, several studies have investigated the effect of PISA's administration mode change on test or item scores (i.a., OECD, 2017b; Jerrim, Micklewright, Heine, Salzer, & McKeown, 2018). The present study adds a new perspective to this by analyzing the mode effect on short text responses on PISA's reading test.

With the recent tendency of shifting from paper- to computer-based assessment, the respective study programmes need to take into account a new potential source for measurement invariance of their temporal trends. In this context, constructed responses are vastly neglected. However, they provide rich indicators for potential mode effects. If text responses differ substantively in aspects germane to the response process, the assessed latent construct might have shifted (Zehner, Goldhammer, Lubaway, & Sälzer, 2019). For example, if we observe that test takers incorporate more pieces of information into their response, they may have carried out different cognitive operations. This can (but does not need to) lead to assessing a shifted construct. In contrast to this perspective, mode effect studies typically investigate scores (e.g., Buerger, Kroehne, & Goldhammer, 2016; Choi & Tinkler, 2002; Clariana & Wallace, 2002; Wang, Jiao, Young, Brooks, & Olson, 2007) or process data (e.g., Kroehne, Hahnel, & Goldhammer, 2019; Piaw, 2011; White, Kim, Chen, & Liu, 2015).

Beyond manual scoring, the fundamental reason that linguistic information buried in text responses has played an inferior role for research so far is their limited accessibility. With the rise of automatic scoring systems (for an overview cf. Burrows, Gurevych, & Stein, 2014 and Galhardi & Brancher, 2018), however, natural language processing techniques now allow information extraction from text data at a large scale for diverse purposes.

A previous study (Zehner et al., 2019) revealed substantial differences between short text responses written (a) by students who responded to the *paper*-based PISA 2012 reading test and those written (b) by students in the *computer*-based PISA 2015 reading test. Indicating mode effects, this analysis was based on a natural experiment and investigated the differences occurring together with the mode change in two different cohorts. In the present study, we experimentally manipulated the administration mode of the PISA reading test in a randomized design. Therefore, it allows drawing causal inferences with respect to the administration mode's effect on text responses. Additionally, we aimed at reproducing the findings of Zehner et al. (2019).

For this, we investigated data from a German add-on study during PISA 2012 that randomly assigned students to computer- or paper-based assessment, or both. From the open-ended short text responses, we extracted the same linguistic indicators like the study that was to be reproduced: the *Proposition Entity Count* (information quantity) and *Relevance Proportion* (information relevance). For doing so, we utilized baseline natural language processing techniques and compared the characteristics of responses in both modes.

Theoretical context

Before diving into theory, let us engage in a thought experiment about why text responses were an attractive observational unit to study mode effects. Assume we have an open-ended item asking for the purpose of a given stimulus text. The scoring allows different lines of reasoning. For full credit, test takers can refer to the decorative pictures *or* the text semantics, suggesting different conclusions. In such cases, it is quite common that the majority of test takers uses the same line of reasoning. Let us now change from paper- to computer-based assessment. It is possible that the pictures suddenly attract more attention because they are brighter on screen than on paper and reading habits differ between screen and paper (Delgado, Vargas, Ackerman, & Salmerón, 2018). More responses might now refer to the pictures. Thus, the mode change could lead to a change of the dominating line of reasoning, and most test takers might carry out different cognitive operations than before. While these operations would still be part of the overall construct, the construct's operationalization could have changed. Analyses based on the plain score would not be able to identify such a mode effect if it did not affect typically investigated measurement properties such as item difficulty. While text responses are not the new panacea for studying mode effects, they provide a rich set of new data points for doing so.

Mode effects and PISA

Mode effects address differences between assessed latent constructs being measured by two implementations of the same test (Kroehne & Martens, 2011). In 2015, PISA introduced computer-based assessment as the main mode for the first time. It is now the question as to whether data from earlier PISA assessments can be directly compared. Typically, this is studied at the level of test and item scores (e.g., Buerger et al., 2016; Choi & Tinkler, 2002; Clariana & Wallace, 2002; Jerrim et al., 2018; Kroehne, Buerger, Hahnel, & Goldhammer, 2019; Robitzsch et al., 2017; OECD, 2017b; Wang et al., 2007) or process data (e.g., Kroehne, Hahnel, & Goldhammer, 2019; Piaw, 2011; White et al., 2015). Compared to PIRLS–where relevant mode effects were found and their estimation was directly incorporated into the study design (Fishbein, Martin, Mullis, & Foy, 2018)–PISA's initiator, the *Organisation for Economic Co-Operation and Development* (OECD), acknowledges the possibility of mode effects in a subset of items in their assessment, though, assuming a sufficient number of invariant items (OECD, 2016). This was concluded from a between-subjects mode effect study on the international level in PISA's field trial that identified some items requiring mode-specific parameters for the scaling model (OECD, 2017b). In addition, the OECD did not find differential effects between subgroups (e.g., gender), but only investigated these on the basis of field trial data (OECD, 2016). Robitzsch et al. (2017) found small overall mode effects in the German PISA Field Trial 2015. Zehner et al. (2019) demonstrated that this is mirrored in text responses on the aggregate level, but even larger differences could be found on the item level.

It is important to note that, for linking scales, every new instrument implementation has to be checked for equivalence with respect to relevant criteria (Kolen & Brennan, 2014). However, for assessing reading, there seems to be the general trend that computer-based tests are harder and completed more quickly than paper-based ones (Kolen & Brennan, 2014; Kroehne, Buerger, et al., 2019; Kroehne, Hahnel, & Goldhammer, 2019; Robitzsch et al., 2017). The differences might stem from several components that can be coarsely distinguished in two categories: properties of the test administration and test taker characteristics (Kroehne & Martens, 2011). We only name a few examples. In a computer-based reading assessment, reading takes place on screen, which can be impacted by screen resolution (Bridgeman, Lennon, & Jackenthal, 2001), and involves computer navigation, which can be challenging to varying degrees for different test takers (Wang et al., 2007). A recent meta-analysis found no differences in reading speed, but comprehension was better on paper (Kong, Seo, & Zhai, 2018). Traditionally, reading on screen was assumed to be slower (Noyes & Garland, 2008). While paper-based assessment uses handwriting for responses, the students need to write via keyboard on the computer. Finally, writing speed and fluency are only moderately related across these input modes (Feng, Lindner, Ji, & Malatesha Joshi, 2017). Given all these complex mechanisms, the importance of verifying construct equivalence between a test in two different modes is evident.

