Comparing Linear and Nonlinear Reading Assessments in PIRLS 2016

Purya Baghaei (corresponding author)

Research and Analysis Unit, IEA Hamburg, Hamburg, Germany

Email: purya.baghaei@iea-hamburg.de

Elpis Grammatikopoulou

Department of Education and Special Education, University of Gothenburg, Sweden

Stefan Johansson

Department of Education and Special Education, University of Gothenburg, Sweden

Rolf Strietholt

Research and Analysis Unit, IEA Hamburg, Hamburg, Germany

Abstract

Are linear paper-based and nonlinear computer-based reading comprehension measuring the same cognitive constructs? Nonlinear hypertext reading may demand different skills than linear text comprehension, as it requires synthesizing and integrating information from multiple sources. However, research on whether these two forms of comprehension represent distinct constructs remains limited. While conceptual distinctions exist between understanding a single linear text and processing multiple texts, empirical evidence supporting these differences is scarce. To address this gap, we examined the dimensional structure of paper-based linear and computer-based nonlinear text comprehension using data from the PIRLS 2016 international large-scale assessment. PIRLS 2016 evaluated reading comprehension at the end of grade 4, with each student completing both a linear paper-based test and a nonlinear electronic test. Comparing a unidimensional model, a 2-dimensional correlated factor model, and a bifactor IRT model across 16 countries and benchmarking entities showed that the bifactor model, which accommodates a general reading comprehension factor along with nonlinear reading as a specific factor, has a better fit than the other models. The ECVs of the general factor ranged from .80 to .89 across all the countries while the ECVs of the specific factor were between .07 and .18. Nevertheless, the specific factor reliabilities ranged from .36 to .58 which indicates that the variations in the specific factor cannot be just noise and could reflect variations in reading nonlinear texts. These findings revealed that although linear paper-based and nonlinear computer-based reading share a substantial amount of variance, as evidenced by a strong general reading comprehension factor, understanding nonlinear hypertexts requires specific skills that cannot be overlooked and should be explicitly addressed. The implications of the study for the testing and teaching of reading comprehension are discussed.

Keywords: linear reading, nonlinear reading, dimensionality, bifactor model

Introduction

The current study investigates whether sequential linear reading on paper is the same as nonlinear reading on screen. Reading comprehension is a fundamental skill critical for students' academic success, personal growth, and lifelong learning (National Institute of Child Health and Human Development, 2000). Beyond its immediate academic importance, proficient reading fosters cognitive development (including language acquisition, comprehension, and critical thinking), benefits individuals and society by promoting empathy and communication (Cunningham & Stanovich, 1997; Kidd & Castano, 2013), and supports mental, social, and emotional well-being (Billington et al., 2010; Zuckerman & Augustyn, 2011).

In recent decades, reading practices have evolved, largely driven by the internet and the widespread use of digital devices (Empson, 2014; Cartelli, 2012). While paper-based reading remains prevalent in many schools, the availability of digital learning platforms such as Moodle, Brightspace, and Blackboard has fueled an increase in screen-based reading in educational contexts. Additionally, research indicates that around 90% of teenagers are regularly online, relying on devices such as computers, smartphones, and tablets for both leisure and educational activities (Madden et al., 2013; OECD, 2011). As a result, nonlinear digital reading is becoming increasingly important alongside traditional linear reading.

The distinction between linear and nonlinear texts is also increasingly reflected in assessment practices. For example, as digital skills become increasingly vital in society and the workforce, major international reading assessments—including the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA)—have begun evaluating students' competencies in digital reading (OECD, 2023; Mullis et al., 2017). Meanwhile, many educational systems are transitioning at least some of their national exams to digital formats, reflecting the growing importance of screen-based literacy in contemporary education (e.g., Daane et al., 2005; KMK, 2016).

Comparing paper and digital reading

Digital reading is often distinguished from traditional reading due to key differences in format, components, and effects (Leu et al., 2004; Coiro, 2003). Physically, printed texts have fixed pages that act as visual boundaries, while digital texts can be continuous, fragmented, or multi-textual (Chen, 2017). Printed texts are limited to basic components like images, whereas digital texts may integrate diverse elements such as images, sound effects, music, and pop-ups, creating a holistic experience for the reader or viewer (Kress, 2003). Distractions also differ between the two modes. In paper-based reading, distractions are typically external, requiring readers to block out their surroundings to maintain focus. Conversely, in digital reading, distractions may be internal, arising from features like notifications, hyperlinks, and multimedia. These internal interruptions, combined with the abundance of engaging content, can make it harder for readers to stay focused, even for highly disciplined learners (Rosén & Gustafsson, 2016). Subsequently, the skills required for these two modes may also differ. Managing distractions in digital texts demands not only focus but also the ability to critically evaluate the relevance of incoming information, requiring a more active engagement than traditional paper-based reading. Coiro (2011) argues that traditional reading skills, while essential, are insufficient for comprehending online information. Digital reading can require additional competencies, including generating research questions, identifying important problems, navigating search engines efficiently, or critically evaluating the reliability of sources. Moreover, readers must also synthesize information from multiple sources and communicate effectively using tools suited to specific platforms. In today's information age, where vast amounts of online content are readily accessible and anyone can publish, understanding the skills and strategies needed for reading literacy is crucial for individuals and educational systems alike. The ability to navigate this evolving landscape of reading is essential, as students' academic success increasingly depends on mastering these digital literacy skills (Coiro, 2003; Coiro, 2011).

