

Evaluating group differences in online reading comprehension: The impact of item properties

Hatice Cigdem Bulut, Okan Bulut & Serkan Arikan

To cite this article: Hatice Cigdem Bulut, Okan Bulut & Serkan Arikan (2023) Evaluating group differences in online reading comprehension: The impact of item properties, International Journal of Testing, 23:1, 10-33, DOI: [10.1080/15305058.2022.2044821](https://doi.org/10.1080/15305058.2022.2044821)

To link to this article: <https://doi.org/10.1080/15305058.2022.2044821>



Published online: 25 Feb 2022.



Submit your article to this journal [↗](#)



Article views: 678



View related articles [↗](#)





View Crossmark data [↗](#)



Citing articles: 2 View citing articles [↗](#)



Evaluating group differences in online reading comprehension: The impact of item properties

Hatice Cigdem Bulut^a , Okan Bulut^b , and Serkan Arikan^c

^aDepartment of Education Sciences, Faculty of Education, Cukurova University, Adana, Turkey; ^bCentre for Research in Applied Measurement and Evaluation, University of Alberta, Edmonton, AB, Canada; ^cDepartment of Mathematics and Science Education, North Campus, Faculty of Education, Bogazici University, Istanbul, Turkey

ABSTRACT

This study examined group differences in online reading comprehension (ORC) using student data from the 2016 administration of the Progress in International Reading Literacy Study (ePIRLS). An explanatory item response modeling approach was used to explore the effects of item properties (i.e., item format, text complexity, and cognitive complexity), student characteristics (i.e., gender and language groups), and their interactions on dichotomous and polytomous item responses. The results showed that female students outperform male students in ORC tasks and that the achievement difference between female and male students appears to change text complexity increases. Similarly, the cognitive complexity of the items seems to play a significant role in explaining the gender gap in ORC performance. Students who never (or sometimes) speak the test language at home particularly struggled with answering ORC tasks. The achievement gap between students who always (or almost always) speak the test language at home and those who never (or sometimes) speak the test language at home was larger for constructed-response items and items with higher cognitive complexity. Overall, the findings suggest that item properties could help understand performance differences between gender and language groups in ORC assessments.

KEYWORDS

Achievement gap;
ePIRLS;
explanatory item
response modeling;
item properties;
online reading
comprehension

CONTACT Okan Bulut  bulut@ualberta.ca  Centre for Research in Applied Measurement and Evaluation, University of Alberta, 6-110 Education Centre North, 11210 87 Ave NW, Edmonton, AB T6G 2G5, Canada.

© 2022 International Test Commission

The advent of the digital age has dramatically affected the ways students read, acquire, and learn information. With the emergence of new skills such as media and information literacy (Potter, 2018), many countries have begun to focus on online reading comprehension (ORC) to equip their students with these skills involved in literacy development. Hence, significant changes seem necessary for adapting the curriculum, teacher practices, and assessments in literacy to new standards of the 21st century (Drew, 2012; Heo & Toomey, 2020). To facilitate the transition to the 21st century standards and curriculum, it is essential to develop and use dynamic assessments measuring how students interact with information in online environments (Leu et al., 2013). Recently, international large-scale assessments, such as the Progress in International Reading Literacy Study (PIRLS), the Programme for International Student Assessment (PISA), and the Programme for the International Assessment of Adult Competencies, have extended their frameworks to assess ORC in addition to traditional literacy skills (Mullis et al., 2017; OECD, 2019, 2021).

To date, researchers have consistently reported achievement gaps in ORC related to student characteristics—such as gender, socioeconomic status (SES), immigration status, and native language (e.g., Kannianen et al., 2019; Leu et al., 2015; Naumann & Sälzer, 2017; Schulz-Heidorf & Støle, 2018). In addition to student characteristics, item properties (e.g., item format and navigation demands of online reading items) have also been associated with student performance in ORC tasks (e.g., Moon et al., 2019; 2020; Naumann, 2015). However, the interaction between student characteristics and item properties has received scant attention from researchers in the ORC literature.

Examining the effects of item properties, student characteristics, and their interactions simultaneously could be significant for several reasons. First, investigating the effects of item properties, in addition to student characteristics, enables researchers to better understand the effects of item properties on students' ORC performance. Such investigations can reveal which item properties are essential regarding the variability in the difficulty of ORC items, and thereby explain additional sources of variance in students' test scores (Buerger et al., 2019; Gorin & Embretson, 2006; Kulesz et al., 2016). Second, grasping the role of item properties could contribute to the procedures of developing new items and assessment designs related to ORC (e.g., Buerger et al., 2019; Collier et al., 2018). Identifying the effect of item properties on student performance, in turn, can elicit what components and features need to be included or changed in the ORC items (e.g., Kulesz et al., 2016; Toyama, 2019). Third, item properties may affect ORC performance differently depending on student characteristics. This information could help test designers modify item difficulty by changing item properties and help practitioners

identify students who have difficulties responding to specific items (Kulesz et al., 2016; Toyama, 2019).

The current study aims to investigate the combined effects of student characteristics and item properties on fourth graders' performance in ORC tasks using student data from the 2016 administration of ePIRLS. An explanatory item response modeling (EIRM) approach was used to examine the effects of student characteristics (gender and speaking the test language at home), item properties (item format, processes of comprehension, and text complexity), and their interactions.

Literature review

Online reading comprehension

The theory of new literacies has become prominent with the Internet and other information and communication technologies (ICT) (Lankshear & Knobel, 2011; Leu et al., 2004, 2013). ORC, which is one of the new literacies, is defined as “a self-directed process of constructing texts and knowledge while engaged in several online reading practices: identifying important problems, locating information, critically evaluating information, synthesizing information, and communicating information” (Leu et al., 2013, p. 163). New literacies, including the domain of ORC, have reshaped and redefined what skills individuals might need when reading from the screen instead of paper (Leu et al., 2013). For example, recent research indicates that traditional reading (also known as offline reading) and ICT skills cannot entirely explain ORC skills (Coiro, 2011; Liu & Ko, 2016). ORC skills need to be built not only on traditional reading skills and ICT skills but also on a complex set of skills such as locating, evaluating, synthesizing, and communicating information on digital platforms (Coiro, 2011; Coiro & Dobler, 2007). ORC skills are beyond simple comprehension because it requires finding helpful information on multiple sources and applying critical thinking to all the information read.

In the digital age, students are expected to have adequate ORC skills to succeed in academic and social tasks (International Reading Association (IRA), 2009; Leu, 2017). To interact with information presented on webpages and other digital platforms, students need to be proficient enough to conduct online research on their own, and they have “digital wisdom” regarding how to learn online (Coiro, 2014). As digital platforms often include a sheer amount of information, finding the most helpful information can be daunting for students (Gao et al., 2022; Leu, 2017). Therefore, it is essential to support students to develop skills for locating and critically evaluating information. Also, there is

a growing need for dynamic assessments and technology-enhanced items that can guide educators and inform their instructional practices regarding ORC.