Linguistic text response features

If we analyze text responses instead of scores for studying mode effects, we need to identify features of the written product that are indicative for the response process and that constitute outcomes related to the construct of interest. Based on cognitive theories, Zehner, Goldhammer, and Sälzer (2018) compiled a framework for text response features that are crucial to the response process in reading tasks. Briefly summarized, the central component is the situation model (Kintsch, 1998) that is mentally built when test takers read a stimulus text. It comprises a set of propositions that reflect the text base (micropropositions), but also address what is associated with or implied by the text base (macropropositions). In order to answer questions like those in the PISA assessment, the students identify the question focus and category (Graesser & Franklin, 1990). Next, they query memory structures, one of which is the just-built situation model, and winnow them down to propositions both relevant and compatible to the question focus and category (Graesser & Franklin, 1990). Determined by the question category, the selected propositions are then concatenated using linguistic structures in order to formulate the final text response (Graesser & Clark, 1985).

Zehner et al. (2019) derived two crucial linguistic features from this framework in order to investigate potential mode effects: *Proposition Entity Count* and *Relevance Proportion*. Note that both features have shortcomings which are critically discussed in the Limitations section.

Information quantity: the Proposition Entity Count (PEC)

The first feature Zehner et al. (2019) chose is called *Proposition Entity Count* (PEC; Zehner et al., 2018). It is rooted in Kintsch and van Dijk (1978) who show that more skilled readers reproduce more propositions. That is, it makes a construct-relevant difference whether a response contains more or less pieces of information. But how can we capture the amount of information? The most proper way would be to identify all propositions in the response. Unfortunately, this endeavor is not feasible due to technical constraints and conceptual issues. First, current natural language processing cannot handle texts with improper language use (i.a., grammar-wise; cf. Dzikovska, Nielsen, & Leacock, 2016; Higgins et al., 2014). With PISA being a low-stakes assessment, the data of the present study contains lots of informal, orthographically and grammatically improper, or hyper-concisely written responses. The second major problem is a conceptual one. Given their informal character, the responses regularly neglect conventions and comprise only single words instead of complete sentences, which typically does not occur in written essays but is comparatively common for short-text responses. Compared to written essays, short text responses can be very minimalistic. Hence, operationalizations have to take single words into account as well.

Therefore, PEC is a proxy that refers to single entities of propositions (Zehner et al., 2018). Hence, a proposition such as LIKE(READER, BOOK)–"the reader likes the book"–is split into its three entities {READER, LIKE, BOOK}. PEC is computed by counting those words in a response that are either nouns, pronouns, non-auxiliary verbs, adjectives, adverbs, or answer particles. These are words that are assumed to refer to genuine elements in the situation model and do not constitute language artifacts (Zehner et al., 2018).

Information relevance: Relevance Proportion (Rel)

While PEC captures the amount of information in the response, the second feature Zehner et al. (2019) chose (*Relevance Proportion*, Rel; Zehner et al., 2018) reflects how relevant this information is for providing a correct response. This feature refers to filtering the relevant propositions during the response process (sensu Graesser & Franklin, 1990, cf. section Linguistic text response features).

For assessing their relevance, the proposition entities are compared pairwise with proposition entities of correct responses in the coding guides. The highest semantic (cosine) similarity determines how close the response's proposition entity is to the one in the correct responses. If a proposition entity's relevance score is within the distribution's lower 25 percent, it is classified as irrelevant (as relevant otherwise). The ratio of relevant proposition entities and the total number of proposition entities constitutes the final measure *Relevance Proportion*. For measuring semantic similarity, one semantic vector space model is computed for each item using Latent Semantic Analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990). An external text corpus, such as Wikipedia, serves as input for the Latent Semantic Analysis.

Research questions and hypotheses

Aiming at reproducing Zehner et al. (2019), the analysis adheres to the same research questions and uses the same two response features. The research questions follow the idea that differences in the response features correspond to differences in the response process, hence, indicating a mode effect. Each research question addresses one of the two features `PEC` and `Rel` . The hypotheses imply a recovery of the results of Zehner et al. (2019).

P1|R1 **Response correctness effect.** We expect `PEC` (P1) and `Rel` (R1) to be higher in correct than in incorrect responses.

P2a|R2a **Mode effect.** Also in line with other previous findings (Horkay, Bennett, Allen, Kaplan, & Yan, 2006; White et al., 2015), we expect a higher `PEC` for responses from the computer-based than for those from the paper-based assessment. For `Rel`, we do not expect to find a mode effect at the aggregate level, which was a result but not the expectation of Zehner et al. (2019).

P2b|R2b **Item-specific mode effect.** Items differ in the kind of response processes they evoke. Thus, we expect the mode effect on `PEC` and on `Rel` to vary across items.

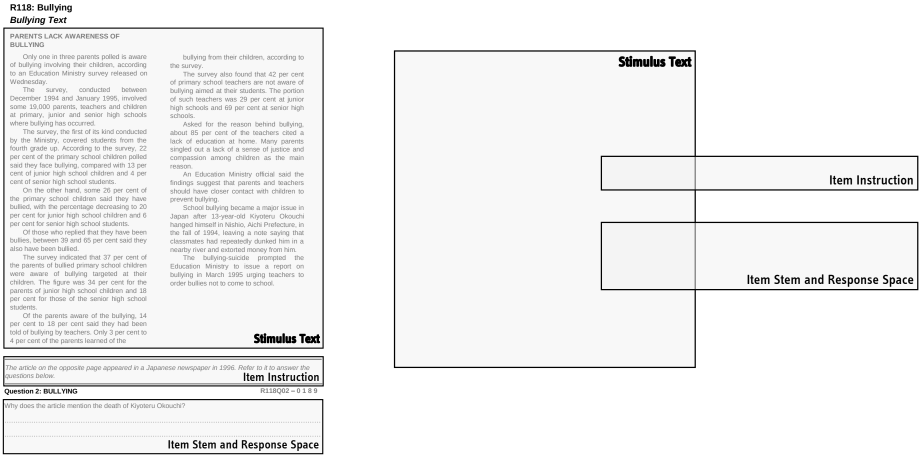
P3a|R3a **Gender effect.** We expect girls to show a higher `PEC` than boys (cf. Zehner et al., 2018). For `Rel`, we do not expect to find a gender effect on the aggregate level, which was a result but not the expectation of Zehner et al. (2019).

P3b|R3b **Gender-Specific mode effect.** Zehner et al. (2019) found a significant, but small overall interaction between mode and gender for `PEC`, but not for `Rel`. For both, `PEC` and `Rel`, the gender-specific mode effects are expected to vary across items.