Linear vs. Nonlinear Reading

In the context of reading digital texts, a distinction is often made between linear and nonlinear reading, and sometimes digital and nonlinear reading are even used interchangeably (Hahnel et al., 2016). Nonlinear reading refers to the process of reading text in a non-sequential manner, where the reader does not follow a traditional, top-to-bottom linear path through the content. This type of reading is common in digital environments, where readers interact with hyperlinked text, images, sidebars, and multimedia. Nonlinear reading allows users to jump between sections, skim content, or explore different parts of a document or webpage based on interest or need, rather than following a fixed order. It contrasts with traditional linear reading, which involves progressing through a text from start to finish.

While the distinction between these modes is clear, not all digital reading is nonlinear. For instance, reading a novel on an e-reader like Amazon's Kindle is linear reading. Nonlinear reading is mainly characterized by hypertext and interactivity. Hypertext refers to a system of managing and organizing information through text that contains links (called "hyperlinks") to other texts or resources. These links allow users to navigate between related documents or pages by simply clicking on them. Hypertext creates a non-linear way of accessing information, where readers can jump from one topic to another based on their interests or needs. While hypertext has been noted for mirroring the associative nature of human memory, providing a dynamic way of accessing and organizing information (Shneiderman, 1989), research has also suggested that nonlinear reading demands additional cognitive effort. Readers must continually make decisions about how to navigate the content and assess the relevance of the information (Zumbach & Mohraz, 2007).

Although researchers have studied linear reading for many decades, there has been far less research on nonlinear computer-based reading, despite the increasing prevalence of digital reading material. Some researchers have argued that "the strategies requisite for comprehending traditional printed text are not the same strategies required to comprehend computerized texts" (Murphy et al., p. 528). However, the psychometric distinctness and separability of linear and nonlinear reading—specifically, reading hypertext versus reading on paper—have not been extensively studied.

Does the mode of reading play a role in comprehension and learning?

Meta-analyses suggest that readers perform differently on reading assessments when reading on paper compared to reading digitally (Furenes et al., 2021; Clinton, 2019; Delgado et al., 2018; Kong et al., 2018). Specifically, Delgado et al. (2018), in their meta-analysis of 54 studies conducted between 2000 and 2017, highlight the advantages of paper-based reading, including better concentration, memory retention, and deeper understanding. Their findings emphasize that paper reading leads to more profound comprehension and learning for both informational and mixed texts, although this advantage does not extend to purely narrative texts. Similarly, Clinton (2019) conducted a meta-analysis and found a performance advantage for paper-based reading over digital reading, particularly for expository or mixed texts. However, like Delgado, Clinton observed no significant difference for purely narrative texts. Kong et al. (2018) in their meta-analysis of 17 studies, found that reading on paper outperformed reading on screen, although they did not distinguish between text genres. In the same vein, Furenes et al. (2021) concluded in their meta-analysis that children generally scored lower in comprehension when reading digital books compared to print books. They also noted that traditional print books were more effective for comprehension, particularly when an adult mediated the reading session**.** However, digital books with story-congruent enhancements could, in some cases, surpass print books in effectiveness. Additionally, Ronconi et al. (2022) conducted a smaller study with 150 eighth-grade participants, focusing on informational texts to examine the effect of reading mode. Their findings revealed that students performed significantly better in comprehending main ideas when reading on paper, although the reading mode had minimal impact on understanding key points.