Item properties and student characteristics in ORC

Both item properties and student characteristics may impact students' reading processes and, ultimately, influence their ORC performance. Previous research on ORC examined the effects of gender, SES, and ICT skills as student characteristics (e.g., Coiro, 2011; Coiro & Dobler, 2007; Kanninen et al., 2019; Mullis et al., 2017; Schulz-Heidorf & Støle, 2018) and the effects of item format and text complexity as item properties (e.g., Moon et al., 2019; Naumann, 2015). However, there has been limited research on the interaction between student characteristics and item properties within the context of ORC. For example, researchers used data from the computer-based reading assessment of PISA 2009 and reported significant interactions between student characteristics (e.g., comprehension skills, enjoyment of reading, and knowledge of reading strategies) and task properties (e.g., difficulty of reading tasks and the average time students spent on each reading task) in the domain of reading digital text (Naumann, 2019; Naumann & Goldhammer, 2017). These studies also revealed that group-level differences based on student characteristics might be associated with item properties, suggesting that item properties may widen the existing achievement gap between student groups.

Previous studies also indicated that the achievement gap in reading among gender, ethnic, language, and SES groups continue to exist in ORC. However, the size and direction of achievement gaps appear to change depending on the type of skills (i.e., traditional reading and ORC) and the type of items involved in assessing students' reading performance (Mullis et al., 2017; Schulz-Heidorf & Støle, 2018). For example, some studies focusing on gender gaps indicated that female students outperform male students in ORC (e.g., Kanninen et al., 2019; Naumann & Sälzer, 2017; Schulz-Heidorf & Støle, 2018), whereas other studies reported the opposite results based on item type (e.g., transfer and retention items) (Heo & Toomey, 2020). Furthermore, item format appears to influence students' reading performance significantly. For example, Schwabe et al. (2015) found that female students performed better on constructed-response items than male students in PISA 2009 and PIRLS 2011.

Researchers reported that female students perform better on constructed-response items related to traditional reading comprehension (Schwabe et al., 2015), depending upon other student characteristics such as text reading comprehension (Kanninen et al., 2019). However, differential

reading performance between gender groups appears to change in online reading assessments (Mullis et al., 2017). For example, Schulz-Heidorf and Støle (2018) found that the performance of both male and female students declined in the digital reading assessment (i.e., ePIRLS 2016) compared with their performance in the traditional reading assessment (PIRLS 2016); however, the performance difference was more negligible for male students due to their enhanced interaction with the digital assessment environment. Similarly, students' ICT skills and familiarity with digital reading also influence their performance in ORC (Gil-Flores et al., 2012; Naumann, 2015, 2019).

Recent studies have shown that high-SES students outperform their low-SES counterparts in ORC (e.g., Jerrim, 2016; Naumann & Sälzer, 2017). However, Jerrim (2016) found that students' SES levels had a weaker relationship with student performance in online assessments compared with student performance in paper-and-pencil assessments. In contrast, Leu et al. (2015) reported a large effect size regarding the achievement gap in ORC, favoring economically advantaged students. Previous research also suggests that immigration status could be a strong indicator of lower SES, fewer home resources, and lower reading skills (e.g., Caro et al., 2009; Strand & Schwippert, 2019). Research suggests that nonimmigrant students tend to perform better in reading comprehension than immigrant students (Naumann & Sälzer, 2017; OECD, 2014; Strand & Schwippert, 2019).

Another reason for achievement gaps in ORC might be text complexity at the item level. Recent research shows that text complexity may significantly influence student performance in reading comprehension tasks (Amendum et al., 2018; Elleman & Oslund, 2019; Spencer et al., 2019). Many large-scale assessments, including statewide assessments using the Common Core State Standards for English language arts and literacy in the United States, consider text complexity a critical feature of the item development process when developing reading tasks. According to Mesmer et al. (2012), the perceived complexity of a given text depends on both text-related features (e.g., vocabulary, word length, and sentence length) and their interactions with readers. Therefore, the impact of text complexity on students' ORC performance may differ based on student characteristics and their interactions with other item properties.

The cognitive complexity of items may also influence students' reading comprehension (see Elleman & Oslund, [2019] for a review). Cognitive processes underlying reading items require students to use different skills in constructing answers from either online or traditional texts (Mullis & Martin, 2015). Previous studies showed that the difficulty of answering items in a digital environment is greater due to the additional cognitive capacity required to understand reading materials (Coiro & Dobler, 2007;

DeStefano & LeFevre, 2007; Lauterman & Ackerman, 2014). To answer digital reading items, students must employ traditional reading skills, at least basic ICT skills, and other skills such as evaluating and synthesizing information (Coiro & Dobler, 2007; Leu et al., 2013). Therefore, it is necessary to consider the effects of cognitive complexity at the item level to better understand how students perform in ORC tasks (Chen et al., 2014). However, no research has investigated the role of cognitive complexity in ORC assessments.

The current study employs the EIRM approach to investigate the effects of item properties and student characteristics on students' ORC performance. Previous studies used linear modeling (Coiro, 2011; Naumann, 2015, 2019; Schulz-Heidorf & Støle, 2018), item response theory (IRT) frameworks (Moon et al., 2019), and causal-comparative methods (Leu et al., 2015) to answer similar research questions. These studies adopted a two-step approach (i.e., first calculating students' scores or estimating item difficulties and then examining the effects of explanatory variables on the estimated parameters). This approach fails to consider the dependency between items and students' scores as it solely focuses on either the items or the scores. Also, each step of the analysis is likely to bring statistical error and contaminate the results. As a more compact and stringent method, EIRM can evaluate the effects of item properties and student characteristics and their interactions within the same model (De Boeck & Wilson, 2004). EIRM also distinguishes the variation among the items from the variation among students. Therefore, compared with the two-step approach, EIRM is more straightforward and is likely to involve less statistical error due to fewer numbers of analyses to be performed.

Present study

This study aims to explain the effects of item properties, student characteristics, and their interactions on students' ORC performance in a digital reading assessment. The following research questions are addressed in the study:

1. Which student characteristics (are gender and speaking the test language at home) affect students' ORC performance?
2. Which item properties (item format, processes of comprehension, and text complexity) affect students' ORC performance?
3. Which interactions between student characteristics and item properties affect students' ORC performance?

Methods

Sample

A total of 13,701 fourth-grade students from 18 countries or benchmarking entities participated in the 2016 administration of ePIRLS. The sample of this study consisted of 2894 students (47.5% female) who answered the released tasks in English because text complexity in English was one of the item properties used in this study, and it was only calculated for the released ePIRLS items. The country distribution of the students is presented in Table 1.

Instruments

Student characteristics used in the study were gender and speaking the test language (i.e., English) at home. Gender information was obtained from the ePIRLS student questionnaire. Speaking the test language at home variable was obtained from the questionnaire item of “How often do you speak the test language at home?” The response options of the language question were combined to create a binary variable (i.e., 0=never or sometimes, 1=always or almost always). Item properties used in the study were item format, processes of comprehension, and text complexity. In the ePIRLS assessment, the tasks were administered via computer, and students provided their answers by either selecting a response option in multiple-choice items or typing the required information in constructed-response items (Mullis et al., 2017). Each ORC task item is presented in a simulated environment similar to an Internet browser, and a teacher avatar guides students through the items. The simulated environments

Table 1. Country information of the participants in ePIRLS 2016.