Methods

Participants, procedure, and materials

Alongside PISA 2012 in Germany, an add-on study implemented an experimental variation of computer- and paper-based assessment in a randomized, balanced within- and between-subjects design (for detail, see Hahnel, Goldhammer, Naumann, & Kroehne, 2016). On a second day after the PISA test, $n = 880$ fifteen-year-olds responded to a set of in total 30 reading literacy items, resulting in $n = 7,963$ German text responses. The reading literacy items were organized in two fixed sets (*clusters*) that were administered either computer- or paper-based. While 49 percent of the students took one cluster either in the computer- or paper-based assessment, the others subsequently worked on both clusters in both modes. This allowed a direct cross-mode comparison of text re-



R118: Bullying

Bullying Text

PARENTS LACK AWARENESS OF BULLYING

Only one in three parents polled is aware of bullying involving their children, according to an Education Ministry survey released on Wednesday.

The survey, conducted between December 1994 and January 1995, involved some 19,000 parents, teachers and children at primary, junior and senior high schools where bullying has occurred.

The survey, the first of its kind conducted by the Ministry, covered students from the fourth grade up. According to the survey, 22 per cent of the primary school children polled said they face bullying, compared with 13 per cent of junior high school children and 4 per cent of senior high school students.

On the other hand, some 26 per cent of the primary school children said they have bullied, with the percentage decreasing to 20 per cent for junior high school children and 6 per cent for senior high school students.

Of those who replied that they have been bullies, between 39 and 65 per cent said they also have been bullied.

The survey indicated that 37 per cent of the parents of bullied primary school children were aware of bullying targeted at their children. The figure was 34 per cent for the parents of junior high school children and 18 per cent for those of the senior high school students.

Of the parents aware of the bullying, 14 per cent to 18 per cent said they had been told of bullying by teachers. Only 3 per cent to 4 per cent of the parents learned of the bullying from their children, according to the survey.

The survey also found that 42 per cent of primary school teachers are not aware of bullying aimed at their students. The portion of such teachers was 29 per cent at junior high schools and 69 per cent at senior high schools.

Asked for the reason behind bullying, about 85 per cent of the teachers cited a lack of education at home. Many parents singled out a lack of a sense of justice and compassion among children as the main reason.

An Education Ministry official said the findings suggest that parents and teachers should have closer contact with children to prevent bullying.

School bullying became a major issue in Japan after 13-year-old Kiyoteru Okouchi hanged himself in Nishio, Aichi Prefecture, in the fall of 1994, leaving a note saying that classmates had repeatedly dunked him in a nearby river and extorted money from him.

The bullying-suicide prompted the Education Ministry to issue a report on bullying in March 1995 urging teachers to order bullies not to come to school.

The article on the opposite page appeared in a Japanese newspaper in 1996. Refer to it to answer the questions below.

Question 2: BULLYING R118Q02 – 0 1 8 9

Why does the article mention the death of Kiyoteru Okouchi?

………………………………………………………………

………………………………………………………………

Figure 1: The paper-based (left) and the computer-based instrument (right).

sponses to twelve open-ended items. Paper-based responses were transcribed by two research assistants. Due to missing information about the gender of 44 students and 79 erroneously missing scores, 468 responses had to be excluded from the analysis, resulting in $n = 7,495$ responses written by $n = 836$ students from 78 schools. Beside the open-ended items that were relevant to the present study, students also answered 18 further PISA reading items and other tests (e.g., for computer-related skills) in a rotated design (Hahnel et al., 2016).

The study administered nine dichotomous and three partial-credit open-ended items that cannot be disclosed due to their confidentiality. There was no overlap with the item set of the study that was to be reproduced. Because PISA 2012 had not been computer-based yet, the items were implemented in the CBA ItemBuilder (Rölke, 2012). Reading items in PISA typically contain a stimulus comprising continuous or non-continuous text, or both. Often, multiple questions refer to such a stimulus, which is called a unit. The twelve analyzed items of this study were nested in seven such units. PISA reading items attempt to assess one of the following three processes (OECD, 2013): *Access & Retrieve* (locate explicit information in the stimulus), *Integrate & Interpret* (incorporate multiple pieces of explicit or implied information), and *Reflect & Evaluate* (apply world-knowledge for reflecting on the stimulus). Figure 1 shows an exemplary item in the paper- and the computer-based version (OECD, 2006, p. 59f.) that was not part of the present study. As a procedural difference, the computer implementation reminded students–but did not force–to give a response if they omitted it in the first place (for a list of differences see Kroehne, Buerger, et al., 2019).

Text responses were compared in two respects: `PEC` and `Rel` (cf. Theoretical Context). Both features were automatically extracted by *ReCo* (Zehner, Sälzer, & Goldhammer, 2016; Zehner et al., 2018). Analogous to the original study, part-of-speech tagging was carried out using the Stuttgart-Tübingen Tagset (STTS; Schiller, Teufel, Stöckert, & Thielen, 1999), and for normalizing the language data, automatic spelling correction (Jazzy; Idzelis, 2005) and stemming (Snowball; Porter, 2001) were applied. For each item, 1658 to 14,077 articles from the German Wikipedia were collected as the text corpus for building semantic spaces using Latent Semantic Analysis (Deerwester et al., 1990; following the methodology of Zehner et al., 2016).

Modeling approach

Identical to Zehner et al. (2019), we specified four (Generalized) Linear Mixed Models (GLMM) with increasing complexity for both response features each. For the sake of simplicity of the following, the dependent variables PEC and Rel are both described as η_{pi}, whereas $\eta_{pi} = \begin{cases} \log(\texttt{PEC}) & \text{for } \texttt{PEC} \\ \texttt{Rel} & \text{for } \texttt{Rel} \end{cases}$.

The models' goodness of fit are reported on the basis of the χ^2-distributed Likelihood Ratio Test statistic (LRT), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Marginal and conditional R^2 are reported where available (Nakagawa, Schielzeth, & O'Hara, 2013).

Model (1) served as the baseline, only representing the data structure.

$$\eta_{pi} \;=\; \beta_0 + \mathrm{t}_{0p} + \mathrm{e}_{0i} + \mathrm{s}_{0k} \tag{1}$$

Here, the feature η (PEC or Rel) was estimated for student p from school k, who responded to item i. It was decomposed into the following components:

- fixed intercept β_0
- random intercept for students t_{0p}
- random intercept for items e_{0i}
- random intercept for schools s_{0k}

Model (2) tested Hypothesis 1 by adding a fixed effect β_1 for response correctness $F_{ip} \in \{0.0, 0.5, 1.0\}$ with $F_{ip} = 0.5$ for *Partial Credit* in non-dichotomous items.

$$\eta_{pi} \;=\; \beta_0 + \mathrm{t}_{0p} + \mathrm{e}_{0i} + \mathrm{s}_{0k} + \beta_1 \mathrm{F}_{ip} \tag{2}$$