While the studies above highlight the benefits of paper-based reading, other research has found favourable results for digital reading, particularly when interactive features are included. Egert et al. (2022) conducted a meta-analysis of 17 studies focusing on researcher-developed e-books with interactive digital features for children aged 3-7 years. Their findings indicated that these digital books, particularly those designed to engage students, resulted in better reading performance compared to traditional print books. The analysis revealed that e-book interventions significantly improved language outcomes in young children, particularly in preschool and kindergarten settings, with notable improvements over regular childcare activities and print book reading. Similarly, Savva et al., (2022) in their meta-analysis of experimental studies, spanning between 2008 and 2021 found a small positive effect for e-books when compared with print books on language and literacy development (alphabet knowledge, phonological awareness, print aware-ness, word reading and writing), though no significant effect on story comprehension. Additionally, Kaman and Ertem (2018) used mixed methods to examine the effect of digital reading on comprehension, fluency, and attitudes. Their study, involving 75 4th grade students, found that digital reading significantly improved both reading comprehension and fluency. Hsiao and Chen (2015) used data from 60 third-grade Chinese pupils to examine the relationship between digital reading and reading comprehension. Results indicated that e-book reading, when compared with paper-based reading, is associated with higher comprehension. Aparicio et al (2022), in a small scale study on the reading mode effect on both expository and narrative texts, involving 52 students in grade five, found no significant effect. Similarly, Schwabe et al. (2022) conducted a meta-analysis of 32 studies examining also the reading mode effect on reading comprehension of narrative texts. These results align with Clinton (2019) and Delgado et al. (2018), who found no significant reading mode effect on comprehension for narrative texts, but also with Egert et al. (2022) who highlight the potential advantage of digital texts when multimedia and interactive functions are involved.

Regarding test results from international large-scale assessments, Mullis and Martin (2015) in their report on the 2016 ePIRLS and PIRLS assessments, observed that students from most participating countries who took both the electronic and paper-based tests achieved higher scores in the electronic assessment compared to the paper-based assessment. However, Jerrim et al.'s (2018) analysis of PISA 2015 field trial data from Germany, Ireland, and Sweden showed an opposite pattern: students scored lower in digital assessments compared to paper-based ones.

While the research on reading mode (digital vs. paper) remains inconclusive, it highlights the need to carefully consider the cognitive processes involved in both *linear* and *nonlinear* text comprehension, particularly as digital texts become increasingly prevalent in educational settings.

The Present Study

Research on the comparability of linear and nonlinear reading assessments is very scarce. This gap in research highlights the need for further investigation into whether linear reading assessments can be considered equivalent to nonlinear reading.

This study examines this issue within the framework of PIRLS, a major international large-scale assessment that evaluates reading comprehension at the end of primary school. Since its introduction as a paper-based assessment in 2001, PIRLS has been conducted in five-year cycles, with up to 66 education systems participating worldwide. As the study is repeated over time, there is strong interest in tracking changes in reading performance. However, PIRLS 2016 introduced ePIRLS, a computer-based nonlinear component designed to assess students' comprehension of online-style informational texts.

This shift raises questions about the dimensionality and comparability of the assessments across delivery modes. It remains unclear whether differences in format affect test performance beyond reading comprehension as measured in the conventional linear paper-based PIRLS assessments. Likewise, ePIRLS, designed for nonlinear reading, includes graphical features, pop-ups, and a teacher avatar, potentially influencing students' engagement and navigation strategies.

To address these uncertainties, this study examines the extent to which paper-based linear and computer-based nonlinear reading assessments measure the same skills. By comparing performance across PIRLS and ePIRLS, it evaluates whether computer-based and paper-based reading comprehension involve the same underlying competencies or if distinct cognitive demands come into play.

Data and Method

Progress in International Reading Literacy Study (PIRLS)

IEA's PIRLS is an international trend study designed to measure and compare the reading ability of fourth-grade students around the globe over time. Launched in 2001 as a follow-up to IEA's 1991 Reading Literacy Study and to the older IEA's 1971 'The Reading Comprehension Study', this assessment is administered every five years to participating countries. Its goal is to provide insights into students' reading achievement and their ability to apply what they understand from texts in real-life contexts or school projects. It also gathers data on factors that influence reading development, such as home environments, school resources, and teaching practices. PIRLS provides comparative data across countries to evaluate and improve educational outcomes related to reading literacy. PIRLS highlights two broad reading purposes (Mullis & Martin, 2015): “literary experience” and reading to “acquire and use information” (AUI) with 50% of the items allocated to each reading purpose. The two reading purposes are fulfilled with the choice of the texts. For literary experience, narrative fiction is used and for acquire and use information, scientific texts are employed (Mullis & Martin, 2015).

In response to the widespread availability of digital media and the importance of nonlinear reading on electronic devices, computer-based reading assessment was incorporated in the fourth cycle of the Progress in International Reading Literacy Study in 2016 (PIRLS, Mullis et al., 2017). On the understanding that accurate assessment of digital reading skill is vital to track students' reading progress and deliver suitable, timely instruction and interventions, ePIRLS was introduced in 2016. ePIRLS is an extension of the PIRLS that focuses specifically on assessing students' ability to read and comprehend online informational texts, without including the ‘literary experience’ purpose. It evaluates how well fourth-grade students can navigate and understand digital environments, such as websites, and locate, interpret, and use information presented online. This assessment reflects the growing importance of digital literacy in modern education and helps measure students' proficiency in reading digital materials, which is increasingly relevant in today's technology-driven world (Mullis et al., 2017). Table 1 highlights the similarities and differences between PIRLS and ePIRLS.