Country	N	Gender		Speaking the test language at home	
		Female	Male	Always and almost always	Never and sometimes
Canada	633	321	304	509	114
Chinese Taipei	13	2	11	8	5
Denmark	26	14	12	20	6
Georgia	44	22	22	37	6
Norway	41	21	20	36	4
Portugal	24	12	12	23	1
Singapore	366	178	188	175	190
Slovenia	61	31	30	52	8
Sweden	53	30	22	45	7
United Arab Emirates	816	376	439	333	472
United States	167	75	91	140	24
Dubai, UAE	454	201	252	214	235
Abu Dhabi, UAE	196	91	105	62	130

of the ePIRLS have various features such as graphics, multiple tabs, links, pop-up windows, and animation (Mullis et al., 2017) (for sample ORC tasks, see <http://timssandpirls.bc.edu/pirls2016/international-results/epirls/take-the-epirls-assessment/>). Each student completes two ORC tasks in 80 minutes (i.e., 40 minutes per task). This study used the two released ORC tasks in the ePIRLS 2016 assessment (Mars and Elizabeth Blackwell) to calculate text complexity for the items. Table 2 shows the item properties of the released tasks.

The ORC items in ePIRLS 2016 were designed to measure four processes of comprehension (Mullis & Martin, 2015): focus on and retrieve explicitly stated information (FRE; 20% of the items); make straightforward inferences (MSI; 30% of the items); interpret and integrate ideas and information (III;

Table 2. Item descriptions in ePIRLS 2016.

Item code in ePIRLS 2016	Item format	Maximum score	Cognitive process	ATOS text complexity
E11B01M	MC	1	MSI	3.2
E11B02M	MC	1	FRE	3.1
E11B03C	CR	1	FRE	3.1
E11B04C	CR	1	FRE	3.1
E11B05M	MC	1	ECC	3.1
E11B06C	CR	2	MSI	2.7
E11B07M	MC	1	ECC	2.7
E11B08C	CR	1	FRE	2.4
E11B09C	CR	1	MSI	2.8
E11B10C	CR	2	III	2.7
E11B11M	MC	1	ECC	2.7
E11B12C	CR	1	III	2.7
E11B13C	CR	1	III	2.7
E11B14C	CR	1	MSI	2.4
E11B15C	CR	2	MSI	2.7
E11B16C	CR	3	III	2.7
E11B17C	CR	3	III	2.7
E11M01M	MC	1	MSI	4.0
E11M02C	CR	1	FRE	2.7
E11M03C	CR	1	MSI	2.8
E11M04C	CR	1	FRE	2.8
E11M05M	MC	1	FRE	2.8
E11M06M	MC	1	MSI	4.2
E11M07M	MC	1	ECC	3.7
E11M08C	CR	1	FRE	3.7
E11M09C	CR	1	FRE	2.4
E11M10M	MC	1	ECC	2.4
E11M11C	CR	2	III	2.4
E11M12M	MC	1	MSI	3.0
E11M13C	CR	1	III	3.0
E11M14C	CR	3	III	3.0
E11M15C	CR	1	MSI	2.7
E11M16C	CR	2	III	2.7
E11M17C	CR	2	ECC	2.7
E11M18C	CR	1	MSI	3.1
E11M19M	MC	1	ECC	3.1
E11M20C	CR	1	ECC	3.1

Note: E11Babc: Items related to Elizabeth Blackwell; E11Mabc: Items related to Mars.

30% of the items); and evaluate and critique content and textual elements (ECC; 20% of the items). The text complexity of the reading passages and items was computed using the ATOS readability formula (Renaissance Learning, 2012). ATOS is a reliable and valid text complexity measure that considers average sentence length, average word length, and word difficulty level (Milone, 2014). ATOS provides a metric-based grade-level scale for the reader. Using the ATOS for Text software program, the analysis of ePIRLS 2016 released items yielded ATOS levels ranging from 2.7 to 4.2.

Data analysis

Data analysis consisted of several steps. First, we checked IRT model assumptions. The ORC items in ePIRLS were jointly calibrated using a unidimensional item response theory approach (Foy & Yin, 2017). Therefore, we examined whether the unidimensionality assumption was reasonable for the released items in ePIRLS 2016. We fit a one-factor confirmatory factor analysis (CFA) model to the data using the lavaan package (Rosseel, 2012) in R (R Core Team, 2020). Given the presence of binary and polytomous items in the response data, a weighted least square mean and variance adjusted (WLSMV) estimator was used because it provides more robust results for non-normally distributed observed variables. We used root mean square error of approximation (RMSEA), Tucker-Lewis Index (TLI), and comparative fit index (CFI) to evaluate model-data fit based on Hu and Bentler's (1999) suggested cutoff values: $RMSEA \leq .06$, $TLI \geq .95$, and $CFI \geq .95$. The model results indicated that the one-factor model fit the data very well ($CFI = 0.989$, $TLI = 0.989$, and $RMSEA = 0.022$), supporting the unidimensionality assumption.

Second, we evaluated the feasibility of the EIRM approach (De Boeck & Wilson, 2004) approach in modeling response data from ePIRLS 2016. Explanatory item response models are extensions of Rasch family models (i.e., the Rasch model [Rasch, 1960/1980] for dichotomous items and the Partial Credit Model [PCM; Masters, 1982] for polytomous items). Therefore, an appropriate Rasch family model must fit the data adequately before implementing EIRM. In this study, we calibrated the items using the PCM because ePIRLS 2016 was a mixed-format test consisting of both dichotomous and polytomous items with either 2 points (i.e., 0–1–2) or 3 points (0–1–2–3). To evaluate the model fit of the PCM, we used Chalmers and Ng (2017) plausible-value variant of the Q_1 statistic ($PV-Q_1$) and obtained the empirical p -value estimates after performing a parametric bootstrapping procedure yielding $PV-Q_1^*$. The estimated item parameters and item fit statistics for the PCM are presented in the appendix. The estimated $\alpha = .01$ values indicated that all items had an acceptable fit at the significance level of $\alpha = .01$.

Third, we used the EIRM framework to examine the effects of student characteristics and item properties on student responses in ePIRLS 2016. EIRM is a flexible approach for analyzing the effects of item-related and person-related predictors simultaneously within a generalized linear mixed model. Unlike traditional IRT models focusing solely on the estimation of item and person parameters, explanatory item response models enable researchers to examine common variability in the response data based on item- and person-level predictors (Briggs, 2008; Stanke & Bulut, 2019). Therefore, the EIRM approach is valuable for providing explanatory inferences regarding the effects of item properties and person characteristics (De Boeck & Wilson, 2004). To date, EIRM has been used for analyzing both dichotomous response data (e.g., French & Finch, 2010; Kan & Bulut, 2014) and polytomous response data (e.g., Kim & Wilson, 2020; Stanke & Bulut, 2019). EIRM can also be applied to mixed-format tests with dichotomous and polytomous items. Using a set of explanatory variables (i.e., covariates), difficulty parameters for dichotomous items and threshold parameters for polytomous items can be estimated within a single explanatory item response model.