Testing Hypotheses 2a and 2b, Model (3) added to Model (2):

$$\eta_{pi} \;=\; \beta_0 + \mathrm{t}_{0p} + \mathrm{e}_{0i} + \mathrm{s}_{0k} + \beta_1 \mathrm{F}_{ip} + (\beta_2 + \mathrm{c}_{0i})\mathrm{M}_p \tag{3}$$

- fixed effect β_2 of mode M_p
- random by-item mode effect c_{0i}

The final Model (4) tested Hypothesis 3a and 3b:

$$\eta_{pi} \;=\; \beta_0 + \mathrm{t}_{0p} + \mathrm{e}_{0i} + \mathrm{s}_{0k} + \beta_1 \mathrm{F}_{ip} + (\beta_2 + \mathrm{c}_{0i})\mathrm{M}_p + (\beta_3 + \mathrm{g}_{0i})\mathrm{G}_p + \beta_4 \mathrm{M}_\mathrm{p}\mathrm{G}_p \tag{4}$$

- fixed effect β_3 of gender G_p
- fixed effect β_4 of the interaction between gender G_p and mode M_p
- random by-item effect of gender g_{0i}

The distribution of the random effects **b** was modeled as a multivariate normal distribution; $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ as the covariance matrix of the random effects. `PEC`'s probability distribution was modeled as a Poisson distribution (cf. Stroup, 2012), `PEC` $\sim Poisson(\lambda)$. Likewise, `Rel`'s probability distribution was modeled as a normal distribution, `Rel` $\sim \mathcal{N}(\mu, \sigma^2)$.

Software

ReCo (Zehner et al., 2016) extracted the response features with its many software components: *DKPro Core* (Gurevych et al., 2007), *DKPro Similarity* (Bär, Zesch, & Gurevych, 2013), *JWPL* (Zesch, Müller, & Gurevych, 2008), *S-Space* (Jurgens & Stevens, 2010), *Snowball* (Porter, 2001), *Stanford NLP Parser* (Rafferty & Manning, 2008).

The statistical analyses were carried out in *R 3.5.3* (R Core Team, 2019), using *snow* for parallel computations (Tierney, Rossini, Li, & Sevcikova, 2018), *lme4* (Bates, Mächler, Bolker, & Walker, 2015) for GLMM estimation, and *r2glmm* (Jaeger, 2017) as well as *MuMIn* (Barton, 2018) for computing R^2.

Results

Before the following subsections depict the results for information quantity (`PEC`) and relevance (`Rel`), Table 1 shows the percentage of correct as well as empty responses by mode. It appears, the twelve open-ended response items are, on average, more difficult on computer than on paper (solved less often by 5%). At the same time, responses that were assessed in paper-based mode were left empty more often by 5 percent on average.

Figure 2 shows the descriptives neglecting the nesting and crossing (of students and items) and outliers. The top row shows the original findings of Zehner et al. (2019), the bottom row the results of the present study. The boxplots in the left column depict `PEC`; `Rel` is shown on the right. In each subplot, the first four bars display incorrect responses, the right ones stand for correct responses. Furthermore, the bars distinguish responses from each gender (♂, ♀) as well as the experimental conditions: paper-based assessment in gray (PBA), computer-based assessment in orange bars (CBA). While the boxes' centers represent the median, the white lines additionally show the arithmetic mean. The figure shows the general tendency of including more proposition entities in computer- than in paper-based modes, which is also true for female compared to male students. For the Relevance Proportion, it is apparent that there is no difference between the modes for correct responses, but there is for incorrect. For better readability, partial credit items are dichotomized in this figure (partial credit considered as correct). The

Table 1: Percentage of Correct and Empty Responses by Administration Mode

	Item ID	**R227Q03**	**R227Q06**	**R111Q02B**	**R111Q06B**	**R055Q02**	**R055Q03**	
Correct	*CBA*	48%	74%	34%	32%	40%	57%	
	PBA	49%	74%	42%	33%	46%	57%	
	Δ	−1%	0%	**−8%**	−1%	−6%	0%	
Empty	*CBA*	10%	4%	11%	12%	15%	10%	
	PBA	25%	8%	18%	17%	21%	13%	
	Δ	**−15%**	**−4%**	**−7%**	−5%	**−6%**	−3%	
	Item ID	**R055Q05**	**R458Q07**	**R447Q06**	**R452Q03**	**R452Q06**	**R414Q06**	*Total*
Correct	*CBA*	53%	52%	42%	14%	36%	35%	43%
	PBA	65%	59%	49%	14%	45%	38%	48%
	Δ	**−12%**	−7%	−7%	0%	**−9%**	−3%	−5%
Empty	*CBA*	16%	13%	13%	13%	23%	24%	14%
	PBA	14%	18%	16%	15%	27%	33%	19%
	Δ	2%	**−5%**	−3%	−2%	−4%	**−9%**	−5%

Note. Significant differences ($\alpha = .05$) at the item level are printed in bold.

close similarity between the descriptives of the original and the present study indicates a stable recovery of many original findings, which is thoroughly analyzed in the following.

Information quantity (PEC)

Table 2 shows the results of the four GLMMs for the responses' Proposition Entity Count (PEC). The information criteria indicate that the inclusion of all proposed effects fits the data best (Model [4]). Beyond Model (2), the inclusion of the fixed mode effect (Model [3]) yields an additional $\Delta R^2_m = 1.7$ percent in explaining the variation of how much information was included in a response. The overall effect of gender and its interaction with mode add another $\Delta R^2_m = 1.7$ percent. Note that the marginal R^2_m does not include random effects.

While the model fit statistics of the different models give a first insight, the effect estimates confirm the hypotheses regarding PEC. All fixed effects differ significantly from 0. The effect $\beta_1 = 0.54[\pm 0.02]$[1] ($z = 44.71, p = .000$), indicates that the largest, positive impact on information quantity is the correctness of a response (P1). An aggregate mode effect on PEC (P2a) is shown by $\beta_2 = 0.22[\pm 0.09]$ ($z = 4.56, p = .000$). Also, the effect of mode on PEC varies substantially across items (P2b), $\text{sd}(c_{0i}) = 0.15$, ranging from $\beta_2 + c_{02} = -0.05$ to $\beta_2 + c_{04} = 0.41$. Likewise, gender had an overall effect on PEC in that girls incorporated more pieces of information into their responses (P3a), $\beta_3 = 0.20[\pm 0.09]$ ($z = 4.54, p = .000$). The gender effect varied across items with $\text{sd}(g_{0i}) = 0.10$, and it correlated with the mode effect within items by $cor(g_{0i}, c_{0i}) = .52$. That is, items with a larger mode effect also tended to show a larger gender effect (P3b). The omission of single random effects in the final model shows that the by-item mode effect is more important to fit the data well than the by-

[1] The brackets indicate 95% confidence intervals.