Table 1

Comparison of PIRLS and ePIRLS

	PIRLS	**ePIRLS**
Format	Paper-based	Computer-based
Structure	Linear	Nonlinear
Purpose	acquire and use information + literary experience[a]	acquire and use information
Material	6 acquire and use information texts + 6 literary experience texts[a]	4 school-based reading tasks each consists of 2-3 different websites totaling 5-10 web pages
Assignment	1 booklet (composed of 1 acquire and use information text and 1 literary experience texts[a])	Online reading tasks
Type of questions	Multiple choice (40%) + open-ended (60%)	Multiple choice (40%) + open-ended (60%)
Testing time	80 minutes	80 minutes

Note. [a] = literary experience material was not used in the present study

Both assessments use a combination of multiple-choice and open-ended questions to evaluate the same four reading comprehension processes: retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas, and evaluating and critiquing content. However, there are also differences across the assessments because ePIRLS is administered in a simulated nonlinear online environment instead of on paper (Mullis & Martin, 2015)[1]. Further, although both assessments target the same reading comprehension processes, they differ in the content of their passages and questions. As such,

[1] For detailed information concerning the distribution of item formats and subskills refer to Mullis and Prendergast (2017).

ePIRLS should not be regarded as merely a digital version of PIRLS with identical materials. It is important to avoid the misconception that ePIRLS is simply the computer-based format of PIRLS.

The technology-enhanced features of ePIRLS introduce new skills not required in paper tests, such as navigating between pages, evaluating digital sources, and managing screen-based distractions. As a result, test performance might reflect not just reading comprehension, but also students' digital literacy and familiarity with computer-based environments. This introduces potential threats to comparability, as differences in digital navigation skills can affect test outcomes even if reading ability is the same.

PIRLS strives for comparable trend measures to trace reading achievement progression over time. The structure of PIRLS 2016, i.e., the combination of paper-based linear and computer-based nonlinear texts, raises questions about the assessments' dimensionality and comparability across delivery modes. While the PIRLS assessment itself is considered linear, the ePIRLS incorporates nonlinear elements, such as clickable images and interactive features. However, despite these nonlinear components, the ePIRLS assessment is structured with linear guidance, as a teacher avatar instructs students to follow a specific linear path/sequence. This means that students do not have the autonomy to navigate freely but are directed by the avatar. This makes the ePIRLS test neither purely linear nor fully nonlinear. It remains uncertain whether the digital mode introduces construct-irrelevant variance compared to the traditional paper-and-pencil format. There is also ambiguity about whether ePIRLS, designed for nonlinear reading with its graphical features, pop-ups, and teacher avatar, adds construct-irrelevant factors beyond reading ability. This research intends to address these uncertainties.

Sample

Publicly available data PIRLS were used for the present study (https://www.iea.nl/data-tools/repository). All the 16 countries and bench marking entities who took the PIRLS and ePIRLS 2016 were included in the analyses. In each country, only the students who had taken both tests was entered into the analyses. The data for Canada (*n*=8871), Denmark (*n*=2506), Georgia (*n*=5557), Ireland (*n*=2473), Israel (*n*=3798), Italy (*n*=3767), Norway (*n*=3610), Portugal (*n*=4558), Singapore (*n*=6320), Slovenia (*n*=4303), Sweden (*n*=3879), Taiwan (*n*=4326), United Arab Emirates (*n*=15566), Abu Dhabi (*n*=3980), Dubai (*n*=7471), and USA (*n*=4090) were analyzed. A pooled analysis with all the countries combined (*n*=85075) was also run.

Instruments

PIRLS 2016 employed multiple-matrix sampling and a rotated booklet design to ensure broad content coverage and overcome limited testing time. In multiple-matrix sampling, frequently used in large-scale assessments, a large pool of items is divided into different "blocks," and each participant is randomly assigned only a subset of these blocks. This reduces the testing load on individuals while still covering a wide range of content across the overall population. In PIRLS 2016, there were 12 blocks of reading passages and items, organized across 16 booklets. Each booklet included one literary and one informational passage with associated items. Some blocks were reused from previous PIRLS assessments, serving as “trend” items to track changes in reading achievement.