In this study, we used the explanatory form of PCM (EPCM; Tuerlinckx & Wang, 2004) for analyzing response data from ePIRLS 2016:

$$\log \left(\frac{P_{ij}}{P_{i(j-1)}} \right) = \mathbf{Z}\theta - \mathbf{X}_i' \delta_i + \tau_{ij} \quad (1)$$

where P_{ij} and $P_{i(j-1)}$ are the probabilities of receiving j and $j-1$ points on item i , θ is a vector of the latent trait (i.e., ORC in ePIRLS 2016), \mathbf{Z} is a matrix of fixed-effects predicting student ability (i.e., gender and speaking the test language at home), δ_i is a vector of the threshold between the first ($j=0$) and second ($j=1$) score categories for item i , \mathbf{X}_i' is the matrix of fixed-effects based on the properties of individual items (i.e., item format, processes of comprehension, and text complexity), and τ_{ij} is the matrix of distances between the $(j-2)/(j-1)$ threshold and the $(j-1)/j$ threshold for item i (i.e., step parameters) for a given person. This model can be further expanded to explain the threshold parameters (i.e., τ_{ij}) using explanatory variables. To analyze dichotomous and polytomous responses within the same model, we recoded the polytomous items as a series of ordered dichotomous items, as demonstrated in Table 3 (see Bulut et al., [2021] for further details on this parameterization).

In this study, three variants of EPCM were used. Model 1 included item format, processes of comprehension (i.e., cognitive complexity), and text complexity to explain the item threshold (i.e., the difficulty parameters for dichotomous items and the first step parameters for polytomous

Table 3. Preparing the item responses for EIRM analysis.

Item	Response	Recoded response 1	Recoded response 2	Recoded response 3
1	0	0	–	–
	1	1	–	–
2	0	0	NA	–
	1	1	0	–
	2	NA	1	–
3	0	0	NA	NA
	1	1	0	NA
	2	NA	1	0
	3	NA	NA	1

items). In addition to the item-level predictors in Model 1, Model 2 also included gender and speaking the test language at home to examine the effects of these variables on students’ ORC performance. The final model, Model 3, included item properties, student characteristics, and their interactions as person-by-item predictors to examine whether student characteristics moderated the effects of item properties.¹ To estimate the fixed effects for the categorical variables (i.e., all the item-level and person-level predictors except for text complexity), the variables were recoded into a set of separate dummy variables where the first level based on alphabetical ordering was used as the reference group (e.g., female for the variable “gender”). In addition to the explanatory item response models, a model with no covariates (Model 0) was estimated as a baseline model to demonstrate the impact of adding item properties and student characteristics as predictors of students’ responses to the ORC tasks in ePIRLS 2016. All models were estimated using the *eirm* package (Bulut, 2021) in R (R Core Team, 2020). Statistical significance of the predictors (i.e., item properties, student characteristics, and their interactions) was evaluated at the significance level of $\alpha = .05$. In addition, the models were compared using the relative fit indices of the Akaike Information Criterion (AIC; Akaike, 1974) and Bayesian Information Criterion (BIC; Schwarz, 1978). Smaller AIC and BIC values indicated a better model-data fit.

Results

The results of the baseline model (Model 0) and the three explanatory item response models (Models 1–3) are presented in Table 4. Compared with Model 0, the explanatory item response models appear to fit the

¹In this study, country was not considered as a third-level clustering variable because of unbalanced and small sample sizes for the selected countries, based on the multilevel modeling guidelines in Maas and Hox (2005) study.

Table 4. Results of explanatory item response models.

Item/Student characteristics	Model 0		Model 1		Model 2		Model 3	
	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>	<i>b</i>	<i>SE</i>
Intercept	0.49	0.01***	−1.40	0.05***	−1.32	0.06***	−1.37	0.08***
Threshold 2			2.63	0.01***	2.63	0.02***	2.63	0.02***
Threshold 3			2.26	0.03***	2.27	0.03***	2.27	0.03***
Item format			0.18	0.02***	0.17	0.02***	0.25	0.03***
Cognitive complexity								
MSI			−0.67	0.02***	−0.66	0.02***	−0.91	0.03***
FRE			−0.71	0.02***	−0.71	0.02***	−0.89	0.04***
III			−0.99	0.02***	−0.99	0.02***	−1.18	0.04***
Text complexity			0.52	0.02***	0.52	0.02***	0.58	0.02***
Gender					−0.09	0.03**	−0.33	0.10**
Language					0.23	0.03***	0.37	0.04***
Gender × Text complexity							0.11	0.03***
Gender × MSI							−0.16	0.04***
Gender × FRE							−0.07	0.04*
Gender × III							−0.01	0.04
Language × Item format							0.12	0.03***
Language × MSI							−0.27	0.04***
Language × FRE							−0.25	0.04***
Language × III							−0.30	0.04***
Model results								
AIC	216239		173984		171916		171839	
BIC	216259		174074		172029		172026	
Deviance	216235		173966		171894		171801	

Note: * $p < .05$. ** $p < .01$. *** $p < .001$. *b*: Estimated parameters for item and person characteristics. *SE*: Standard error of the estimated parameters. MSI: Straightforward Inferences; FRE: Explicit Information; III: Interpret Ideas. Language: Speaking the test language at home. Multiple-choice format was the reference category for item format. Evaluate and critique content and textual elements (ECC) was the reference category for cognitive complexity. Female was the reference category for gender. Never or sometimes was the reference category for speaking the test language at home. Thresholds 2 and 3 represent the distance between scores of 1 and 2 and between scores of 2 and 3, respectively.

data better based on the AIC and BIC values. In Model 1, the effects of item format, the process of comprehension, and text complexity were examined at the item level. The results indicated that the constructed-response items were more difficult than the multiple-choice items; $\exp(0.18) = 1.19$ times. Compared to the items at the ECC level, the items at the remaining levels of cognitive complexity (i.e., MSI, FRE, and III) were significantly easier. A positive coefficient for text complexity indicated that as text complexity of the ePIRLS items increased, the items became more difficult. In Model 2, gender and speaking the test language at home were also added to the model as student-level predictors. The results showed that male students were less likely to answer the items correctly; $\exp(-0.09) = 0.91$ times. Students who always or almost always speak the test language at home were more likely to answer the items correctly, $\exp(0.23) = 1.26$ times. Overall, the effects of student characteristics on students' ORC ability were statistically significant but negligible.