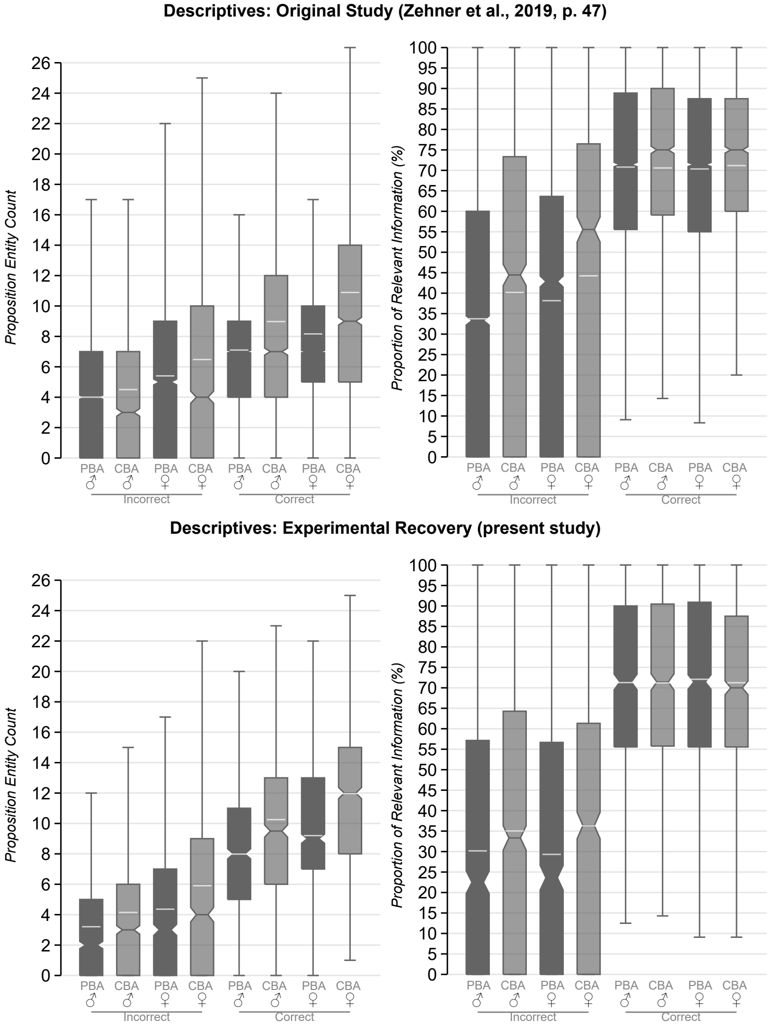


Figure 2: Comprehensive Recovery of Descriptives for PEC (Information Quantity, left) & Rel (Relevance Proportion, right), neglecting the nested structure and outliers. Top row shows the original study that was to be reproduced, bottom row the results of the present study. The transparent white lines represent arithmetic means. Partial credit items are dichotomized here.

Table 2: Proposition Entity Count (`PEC`): Generalized Linear Mixed Models (1)–(4)

	n_{Par}	AIC	BIC	χ^2	Δdf	p_{χ^2}	R^2_m	ΔR^2_m	$p_{\Delta R^2_m}$
Model (1)	4	47232	47259						
Model (2)	5	45305	45339				.080 [±.012]		
Response Correctness							.080 [±.012]		
Model (3)	8	44583	44638				.097 [±.020]	.017 [±.001]	.000
Response Correctness							.085 [±.012]		
Mode							.017 [±.007]		
Model (4)	13	44417	44507				.114 [±.014]	.017 [±.001]	.000
Response Correctness							.086 [±.012]		
Mode							.007 [±.005]		
Gender							.005 [±.003]		
Mode*Gender							.000 [±.001]		
w/o $\vert$Mode and $\vert$Gender	8	44796	44851	389.20	5	.000			
w/o $\vert$Mode	10	44684	44753	273.08	3	.000			
w/o $\vert$Gender	10	44534	44603	122.94	3	.000			

Note. n_{Par} = number of estimated parameters, Δdf = degrees of freedom for the Likelihood Ratio Test (LRT; χ^2), p_{χ^2} = p-value for the LRT, R^2_m = marginal R^2 of fixed effects (Nakagawa et al., 2013), ΔR^2_m = difference of R^2_m with previous model, $p_{\Delta R^2_m}$ = p-value for the R^2 difference test; 95% confidence intervals in brackets; w/o = without; random effects are indicated by pipes (|); the restricted models in the last three lines (excluding the random by-item effects) are each compared to Model (4)

item gender effect is (cf. AIC/BIC in Table 2). Finally, the fixed interaction between mode and gender turned out to be significant, $\beta_4 = 0.05[\pm 0.04]$, $(z = 1.99, p = .046)$, but without any practical importance, given the small magnitude.

Relevance proportion (`Rel`)

Table 3 shows the results of the four LMMs for the responses' Relevance Proportion (`Rel`); among others, the conditional R^2_c (including fixed *and* random effects) and marginal R^2_m (including *only* fixed effects). The baseline Model (1) explains $R^2_c = 34$ percent, response correctness adds roughly $\Delta R^2_c = 6$ percent, whereas the effects of interest, mode and gender, only increase the proportion of explained variance marginally ($\Delta R^2_c = 0.2$ and 0.1%). Model (3), introducing mode effects, shows the best fit to the data, being superior to the previous model, $\chi^2(3) = 898.77, p = .000$, while Model (4) does not further improve the data likelihood, $\chi^2(5) = 0.00, p = 1.000$. Moreover, R^2_m shows that some variance in `PEC` is attributed to the mode effect on the item level, but its contribution is only marginal.

Identical to the information quantity model, the most prominent fixed effect is response correctness, $\beta_1 = 0.32[\pm 0.01]$, supporting hypothesis R1. In contrast to the results of Zehner et al. (2019), $\beta_2 = 0.02[\pm 0.02]$ shows a statistically significant overall mode effect[2], but–reproducing the original study's findings–of hardly any practical relevance. Thus, R2a–based on previous findings, predicting *no* overall mode effect–is confirmed for its effect size and is practically irrelevant. Unexpectedly, the mode effect did not vary substantially across items (R2b), sd$(c_{0i}) = 0.02$, ranging from $\beta_2 + c_{07} = -0.01$ to $\beta_2 + c_{06} = 0.05$. As expected in R3a, no aggregate effect of gender could be found,