By grouping the items into blocks and randomly rotating these blocks within booklets, the testing time per student was limited to 80 minutes while allowing each item to be answered by a representative sample without overburdening any single student. To connect PIRLS booklets internally and with PIRLS Literacy (a simplified version for countries with lower reading proficiency), each literary and informational block appeared in three different booklets, paired with a different block each time. The ePIRLS assessment, grounded in the PIRLS 2016 Assessment Framework, emphasized evaluating students' online reading skills. The ePIRLS 2016 included five tasks (comprising 90 items), with each student completing two tasks, each lasting around 40 minutes. These tasks focused on topics in science and social studies. Two of the tasks are publicly available on the IEA website (https://pirls2016.org/epirls/take-the-epirls-assessment/)

PIRLS 2016 contained 175 items in paper-and-pencil mode, 90 of which targeted the *literary experience* purpose of reading, and 85 items measured the *acquire and use information* (AUI) purpose. The ePIRLS 2016 contained 90 items all measuring the *acquire and use information* reading purpose. To avoid the conflation of construct and testing mode, only the 85 AUI items of PIRLS were used in the analysis. The ePIRLS featured interactive tasks and a modern user interface allowing navigation between texts and answering questions. It contained simulated online projects with multiple websites, videos, and navigational features like links and pop-ups. Thus, all the items in the current study measured the same construct and only differed in testing mode. Items with partial credit were rescored to dichotomous items (correct = 1; incorrect or partially correct = 0) to simplify the analyses. A total of 175 AUI items (85 paper and 90 electronic) were analyzed.

Analyses

To examine the latent structure of the paper-based and digital tests, we estimate a series of Item Response Theory (IRT) models incorporating both general and specific factors. As a reference model, we first estimate a unidimensional 2PL model and compare it to a two-factor correlated multidimensional 2PL model and a bifactor 2PL model. The analyses are conducted separately for each country, followed by a pooled analysis combining data from all participating countries. The following models were estimated:

1. A unidimensional 2PL IRT model where a general reading comprehension factor was

modeled. Here the roles of mode and construct are ignored, and it is assumed that all the variances can be accounted for by a general reading ability factor (Figure 1, Panel A).

2. A correlated two-factor model where linear (paper-and-pencil) items loaded on one factor and nonlinear (digital) items loaded on another factor. This model assumes that the two types of reading are distinct but correlated constructs (Figure 1, Panel B).
3. A bifactor (S-1) model where a general reading factor for all items along with an orthogonal nested factor for the digitally administrated ePIRLS items was modeled (Figure 1, Panel C).

In a traditional bifactor model, there is a general factor (which represents the shared variance across all items in a test or questionnaire) and several specific factors (which represent variance that is unique to subsets of items). Each item loads onto the general factor and one specific factor. The (S-1) bifactor model (Eid et al., 2017) is a particular variant where one specific factor is designated as the general factor for the analysis. The other factors are constrained to be uncorrelated with the designated general factor but can correlated with each other. The model is particularly useful when there is a need to test the dominance of a specific construct (often a well-established dimension of interest) over others in the data, effectively setting one specific construct as the reference domain.

Eid et al. (2017) argue that in a traditional bifactor model, the general factor may not have a clear meaning or interpretation. They illustrate this with an example from medicine and state that, for instance, a G factor composed of height and weight could be modeled, but it may lack clear interpretative value. In contrast, the (S-1) bifactor model defines one domain as a reference, clarifying the G factor's meaning. In our example, using height as a reference, the G factor would reflect an individual's height corrected for measurement error, while the specific factor would capture variations in weight relative to height. This makes the specific factors meaningful as they show differences that remain after the influence of the general, reference-based factor is accounted for. This approach is particularly useful when one domain significantly influences others, such as fluid intelligence in intelligence research. In the current study, the ability to read linear texts on paper is designated as the reference domain which specifies the G factor while electronic nonlinear reading factor indicates whether examinees have a higher or lower nonlinear reading ability than expected given their linear reading scores. The *mirt* package (Chalmers, 2012) in R (R Core Team, 2023) was used to estimate the models.

To compare the three candidate models depicted in Figure 1, we use the Akaike Information Criterion (AIC, Akaike, 1974) and Bayesian Information Criterion (BIC, Schwarz, 1978). For model evaluation, we assess how much variance the different factors explain in the items. To this end, we compute the Explained Common Variance (ECV; Bentler, 2009), which represents the proportion of common variance explained by each factor. An ECV of 1 indicates that a factor accounts for all variance in the items it loads on. ECV for specific group factors refers to the proportion of common variance explained by a specific factor. Dueber and Toland (2023) suggest that for low sub-score reliabilities (ω=.60), ECV should be .45 to have added value. Bonifay et al. (2015) demonstrated that when ECV of the general factor is >.70, the relative bias in factor loadings is less than 10% which is trivial according to Reise et al. (2013). They also showed that when ECV>.80, the relative bias is less than 5%. Bias here is defined as the difference between the true loadings on the general factor in the bifactor model and the loadings in the unidimensional model when multidimensional data are forced to a unidimensional structure.