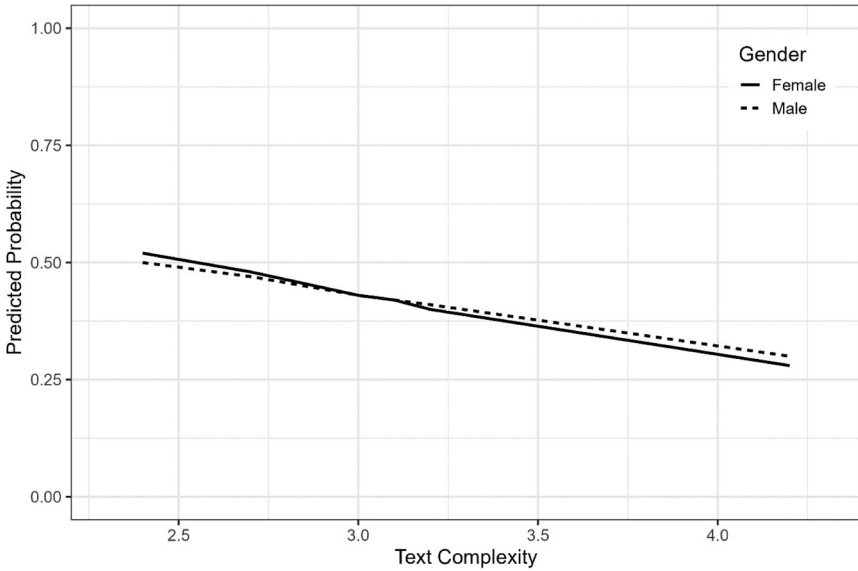


Figure 1. The interaction between gender and text complexity.

Model 3 included item properties, student characteristics, and their interactions.² The results showed that the main effects of item properties and student characteristics remained statistically significant in the model. With the inclusion of person-by-item interactions in the model, the magnitude of the estimated main effects for gender and speaking the test language at home increased. Figure 1 shows the interaction between gender and text complexity when responding to the ORC items. As text complexity in the items increased, the probability of answering the ORC items correctly decreased for both male and female students. However, the negative impact of text complexity appears to be slightly larger for female students. Figure 2 shows the interaction between gender and cognitive complexity of the items. Compared with female students, male students had a lower predicted probability of answering the ORC items across all levels of cognitive complexity. The largest difference between the male and female students occurred in the MSI items that require making straightforward inferences, followed by the difference in the FRE items that require focusing on and retrieving explicitly stated information.

Another significant interaction in Model 3 occurred between speaking the test language at home and the item format. Figure 3 shows that the

²The interactions of both gender-item format and speaking the test language at home-text complexity were not statistically significant, and thus removed from the final version of Model 3.

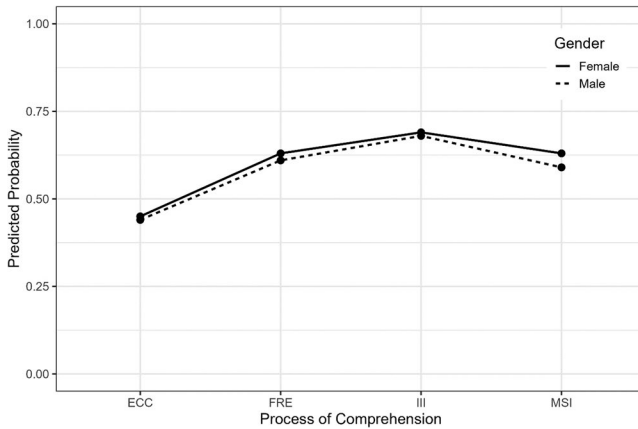


Figure 2. The interaction between gender and cognitive complexity.

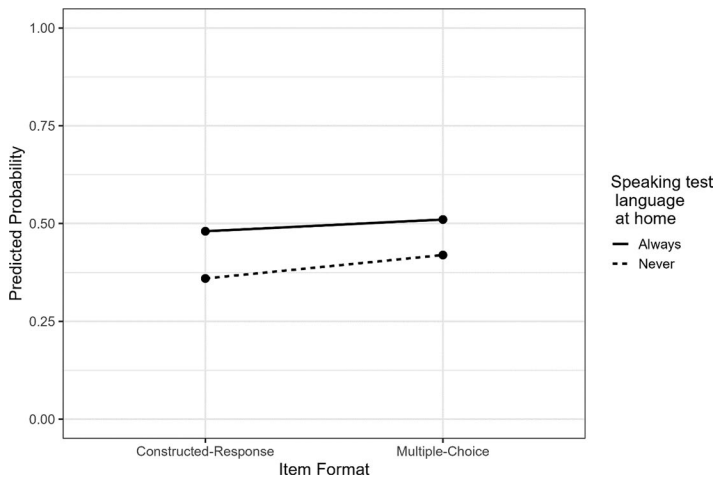


Figure 3. The interaction between speaking the test language at home and item format.

predicted probability of answering the multiple-choice items was higher than that of answering the constructed-response items for both students who always (or almost always) speak the test language at home and those who never (or sometimes) speak the test language at home. However, the increase in the predicted probability was slightly larger for students who never (or sometimes) speak the test language at home. Finally, Figure 4 shows the predicted probability for answering the ORC items as a function of cognitive complexity and speaking the test language at home. The ECC items on analyzing and evaluating content and textual elements had the lowest predicted probability among the four levels of

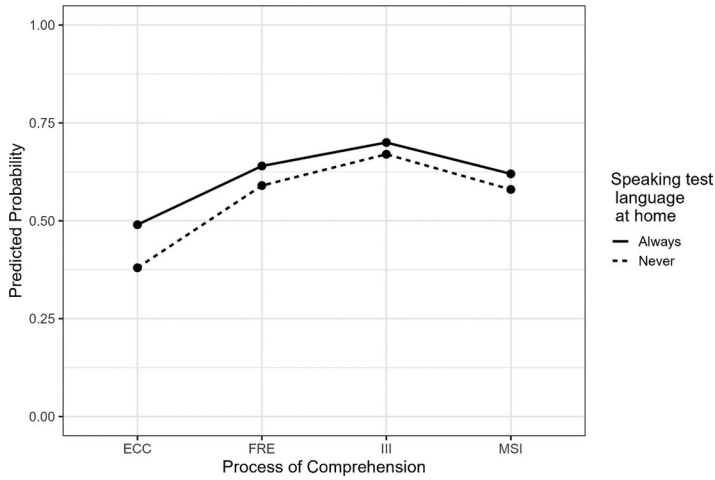


Figure 4. The interaction between speaking the test language at home and processes of comprehension.

cognitive complexity and the largest gap between the two language groups. There was a smaller gap between the language groups in the FRE, III, and MSI items compared with the ECC items. Students who always (or almost always) speak the test language at home seemed to have a higher predicted probability across all levels of cognitive complexity.

When the effects of item properties and student characteristics were evaluated within the same model (i.e., Model 2), the findings provided general insight into understanding the unique effects of these predictors at the item and student levels. However, when the interactions between item properties and student characteristics were included in the model as person-by-item predictors (i.e., Model 3), the results indicated that student characteristics moderated the effects of item properties, with a few exceptions (e.g., gender-item format and speaking the test language at home-text complexity). Unlike the student characteristics and item properties, interactions vary both within and between students. Thus, the inclusion of interactions between item properties and student characteristics revealed the differential impact of student characteristics on how students responded to the items in ePIRLS 2016. The findings showed that the difficulty levels of the ORC items seemed to vary between gender and language groups depending on item properties (i.e., item format, text complexity, and cognitive complexity).