[2] Rounding with more digits: $\beta_2 = 0.023[\pm 0.022]$

Table 3: Relevance Measure (`Rel`): Linear Mixed Models (1)–(4)

	n_{Par}	AIC	BIC	R^2_c	R^2_m	ΔR^2_m	$p_{\Delta R^2_m}$
Model (1)	5	3199	3234	.336			
Model (2)	6	1576	1618	.393	.224 [±.016]		
Response Correctness					.224 [±.016]		
Model (3)	9	1563	1625	.395	.227 [±.016]	.003 [±.001]	.000
Response Correctness					.226 [±.016]		
Mode					.003 [±.003]		
Model (4)	14	1577	1674	.396	.226 [±.016]	−.001 [±.001]	.076
Response Correctness					.224 [±.016]		
Mode					.001 [±.002]		
Gender					.000 [±.001]		
Mode*Gender					.000 [±.001]		

Note. n_{Par} = number of estimated parameters, R^2_c = conditional R^2 of fixed & random effects, R^2_m = marginal R^2 of fixed effects (aggregated by model as well as for every effect; Nakagawa et al., 2013), ΔR^2_m = difference of R^2_m with previous model, $p_{\Delta R^2_m}$ = p-value for the R^2 difference test; 95% confidence intervals in brackets

$\beta_3 = -0.01[\pm 0.02]$. Supporting R3b, gender and its interaction with mode have estimates not significantly different from 0, $\beta_4 = 0.02[\pm 0.03]$. The expected item-wise variation of the gender effect, however, was only small, sd$(g_{0i}) = 0.02$. Nevertheless, the gender effect estimates correlated moderately and negatively with the mode effect estimates, $cor(g_{0i}, c_{0i}) = -.47$. That is, items that tended to come with larger mode effects tended to compensate this with a more negative gender effect. Thus, if an item is more prone to being affected by mode, it is likely that it will also evoke larger (negative) gender differences. However, given the effects' small variation around 0, the practical relevance of this finding is marginal for `Rel`. In addition, the negative correlation (opposed to a positive one in the original study) hints at either unstable mechanisms or a strong dependency of item characteristics. The correlation coefficients of effect estimates might be somewhat unreliable in these studies given the relatively small number of items. Still, across items, the studies demonstrate an interplay of mode and item characteristics related to subgroups, while they do not allow a generic mechanism to be concluded.

Discussion

For identifying potential mode effects on text responses to PISA reading items, we compared these in a paper- and a computer-based administration. Previous research (Zehner et al., 2019) had found differences in linguistic features of text responses between the paper-based PISA 2012 and the computer-based PISA 2015, which could stem from observing different cohorts. Attempting to reproduce these findings in an experimental setting, we analyzed data from a German add-on study to PISA 2012 and extracted the same linguistic indicators by means of natural language processing techniques: the Proposition Entity Count (`PEC`; information quantity) and Relevance Proportion (`Rel`; information quality). This way, the text response features are considered outcomes that might be affected by administration mode. Hence, as indicators for characteristics of the process, they provide new insights for potential mode effects, whereas they only constitute an addition to traditional mode effect analyses.

The present study reproduced the overall picture of Zehner et al. (2019) surprisingly well, given the fact that part of the research questions addressed item-specific differences which were analyzed with disjoint sets of items. The relationship between response correctness and information quantity as well as relevance turned out to again be crucial. Correct responses tended to contain more pieces of information and larger proportions of relevant information. In case of the latter, this also serves as validity evidence for the indicator: Correct responses contain substantially larger proportions of relevant information. Another necessary, though insufficient, requirement for the indicators' validity is their stability, which is here demonstrated with a different item set and an independent data collection at a fine-grained level (grouping by gender, mode, and response correctness). These findings thus support the utility and accuracy of the employed measures.

Computer-based items turned out to be slightly harder as can be seen from the item scores. This is in line with most mode effect studies (e.g., Kolen & Brennan, 2014; Kroehne, Buerger, et al., 2019). Interestingly, though, this ran contrary to the rate of empty responses, which was lower in the open-ended computer-based items. That means, even though students were attempting to answer an item more often in this particular computer-based assessment, they were less often successful than on paper. It is likely that the attempt rate was influenced by the fact the computer-based implementation reminded students of empty text fields.

For information quantity, a moderate overall effect of the administration mode could be found. Students tended to incorporate more pieces of information when responding on computer than on paper. This does not only support the original study, but also White et al. (2015) who observed responses of eighth graders to be twice as long when writing on computer. The reasons for this remain a matter of speculation so far. Prominent potential sources, however, could be the input mode of keyboarding opposed to pen writing and a possibly higher motivation for computer-based testing. Regarding the latter, with the add-on study being placed after a whole day of paper-based assessment, the computer-based testing might have been perceived to be very distinct with corresponding effects on motivation, attention, and memory (sensu Eysenck & Eysenck, 1980). For the same data set, discrepancies in the response process were demonstrated by shorter response times in the computer-based assessment (Kroehne, Hahnel, & Goldhammer, 2019). Keyboarding and handwriting are only moderately correlated with respect to writing fluency and speed (Feng et al., 2017), which might be a prominent source for differences between the input modes.

Like in the original study, the overall mode effect could not be found for Relevance Proportion. Also, while the original study found quite some variation of the mode effect on the item level for Relevance Proportion, the dispersion of item-wise mode effects was very low for the present study. This means, with different items, cohorts, computer platforms, and randomized conditions, the mode effect on proportion relevance could not be recovered. However, identical to the original study, the descriptives showed that there appeared to be substantial differences induced by mode for incorrect responses, but none for correct responses. This difference was masked at the aggregate level. For information quantity, however, the items varied in the degree in which mode took effect. While two items showed hardly any mode effect in this respect, three others were affected by almost the magnitude of the relationship between information quantity and response correctness.

Altogether, the combination of lower percentages of correct responses and shorter response times on the one hand and more information, less empty responses (possibly, induced by feedback if a response was missing), and an unchanged proportion of relevant information on the other hand draws a complex pattern that is reduced when only looking at scores. The mechanisms that lead to these products will be at the core of future studies that should combine process and product data.

Finally looking at potential differential effects, we could reproduce that there is no interaction between gender and mode at the aggregate level. Switching to the item level, however, we could reproduce the original findings that an items' mode effect is related to its characteristic of being influenced by gender with respect to the Proposition Entity Count as well as Relevance Proportion.

In conclusion, there is strong evidence for supporting the original findings on text responses' differences between computer- and paper-based assessment (Zehner et al., 2019). That is, there seem to be mode effects on text responses in PISA's reading assessment which have not been taken into account before. At the aggregate level, the differences are partly, though not completely masked, but they become obvious at the item level. This can harm trend interpretations if they mirror a systematic shift in the assessed construct which was not, or differently, captured in scores. Changes in the assessment framework over time require newly constructed items and can accordingly affect construct facets in parts, challenging the interpretation of trends. If a mode effect however, systematically changes these facet weights across many items–for example if the response production is more emphasized through keyboarding instead of handwriting–, the overall construct has shifted further and trend comparisons were even more difficult to interpret. Furthermore, such a change could, theoretically, be invisible at the score level if different shifts in the facets' weightings compensate each other at the level of difficulty. Contrary to score analyses, text responses might carry corresponding information for identifying such a construct shift.