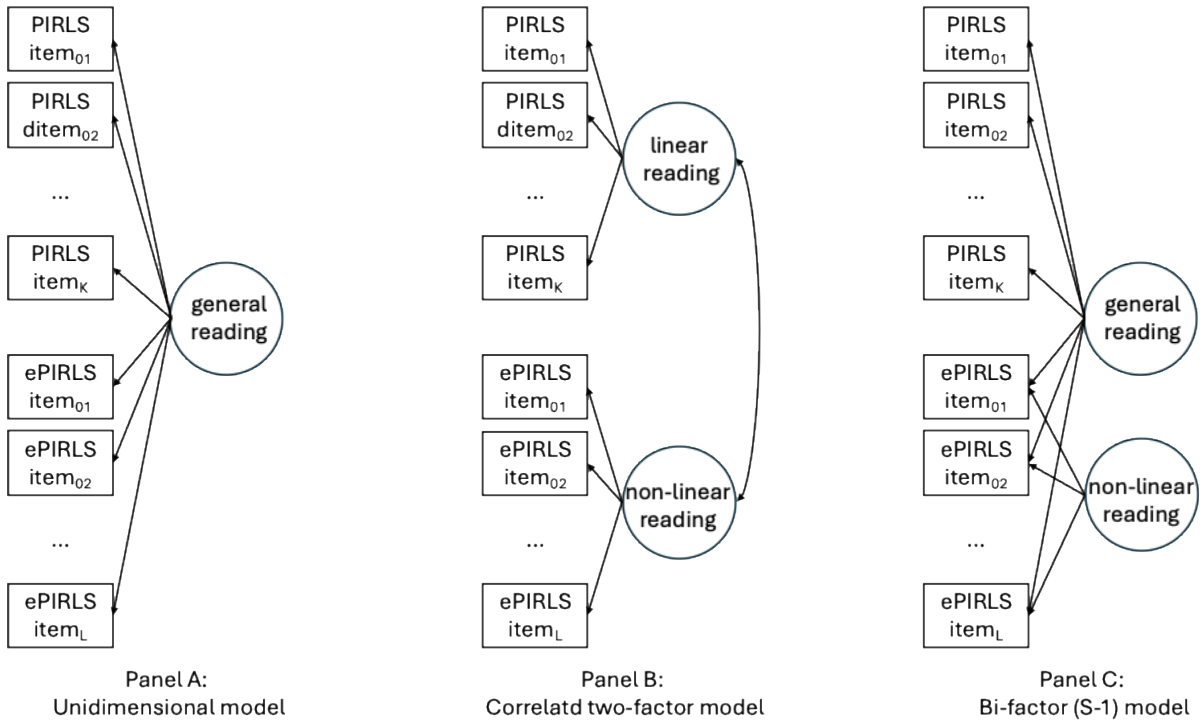

Figure 1.

Candidate Models

Results

Table 2 shows the information criteria (AIC and BIC) and factor reliabilities for the three estimated models across the countries. Information criteria show that the unidimensional model has the worst fit and the bifactor (S-1) model has the best fit. The correlated factor model has a better fit compared to the unidimensional model which is evidence that the two factors of linear and nonlinear reading can be better represented with two correlated factors rather than a single factor. Nevertheless, the better fit of the bifactor model compared to the correlated two-factor model indicates that perhaps the best representation of the data is a general reading factor along with a specific factor to account for nonlinear reading. This trend was uniformly observed across all the countries and in the pooled data. Specific factor reliabilities in the bifactor model ranged from .36 to .58 which indicate that the specific factor is not redundant. In fact, the reliable variance left in the specific factor after the influence of the general factor is conditioned out is not negligible and probably arises from differences in the ability to read nonlinear texts. Specific factor reliabilities show the proportion of variance in the items within a factor that is attributable to the factor itself, independent of the general factor. A high specific factor reliability means that the specific factor explains a considerable amount of variance in its items, beyond what is explained by the general factor (Baghaei, 2016).