Discussion

The increasing use of digital reading materials in education has changed how students read and learn (Coiro, 2011; Naumann, 2010). Hence,

researchers continue to investigate the effects of various predictors related to both test items and students to better understand the factors affecting students' performance in digital reading and comprehension. However, the interaction between student characteristics and item properties has received little attention due to the lack of statistical modeling approaches for estimating the effects of both student-level and item-level predictors together (De Boeck & Wilson, 2004; Stanke & Bulut, 2019). To address this gap, the current study examined the interactions between student characteristics (i.e., gender and speaking the test language at home) and item properties (item format, text complexity, and cognitive complexity of items) using students' responses to ORC items in ePIRLS 2016. The explanatory form of the PCM was used to estimate the main and interaction effects of student characteristics and item properties.

Consistent with the literature (e.g., Kanninen et al., 2019; Naumann & Sälzer, 2017; Schulz-Heidorf & Støle, 2018), this study found that female students were more likely to respond to the ORC items correctly than male students in ePIRLS 2016. Furthermore, the results showed that the probability of answering the ORC items correctly was significantly lower for students who never (or sometimes) speak the test language at home, compared with students who always (or almost always) speak the test language at home. This finding is in line with Naumann and Sälzer (2017) who examined the effect of language status using the digital reading items in PISA 2012. Regarding item properties, constructed-response items were more difficult than multiple-choice items in ePIRLS 2016. As text complexity increased, the ORC items became more difficult. These results corroborate the findings of previous studies regarding the higher difficulty of constructed-response items than multiple-choice items (Schulz-Heidorf & Støle, 2018) and the positive association between text complexity and item difficulty (Amendum et al., 2018; Spencer et al., 2019). Furthermore, items focusing on "evaluating and critiquing a text" were more challenging than items on "retrieving explicitly stated information," "making straightforward inferences," and "interpreting information." This finding is consistent with the anticipated relationship between cognitive complexity and item difficulty, as highlighted in the ePIRLS 2016 assessment framework (Mullis & Martin, 2015).

The main contribution of this study comes from the findings regarding the interactions between student characteristics and item properties. The presence of significant person-by-item interactions indicates that the probability of answering the items correctly (or receiving partial credit) may not be the same across all levels of item properties and student characteristics. Also, investigating item properties and their interactions with student characteristics extends the ORC literature (Buerger et al., 2019; Kulesz et al., 2016; Toyama, 2019). The findings of this study show

that female students were more successful than male students in answering the ORC items. However, the gap between male and female students appears to change depending on text and cognitive complexity of the ORC items. Unlike text and cognitive complexity, item format (i.e., multiple-choice vs. constructed-response items) did not seem to affect how male and female students responded to the ORC items. Overall, these findings demonstrate the necessity of investigating interactions to better understand the disparity in ORC performance between male and female students.

This study also found significant results regarding the effects of speaking the test language at home. Students who never (or sometimes) speak the test language at home experienced difficulties in answering the ORC items, compared to students who always (or almost always) speak the test language at home (Model 2). For example, when answering multiple-choice items, students who never (or sometimes) speak the test language at home appeared to improve their performance more significantly compared with students who always (or almost always) speak the test language at home. Furthermore, the gap between the language groups was more prominent in the ECC items that required a critique of the content and textual elements. In contrast, the gap became much smaller in the items with relatively lower cognitive complexity (i.e., FRE, III, and MSI).

Implications for practice

This study has several implications for practice. First, this study reiterates the importance of controlling for item properties to mitigate construct-irrelevant variation in test scores based on student characteristics such as gender and language. The findings of our study revealed that item properties matter when examining group differences in ORC regarding gender and language. For example, the achievement gaps between gender and language groups seem to vary depending on text complexity and format of the ORC tasks. This finding suggests that the interpretation of differences between gender and language groups might be incomplete when the effects of item properties are neglected. Standard 3.1 from the *Standards for Psychological and Educational Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (AERA, APA, & NCME), 2014) emphasizes the importance of “being knowledgeable about group differences that may interfere with the precision of scores and the validity of test score inferences” (p. 63). The *Standards* further stress that test developers need to avoid item characteristics that might disrupt the performance of relevant subgroups of examinees. Thus, when evaluating whether

ORC items are psychometrically sound, the interactions between item properties and student characteristics should be carefully examined to ensure that the items are appropriate for all subgroups of the target population.

Second, the findings of this study show that compared with conventional statistical approaches (e.g., linear modeling), the EIRM framework offers several methodological advantages in extracting meaningful and practical information from response data based on the interactions between item properties and student characteristics. With increasing numbers of educational assessments using mixed-format tests with dichotomous and polytomously items, explanatory item response models can help researchers explore the impact of item properties (e.g., text complexity and item format) on responses processes underlying each item across different subgroups of examinees. Furthermore, when piloting or field-testing ORC items, test developers can employ the EIRM approach to evaluate the appropriateness of ORC items for the target examinee population. For example, one could evaluate the impact of text complexity of ORC items, as well as stimuli included in the items (e.g., reading passages and any associated graphics), on the performance of English language learners and their native-English-speaking peers and remove items or stimuli leading to unintended group differences from the assessment.

Third, this study shows that cognitive complexity (i.e., the process of comprehension) of ORC items is a significant predictor of the achievement gap between gender and language groups. For example, the findings of this study indicate that the language-related performance gap in ORC becomes more evident in constructed-response items and items with higher cognitive complexity. Therefore, researchers and educators should help struggling students improve their ORC skills, particularly in cognitively complex texts. Previous research suggests that the use of online reading environments with digital media and texts can improve students' reading comprehension skills (). Thus, instead of using printed reading materials, educators should consider using online learning environments in which students can practice their ORC skills more effectively with engaging test items at different levels of cognitive complexity.

Limitations and future research

The current study has several limitations. First, the participants of this study were fourth graders from different countries who completed ePIRLS 2016 in English. Therefore, the findings of this study may not be generalized to student populations who completed ePIRLS 2016 in other languages. Second, considering potential differences among

education systems of the participating countries in ePIRLS 2016, country-specific variables (e.g., the availability of ORC-focused instruction) could further explain complex interactions between item properties and student characteristics. Future research should harness country-level variables when examining student performance in ORC. Third, this study only examined the two released tasks in ePIRLS 2016. Future research could employ ORC assessments involving more reading tasks in order to investigate the effects of task-related variables (e.g., time spent on each task, content of reading tasks, and the number of interactions with readings tasks). Lastly, this study utilized several item properties as fixed effects to account for item difficulties in ePIRLS 2016. A recent study by Kim and Wilson (2020) indicated that explanatory item response models with a residual term based on random item effects could enhance the prediction of the item difficulty parameters. Therefore, future studies could employ explanatory item response models with random item effects when investigating the impact of item properties on ORC performance.