That being said, the intention of modernizing assessment frameworks in contexts such as PISA is coherent. The comparability of resulting scales, however, must be verified if they are to be linked. New data sources and analyses, such as process data or the methodology presented here, can be used to better understand the underlying mechanisms for further controlling these in the future, either in test development or statistically.

Beside reporting further evidence for mode effects in PISA, the methodology used in this study can help establish more proper comparability in international large-scale assessment by providing new information. Text responses provide more fine-grained information on the respondents' understanding while natural language processing allows this information to be extracted objectively and at a large scale.

Limitations

This study recovered the findings of Zehner et al. (2019) to a large extent, but both studies have to be interpreted in the context of several limitations. First of all, the findings are limited to Germany, the reading domain, and the analyzed open-ended items are only one part of the PISA reading test. While both studies attempt to shed light on PISA's computerization, the present experimental design used a different computer environment as the OECD's implementation had not been available yet.

Beside these design-driven constraints, caution is required regarding the two automatically extracted text response features. Though this follow-up study provides some evidence for the stability of both and validity of the `Rel` measure, a thorough validation remains to be done. Their shortcomings are obvious at the conceptual level. Both of them are proxies for what is intended to be measured.

First, `PEC` does not reflect the actual number of propositions, but only addresses words that, potentially, are part of a stated proposition. Which parts of speech to use is ambiguous in few cases (cf. Zehner et al., 2018). However, the measure's design is comparable to proposition density in the software CPIDR (Brown, Snodgrass, Kemper, Herman, & Covington, 2008). Overall, analyzing the word-level when looking for propositions is conceptually problematic because what constitutes a relevant propositional element is sometimes depending on the context. For example, it is state of the art to neglect auxiliary verbs as they do not point at some meaningful element in the situation model on their own. However, if tense plays a role in an item, it is likely that the tense is captured in an auxiliary verb and, hence, this adds to the semantic of the statement in a crucial manner.

Second, for `Rel`, the PISA coding guides constituted the gold standard for exclusively correct responses. Since a previous study had already shown the reference texts in the coding guides do not exhaustively mirror the empirical response types (Zehner, Goldhammer, & Sälzer, 2015), the measure can be considered to be underestimating the true values. Furthermore, `Rel` uses norm-referenced comparisons so that they only allow within-sample comparisons, but the magnitude of the absolute values cannot be interpreted. Both employed measures can be regarded as first attempts that have to be built on in the future.

Finally, the procedure included automatic spelling correction using Jazzy spellchecker (Idzelis, 2005), which comes with the risk of correcting words incorrectly. From our point of view, the benefit of normalizing language in order to avoid bias for students with low writing skills outweighs the risk when a word is changed improperly. This case should be distributed randomly and not introduce a systematic bias. Jazzy's accuracy is strongly dependent on the dictionaries used, so it is unknown for our application. However, Horbach, Ding, and Zesch (2017) report a precision of $P = .86$, recall rate of $R = .94$, and resulting F-score of $F = .90$ for dictionaries with smaller size. Jazzy corrected 3.6% of all words written in the paper- and 4.0% in the computer-based mode, which is a reasonably small increase given the latter adds typos as a source of misspellings.

References

Bär, D., Zesch, T., & Gurevych, I. (2013). DKPro Similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 121–126). Sofia, Bulgaria: Association for Computational Linguistics.

Barton, K. (2018). MuMIn: Multi-model inference [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=MuMIn` (R package version 1.42.1)

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Bridgeman, B., Lennon, M. L., & Jackenthal, A. (2001). *Effects of screen size, screen resolution, and display rate on computer-based test performance* (RR-01-23 ed.). doi: 10.1002/j.2333-8504.2001.tb01865.x

Brown, C., Snodgrass, T., Kemper, S. J., Herman, R., & Covington, M. A. (2008). Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior research methods*, *40*(2), 540–545.

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.

Burrows, S., Gurevych, I., & Stein, B. (2014). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, 1–58. doi: 10.1007/s40593-014-0026-8

Choi, S. W., & Tinkler, T. (2002). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting.* Paper presented at the Annual Meeting of the National Council on Measurement in Education. New Orleans, LA.

Clariana, R., & Wallace, P. (2002). Paper-based versus computer-based assessment: Key factors associated with the test mode effect. *British Journal of Educational Technology*, *33*(5), 593–602. doi: 10.1111/1467-8535.00294

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, *25*, 23–38. doi: https://doi.org/10.1016/j.edurev.2018.09.003

Dzikovska, M. O., Nielsen, R. D., & Leacock, C. (2016). The joint student response analysis and recognizing textual entailment challenge: Making sense of student responses in educational applications. *Language Resources and Evaluation*, *50*(1), 67–93. doi: 10.1007/s10579-015-9313-8

Eysenck, M. W., & Eysenck, M. C. (1980). Effects of processing depth, distinctiveness, and word frequency on retention. *British journal of psychology*, *71*(2), 263–274. doi: 10.1111/j.2044-8295.1980.tb01743.x

Feng, L., Lindner, A., Ji, X. R., & Malatesha Joshi, R. (2017). The roles of handwriting and keyboarding in writing: A meta-analytic review. *Reading and Writing*, *85*, 1–31. doi: 10.1007/s11145-017-9749-x

Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018, Oct 29). 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. doi: 10.1186/s40536-018-0064-z

Galhardi, L. B., & Brancher, J. D. (2018). Machine learning approach for automatic short answer grading: A systematic review. In *Ibero-american conference on artificial intelligence* (pp. 380–391). doi: 10.1007/978-3-030-03928-8_31

Graesser, A. C., & Clark, L. F. (1985). *Structures and procedures of implicit knowledge* (Vol. 17). Norwood, NJ: Ablex.

Graesser, A. C., & Franklin, S. P. (1990). QUEST: A cognitive model of question answering. *Discourse Processes*, *13*(3), 279–303. doi: 10.1080/01638539009544760

Gurevych, I., Mühlhäuser, M., Müller, C., Steimle, J., Weimer, M., & Zesch, T. (2007). Darmstadt knowledge processing repository based on UIMA. In *Proceedings of the First Workshop on Unstructured Information Management Architecture at Biannual Conference of the Society for Computational Linguistics and Language Technology.* Tübingen, Germany.