Table 2. *Information criteria for the three models by country*

	Unidimensional model		Correlated two-factor model		Bifactor (S-1) model	
	AIC	BIC	AIC	BIC	AIC	BIC
Canada	13,544,300	13,547,957	13,519,330	13,522,998	13,504,515	13,509,113
Denmark	2,801,675	2,804,796	2,797,102	2,800,233	2,793,010	2,796,934
Georgia	2,163,657	2,166,665	2,151,732	2,154,749	2,149,639	2,153,416
Ireland	3,101,402	3,104,562	3,098,238	3,101,406	3,094,513	3,098,485
Israel	5302191	5,305,531	5,292,696	5,296,045	5,288,762	5,292,961
Italy	27,652,174	27,656,082	27,584,582	27,588,501	27,555,834	27,560,747
Norway	2,871,688	2,874,802	2,867,499	2,870,621	2,864,272	2,868,191
Portugal	4,883,162	4,886,467	4,875,825	4,879,139	4,868,837	4,872,992
Singapore	1,878,652	1,881,636	1,877,321	1,880,313	1,874,615	1,878,371
Slovenia	1,018,375	1,021,132	1,016,657	1,019,422	1,015,789	1,019,255
Sweden	4,678,900	4,681,615	4,673,868	4,676,592	4,667,740	4,671,316
Taiwan	10,013,606	10,017,174	10,003,449	10,007,028	9,992,276	9,996,762
UAE	3,789,820	3,793,051	3,778,659	3,781,900	3,776,918	3,780,981
Abu Dhabi	1,287,067	1,289,943	1,284,517	1,287,402	1,283,668	1,287,284
Dubai	1,130,711	1,133,503	1,127,460	1,130,260	1,126,978	1,130,489
USA	186,008,472	186,013,054	185,660,940	185,665,535	185,473,713	185,479,473
Pooled	277,737,204	277,741,920	276,988,499	276,993,228	276,868,789	276,874,717

Table 3 shows the Explained Common Variance for the bifactor model across the countries and in the pooled data. The overview suggests that the general factor in the bifactor model, across all the counties and in the pooled data, accounts for the highest proportion of the variance while the specific factor for the nonlinear items have smaller ECVs. The ECVs of the general factor are greater than .80 across all the countries and the ECVs of the specific factor are smaller than the criteria of .45 set by Dueber and Toland (2023). This is evidence for a strong general factor and weaker specific factors. In other words, the specific factor for the electronic test mode (or nonlinear reading) accounts for a small percentage of the entire variance in the item responses and the test is essentially unidimensional (Liaw et al., 2025). The highest specific factor ECV and reliability was observed for Georgia and the lowest for Singapore. This may be interpreted as the sharper distinctness between linear and nonlinear reading in Georgia and an overlap of the two types of reading in Singapore. The differences in countries may be explained by students' computer familiarity back in 2016.

Table 3. *Reliabilities, correlation, and explained common variance for the three models by country*

	Unidimensional model	Correlated two-factor model			Bifactor (S-1) model			
	R_G	R_N	R_L	Cor_{N*L}	R_G	R_N	ECV_G	ECV_N
Canada	.912	.902	.844	.883	.861	.425	0.894	.105
Denmark	.899	.889	.829	.885	.846	.439	0.891	.108
Georgia	.891	.875	.796	.767	.807	.584	0.814	.185
Ireland	.903	.894	.851	.919	.867	.395	0.928	.071
Israel	.922	.913	.870	.909	.881	.415	0.921	.078
Italy	.886	.874	.791	.822	.800	.507	0.843	.156
Norway	.889	.878	.824	.890	.858	.389	0.921	.078
Portugal	.892	.882	.817	.874	.835	.451	0.880	.110

Singapore	.908	.902	.867	.942	.894	.360	0.950	.050
Slovenia	.900	.889	.831	.882	.849	.429	0.896	.103
Sweden	.893	.887	.799	.876	.842	.392	0.881	.118
Taiwan	.894	.886	.882	.900	.850	.390	0.915	.084
UAE	.927	.917	.872	.894	.877	.462	0.905	.094
Abu Dhabi	.916	.906	.865	.910	.875	.403	0.920	.079
Dubai	.926	.916	.865	.881	.870	.474	0.894	.105
USA	.910	.900	.845	.887	.867	.431	0.908	.091
Pooled	.920	.910	.853	.857	.855	.504	0.870	.120

Note: R=reliability, ECV= explained common variance, Cor=correlation, subscripts G=general factor, subscripts N=nonlinear/electronic factor, subscripts L= linear/paper factor.

Table 4 shows the means of the Rasch model item difficulty parameters for the two sections of the test, i.e. electronic nonlinear items and paper-and-pencil linear items. It is important to reiterate that the test items differ in content. As the table shows, the nonlinear items are more difficult in most of the countries except for Sweden. A t-test was run to compare the two means across the countries. We recommend caution in interpreting the differences in difficulty between paper-based and computer-based tests, as variations in test content make it impossible to separate administration mode from content effects. However, interestingly, the mean differences vary considerably across countries: in all the countries (except for Sweden), computer-based items are more challenging than the paper-based ones. The mean differences were statistically significant for Georgia, Italy, Portugal, Slovenia, and Taiwan (p<.05).