ORCID

Hatice Cigdem Bulut  <http://orcid.org/0000-0003-2585-3686>

Okan Bulut  <http://orcid.org/0000-0001-5853-1267>

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–723. <https://doi.org/10.1109/TAC.1974.1100705>
- Amendum, S. J., Conradi, K., & Hiebert, E. (2018). Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension. *Educational Psychology Review*, 30(1), 121–151. <https://doi.org/10.1007/s10648-017-9398-2>
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
- Briggs, D. C. (2008). Using explanatory item response models to analyze group differences in science achievement. *Applied Measurement in Education*, 21(2), 89–118. <https://doi.org/10.1080/08957340801926086>
- 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>
- Bulut, O. (2021). *eirm: Explanatory item response modeling for dichotomous and polytomous item responses* [Computer software]. <http://CRAN.R-project.org/package=eirm>

- Bulut, O., Gorgun, G., & Yildirim-Erbasli, S. N. (2021). Estimating explanatory extensions of dichotomous and polytomous Rasch models: The eirm package in R. *Psych*, 3(3), 308–321. <https://doi.org/10.3390/psych3030023>
- Caro, D. H., McDonald, T. J., & Williams, J. (2009). Socio-economic status and academic achievement trajectories from childhood to adolescence. *Canadian Journal of Education*, 32(3), 558–559.
- Chalmers, R. P., & Ng, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, 41(5), 372–387. <https://doi.org/10.1177/0146621617692079>
- Chen, G., Cheng, W., Chang, T. W., Zheng, X., & Huang, R. (2014). A comparison of reading comprehension across paper, computer screens, and tablets: Does tablet familiarity matter? *Journal of Computers in Education*, 1(2-3), 213–225. <https://doi.org/10.1007/s40692-014-0012-z>
- Coiro, J. (2011). Predicting reading comprehension on the Internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research*, 43(4), 352–392. <https://doi.org/10.1177/1086296X11421979>
- Coiro, J. (2014). Online reading comprehension: Challenges and opportunities. *Texto Livre: Linguagem e Tecnologia*, 7(2), 30–43. <https://doi.org/10.17851/1983-3652.7.2.30-43>
- Coiro, J., & Dobler, E. (2007). Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the Internet. *Reading Research Quarterly*, 42(2), 214–257. <https://doi.org/10.1598/RRQ.42.2.2>
- Collier, T., Morell, L., & Wilson, M. (2018). Exploring the item features of a science assessment with complex tasks. *Measurement*, 114, 16–24. <https://doi.org/10.1016/j.measurement.2017.08.039>
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. *Statistics for social science and public policy*. Springer.
- DeStefano, D., & LeFevre, J. A. (2007). Cognitive load in hypertext reading: A review. *Computers in Human Behavior*, 23(3), 1616–1641. <https://doi.org/10.1016/j.chb.2005.08.012>
- Drew, S. V. (2012). Open up the ceiling on the Common Core State Standards: Preparing students for 21st-century literacy. *Journal of Adolescent & Adult Literacy*, 56(4), 321–330. <https://doi.org/10.1002/JAAL.00145>
- Elleman, A. M., & Oslund, E. L. (2019). Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences*, 6(1), 3–11. <https://doi.org/10.1177/2372732218816339>
- French, B. F., & Finch, W. H. (2010). Hierarchical logistic regression: Accounting for multilevel data in DIF detection. *Journal of Educational Measurement*, 47(3), 299–317. <https://doi.org/10.1111/j.1745-3984.2010.00115.x>
- Foy, P., & Yin, Y. (2017). Scaling the PIRLS 2016 achievement data. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 12.1–12.38). Boston College, TIMSS & PIRLS International Study Center website. <https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-12.html>
- Gao, Y., Cui, Y., Bulut, O., Zhai, X., & Chen, F. (2022). Examining adults' web navigation patterns in multi-layered hypertext environments. *Computers in Human Behavior*, 129, 107142. <https://doi.org/10.1016/j.chb.2021.107142>
- Gil-Flores, J., Torres-Gordillo, J. J., & Perera-Rodríguez, V. H. (2012). The role of online reader experience in explaining students' performance in digital

- reading. *Computers & Education*, 59(2), 653–660. <https://doi.org/10.1016/j.compedu.2012.03.014>
- Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, 30(5), 394–411. <https://doi.org/10.1177/0146621606288554>
- Heo, M., & Toomey, N. (2020). Learning with multimedia: The effects of gender, type of multimedia learning resources, and spatial ability. *Computers & Education*, 146, 103747. <https://doi.org/10.1016/j.compedu.2019.103747>
- International Reading Association. (2009). *IRA position statement on new literacies and 21st century technologies*. Author. www.reading.org/General/AboutIRA/PositionStatements/21stCenturyLiteracies.aspx
- Jerrim, J. (2016). PISA 2012: How do results for the paper and computer tests compare? *Assessment in Education: Principles, Policy & Practice*, 23(4), 495–518. <https://doi.org/10.1080/0969594X.2016.1147420>
- Kan, A., & Bulut, O. (2014). Examining the relationship between gender DIF and language complexity in mathematics assessments. *International Journal of Testing*, 14(3), 245–264. <https://doi.org/10.1080/15305058.2013.877911>
- Kannianen, L., Kiili, C., Tolvanen, A., Aro, M., & Leppänen, P. H. (2019). Literacy skills and online research and comprehension: Struggling readers face difficulties online. *Reading and Writing*, 32(9), 2201–2222. <https://doi.org/10.1007/s11145-019-09944-9>
- Kim, J., & Wilson, M. (2020). Polytomous item explanatory item response theory models. *Educational and Psychological Measurement*, 80(4), 726–755. <https://doi.org/10.1177/0013164419892667>
- Kulesz, P. A., Francis, D. J., Barnes, M. A., & Fletcher, J. M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology*, 108(8), 1078–1097. <https://doi.org/10.1037/edu0000126>
- Lankshear, C., & Knobel, M. (2011). *New literacies*. Open University Press.
- Lauterman, T., & Ackerman, R. (2014). Overcoming screen inferiority in learning and calibration. *Computers in Human Behavior*, 35, 455–463. <https://doi.org/10.1016/j.chb.2014.02.046>
- Leu, D. J., Forzani, E., Rhoads, C., Maykel, C., Kennedy, C., & Timbrell, N. (2015). The new literacies of online research and comprehension: Rethinking the reading achievement gap. *Reading Research Quarterly*, 50(1), 37–59. <https://doi.org/10.1002/rrq.85>
- Leu, D. J., Kinzer, C. K., Coiro, J. L., & Cammack, D. W. (2004). Toward a theory of new literacies emerging from the Internet and other information and communication technologies. *Theoretical Models and Processes of Reading*, 5(1), 1570–1613. <https://doi.org/10.1598/0872075028.54>
- Leu, D. J., Kinzer, C. K., Coiro, J., Castek, J., & Henry, L. A. (2013). New literacies and the new literacies of online reading comprehension: A dual level theory. In N. Unrau & D. Alvermann (Eds.), *Theoretical models and process of reading* (6th ed., pp. 1150–1181). IRA.
- Leu, J. D. (2017). *ePIRLS: An international assessment of reading for new times*. Boston College, TIMSS & PIRLS International Study Center. <http://timssand-pirls.bc.edu/pirls2016/international-results/epirls/foreword/>