Hahnel, C., Goldhammer, F., Naumann, J., & Kroehne, U. (2016). Effects of linear reading, basic computer skills, evaluating online information, and navigation on reading digital text. *Computers in Human Behavior*, *55*, 486 - 500. doi: 10.1016/j.chb.2015.09.042

Higgins, D., Brew, C., Heilman, M., Ziai, R., Chen, L., Cahill, A., … Blackmore, J. (2014). Is getting the right answer just about choosing the right words? The role of syntactically-informed features in short answer scoring. *CoRR*, *abs/1403.0801*.

Horbach, A., Ding, Y., & Zesch, T. (2017). The influence of spelling errors on content scoring performance. In *Proceedings of the 4th workshop on natural language processing techniques for educational applications (nlptea 2017)* (pp. 45–53). Taipei, Taiwan: Asian Federation of Natural Language Processing.

Horkay, N., Bennett, R. E., Allen, N., Kaplan, B., & Yan, F. (2006). Does it matter if I take my writing test on computer? An empirical study of mode effects in NAEP. *Journal of Technology, Learning, and Assessment*, *5*(2).

Idzelis, M. (2005). *Jazzy: The java open source spell checker.* Retrieved 2019/10/10, from `http://jazzy.sourceforge.net/`

Jaeger, B. (2017). *r2glmm: Computes R squared for mixed (multilevel) models.* Retrieved from `https://CRAN.R-project.org/package=r2glmm`

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. doi: 10.1080/03054985.2018.1430025

Jurgens, D., & Stevens, K. (2010). The S-Space package: An open source package for word space models. In Association for Computational Linguistics (Ed.), *48th Annual Meeting of the Association for Computational Linguistics* (pp. 30–35).

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge University Press.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363. doi: 10.1037/0033-295X.85.5.363

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3. ed. ed.). New York, NY: Springer. doi: 10.1007/978-1-4757-4310-4

Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, *123*, 138–149. doi: https://doi.org/10.1016/j.compedu.2018.05.005

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. doi: 10.1111/emip.12280

Kroehne, U., Hahnel, C., & Goldhammer, F. (2019). Invariance of the response processes between gender and modes in an assessment of reading. *Frontiers in Applied Mathematics and Statistics*, *5:2*. doi: 10.3389/fams.2019.00002

Kroehne, U., & Martens, T. (2011). Computer-based competence tests in the National Educational Panel Study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, *14*(S2), 169–186. doi: 10.1007/s11618-011-0185-4

Mullis, I. V. S., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading.* Boston College: Chestnut Hill, MA.

Nakagawa, S., Schielzeth, H., & O'Hara, R. B. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142. doi: 10.1111/j.2041-210x.2012.00261.x

Noyes, J. M., & Garland, K. J. (2008). Computer- vs. paper-based tasks: Are they equivalent? *Ergonomics*, *51*(9), 1352–1375. doi: 10.1080/00140130802170387

OECD. (2006). *PISA released items - reading*. Retrieved 18.02.2016, from `http://www.oecd.org/pisa/38709396.pdf`

OECD. (2013). *PISA 2012 assessment and analytical framework: Mathematics, reading, science, problem solving and financial literacy*. OECD Publishing. doi: 10.1787/9789264190511-en

OECD. (2016). *PISA 2015 results (volume I)*. Paris: OECD Publishing. doi: 10.1787/19963777

OECD. (2017a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving*. Paris: OECD Publishing. doi: 10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 technical report*. Paris: OECD Publishing.

Piaw, C. Y. (2011). Comparisons between computer-based testing and paper-pencil testing: Testing effect, test scores, testing time and testing motivation. In *Proceedings of the Informatics Conference* (pp. 1–9).

Porter, M. (2001). *Snowball: A language for stemming algorithms*. Retrieved from `http://snowball.tartarus.org/texts/introduction.html`

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

Rafferty, A. N., & Manning, C. D. (2008). Parsing three German treebanks: Lexicalized and unlexicalized baselines. In *Proceedings of the Workshop on Parsing German* (pp. 40–46). doi: 10.3115/1621401.1621407

Robitzsch, A., Lüdtke, O., Köller, O., Kroehne, U., Goldhammer, F., & Heine, J.-H. (2017). Herausforderungen bei der Schätzung von Trends in Schulleistungsstudien [Challenges for Trend Estimation in Educational Assessments]. *Diagnostica*, *63*(2), 148–165. doi: 10.1026/0012-1924/a000177

Rölke, H. (2012). The ItemBuilder: A graphical authoring system for complex item development. In T. Bastiaens & G. Marks (Eds.), *Proceedings of e-learn: World conference on e-learning in corporate, government, healthcare, and higher education 2012* (pp. 344–353). Montréal, Quebec, Canada: Association for the Advancement of Computing in Education (AACE).

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textkorpora mit STTS*. University of Stuttgart and University of Tübingen.

Stroup, W. W. (2012). *Generalized Linear Mixed Models: Modern concepts, methods and applications*. Hoboken: CRC Press.

Tierney, L., Rossini, A. J., Li, N., & Sevcikova, H. (2018). snow: Simple network of workstations [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=snow` (R package version 0.4-3)

Wang, S., Jiao, H., Young, M. J., Brooks, T., & Olson, J. (2007). Comparability of computer-based and paper-and-pencil testing in K–12 reading assessments. *Educational and Psychological Measurement*, *68*(1), 5–24. doi: 10.1177/0013164407305592

White, S., Kim, Y. Y., Chen, J., & Liu, F. (2015). *Performance of fourth-grade students in the 2012 NAEP computer-based writing pilot assessment: Scores, text length, and use of editing tools.* Working Paper Series, NCES 2015-119.

Zehner, F., Goldhammer, F., Lubaway, E., & Sälzer, C. (2019). Unattended consequences: How text responses alter alongside PISA's mode change from 2012 to 2015. *Education Inquiry*, *10*(1), 34–55. doi: 10.1080/20004508.2018.1518080

Zehner, F., Goldhammer, F., & Sälzer, C. (2015). Using and improving coding guides for and by automatic coding of PISA short text responses. In *Proceedings of the IEEE ICDM Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2015).* doi: 10.1109/icdmw.2015.189

Zehner, F., Goldhammer, F., & Sälzer, C. (2018). Automatically analyzing text responses for exploring gender-specific cognitions in PISA reading. *Large-scale Assessments in Education*, *6:7*. doi: 10.1186/s40536-018-0060-3

Zehner, F., Sälzer, C., & Goldhammer, F. (2016). Automatic coding of short text responses via clustering in educational assessment. *Educational and Psychological Measurement*, *76*(2), 280–303. doi: 10.1177/0013164415590022

Zesch, T., Müller, C., & Gurevych, I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech and Morocco.