- Liu, I., & Ko, H. (2016). The relationship among ICT skills, traditional reading skills and online reading ability. *International Association for Development of the Information Society*, 287–291.
- Maas, C., & Hox, J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86–92. <https://doi.org/10.1027/1614-2241.1.3.86>
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149–174. <https://doi.org/10.1007/BF02296272>
- Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, 47(3), 235–258. <https://doi.org/10.1002/rrq.019>
- Milone, M. (2014). *Development of the ATOS: Readability formula*. Renaissance Learning, Inc. <http://doc.renlearn.com/KMNet/R004250827GJ11C4.pdf>
- Moon, J. A., Keehner, M., & Katz, I. R. (2019). Affordances of item formats and their effects on test-taker cognition under uncertainty. *Educational Measurement: Issues and Practice*, 38(1), 54–62. <https://doi.org/10.1111/emip.12229>
- Moon, J. A., Sinharay, S., Keehner, M., & Katz, I. R. (2020). Investigating technology-enhanced item formats using cognitive and item response theory approaches. *International Journal of Testing*, 20(2), 122–124. <https://doi.org/10.1080/15305058.2019.1648270>
- Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 assessment framework* (2nd ed.). Boston College, TIMSS & PIRLS International Study Center. <http://timssandpirls.bc.edu/pirls2016/framework.html>
- Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *ePIRLS 2016 international results in online informational reading*. Boston College, TIMSS & PIRLS International Study Center. <http://timssandpirls.bc.edu/pirls2016/international-results/>
- Naumann, J. (2010, May). *Predicting comprehension of electronic reading Tasks: The impact of computer skills and reading literacy*. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO.
- Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior*, 53, 263–277. <https://doi.org/10.1016/j.chb.2015.06.051>
- Naumann, J. (2019). The skilled, the knowledgeable, and the motivated: Investigating the strategic allocation of time on task in a computer-based assessment. *Frontiers in Psychology*, 10, 1429 <https://doi.org/10.3389/fpsyg.2019.01429>
- Naumann, J., & Goldhammer, F. (2017). Time-on-task effects in digital reading are non-linear and moderated by persons' skills and tasks' demands. *Learning and Individual Differences*, 53, 1–16. <https://doi.org/10.1016/j.lindif.2016.10.002>
- Naumann, J., & Sälzer, C. (2017). Digital reading proficiency in German 15-year olds: Evidence from PISA 2012. *Zeitschrift für Erziehungswissenschaft*, 20(4), 585–603. <https://doi.org/10.1007/s11618-017-0758-y>
- OECD. (2014). *PISA 2012 results: Creative problem solving. Students' skills in tackling real-life problems* (Vol. V). OECD Publishing. <https://doi.org/10.1787/9789264208070-en>
- OECD. (2019). *Skills matter: Additional results from the survey of adult skills. OECD skills studies*. OECD Publishing. <https://doi.org/10.1787/1f029d8f-en>
- OECD. (2021). *21st-century readers: Developing literacy skills in a digital world, PISA*. OECD Publishing. <https://doi.org/10.1787/a83d84cb-en>

- Potter, W. J. (2018). *Media literacy*. Sage Publications.
- Rasch, G. (1960/1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) with foreword and afterword by B.D. Wright. The University of Chicago Press.
- R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Renaissance Learning. (2012). *Text complexity: Accurate estimates and educational recommendations*. Wisconsin Rapids. <http://doc.renlearn.com/KMNet/R00548821C95879F.pdf>
- Rosseel, Y. (2012). lavaan : An R package for structural equation modeling. *Journal of Statistical Software*, 48(2). <https://doi.org/10.18637/jss.v048.i02>
- Schulz-Heidorf, K., & Støle, H. (2018). Gender differences in Norwegian PIRLS 2016 and ePIRLS 2016 results at test mode, text and item format level. *Nordic Journal of Literacy Research*, 4(1), 167–183. <https://doi.org/10.23865/njlr.v4.1270>
- Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly*, 50(2), 219–232. <https://doi.org/10.1002/rrq.92>
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Spencer, M., Gilmour, A. F., Miller, A. C., Emerson, A. M., Saha, N. M., & Cutting, L. E. (2019). Understanding the influence of text complexity and question type on reading outcomes. *Read Writ*, 32(3), 603–637. <https://doi.org/10.1007/s11145-018-9883-0>
- Stanke, L., & Bulut, O. (2019). Explanatory item response models for polytomous item responses. *International Journal of Assessment Tools in Education*, 6(2), 259–278. <https://doi.org/10.21449/ijate.515085>
- Strand, O., & Schvippert, K. (2019). The impact of home language and home resources on reading achievement in ten-year-olds in Norway, PIRLS 2016. *Nordic Journal of Literacy Research*, 5(1), 1–17. <https://doi.org/10.23865/njlr.v5.1260>
- Toyama, Y. (2019). *What makes reading difficult? An investigation of the contribution of passage, task, and reader characteristics on item difficulty, using explanatory item response models* [Doctoral dissertation]. University of California, Berkeley.
- Tuerlinckx, F., & Wang, W.-C. (2004). Models for polytomous data. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 75–109). Springer-Verlag.

Appendix

Item parameters and fit indices for the ePIRLS 2016 sample used in this study.

Items	b_1	b_2	b_3	$PV - Q_1$	$PV - Q_1^*$
E11B01M	-1.187			97.57	0.01
E11B02M	-2.268			17.81	0.01
E11B03C	0.918			14.57	0.02
E11B04C	-0.778			20.51	0.01
E11B05M	0.427			20.53	0.01
E11B06C	-0.605	0.000		22.67	0.12
E11B07M	-0.193			33.50	0.02
E11B08C	-1.177			56.24	0.01
E11B09C	-0.870			125.60	0.01
E11B10C	-0.144	0.813		41.88	0.02
E11B11M	0.790			19.21	0.01
E11B12C	0.394			47.87	0.01
E11B13C	-0.494			166.33	0.01
E11B14C	0.539			30.12	0.01
E11B15C	-0.579	-0.351		83.67	0.01
E11B16C	-0.977	-0.581	-0.351	428.24	0.01
E11B17C	-0.756	-0.237	0.733	145.63	0.01
E11M01M	0.795			399.72	0.01
E11M02C	-2.330			20.29	0.01
E11M03C	-0.934			14.44	0.02
E11M04C	-0.582			96.08	0.02
E11M05M	-2.154			50.35	0.01
E11M06M	-0.371			23.07	0.01
E11M07M	0.038			57.50	0.01
E11M08C	-0.957			73.05	0.01
E11M09C	-0.477			12.76	0.08
E11M10M	-1.045			21.69	0.01
E11M11C	-0.267	0.982		52.45	0.01
E11M12M	-0.899			27.51	0.01
E11M13C	0.989			27.37	0.01
E11M14C	0.225	-0.530	0.636	33.23	0.04
E11M15C	-1.359			60.83	0.01
E11M16C	0.427	-0.904		26.20	0.02
E11M17C	0.246	-0.055		166.57	0.02
E11M18C	0.175			27.12	0.01
E11M19M	0.372			40.49	0.01
E11M20C	0.730			16.08	0.01

Note: $PV - Q_1$ is the plausible-value imputation variant of the Q_1 statistic. $PV - Q_1^*$ is the empirical p value estimate for $PV - Q_1$ across 1000 parametric bootstrap samples.