Table 4. *Mean item difficulty parameters for linear and nonlinear items by country*

	M_N	M_L	M_N-M_L	t	p
Canada	-.274	-.618	.345	1.82	.07
Denmark	-.537	-.643	.105	.49	.61

Georgia	.644	.128	.515	2.58	.01
Ireland	-.643	-1.010	.366	1.75	.08
Israel	-.220	-.452	.231	1.17	.24
Italy	-.130	-.796	.665	3.02	.00
Norway	-.708	-.951	.242	1.11	.26
Portugal	.009	-.461	.470	1.99	.04
Singapore	-.958	-1.216	.257	1.31	.19
Slovenia	-.009	-.689	.680	3.17	.00
Sweden	-.564	-.488	-.076	-.29	.77
Taiwan	-.295	-1.061	.766	3.50	.00
UAE	.649	.437	.212	1.17	.74
Abu Dhabi	1.079	.901	.178	.96	.33
Dubai	-.098	-.395	.297	1.60	.11
USA	-.498	-.654	.156	.82	.41
Pooled	-.386	-.594	.208	1.13	.25

Note: M_N= mean of nonlinear/electronic items; M_L= mean of linear/paper items

Discussion

The emergence of new digital technologies has consequences for how we access, consume, and interact with information. Unlike traditional linear forms of reading, where content is presented in a fixed sequence, the digitization enables nonlinear reading through the use of hypertext—a system that links various texts, media, and resources.

This shift from linear to nonlinear reading has encouraged reading researchers to measure nonlinear reading ability alongside the traditional linear reading competence. PIRLS 2016 incorporated nonlinear reading into its assessment by launching the computer-based ePIRLS. The ePIRLS simulated an interactive web environment with several texts connected with hyperlinks, popups, and other graphical features to measure students' ability in reading nonlinear text.

Summary of key findings

A major concern with these new developments is the psychometric dimensionality of the reading tests composed of both linear and nonlinear texts and the equivalence of the constructs measured by the two types of reading measures. This study aimed to examine the psychometric distinctness and separability of linear versus nonlinear reading—that is, reading hypertext compared to reading on paper. To achieve this, we used data from the 2016 cycle of PIRLS and ePIRLS, which provided a large, cross-national sample of 4th grade students who completed both paper-based and digital assessments. This design allowed us to explore potential mode and construct differences across countries. In this study, the dimensionality was examined using bifactor IRT models. A unidimensional IRT model, a correlated two-factor model, and a bifactor (S-1) IRT model was estimated and compared in terms of overall fit, factor reliabilities, and the proportion of the variance they explain.

Findings showed that the bifactor (S-1) model which accounted for a general reading factor along with a specific factor for electronic test mode (nonlinear) had the best fit. This pattern was observed uniformly across all the 16 countries/benchmarking entities with very divergent school systems and learning experiences and in the pooled data. This suggests that it is plausible to assume that reading comprehension is composed of a general linear reading factor along with a specific ability for reading nonlinear texts. This specific factor could be an additional skill related to the ability to navigate and find information in online environments. However, variations in reliabilities and ECVs for the specific factor across countries indicate that the separability of the nonlinear reading factor varies across countries. Findings also showed that the correlated two-factor model had a better fit compared to the unidimensional model. This suggests that representing the data with two separate, but correlated factors is a better representation than a unidimensional structure. Nevertheless, this conclusion is scaled down when the factor correlations are examined. The correlations between paper-based linear and computer-based nonlinear reading ranged from .76 (Georgia) to .94 (Singapore) with a mean of .88 (SD=.03). This finding suggests that the distinction between linear and nonlinear varies across countries. A possible explanation is that computer familiarity might moderate the correlation between linear reading on paper and nonlinear reading on the computer (Cho et al., 2021). However, we could not test this hypothesis due to the absence of a computer familiarity measure. Future research should explore this further, particularly in educational systems where a digital gap exists. Additionally, differences across countries could be influenced by variations in keyboard layouts and typing difficulty, as some languages require different input methods. Bi et al. (2012) highlight the challenges posed by existing keyboard layouts for different language users, suggesting that keyboard design could affect typing performance and, potentially, digital reading tasks as well.

This study explored similarities and differences between paper-based linear and computer-based nonlinear reading. Our findings, based on high ECVs (> .80) for the general factor across all countries, indicates that general reading comprehension is fundamental for both reading linear texts on paper and nonlinear texts on a computer. While our analyses show fundamental similarities between linear and nonlinear reading, we also find notable differences. The nonlinear factor for nonlinear items have ECVs ranging from .07 to .18 with reliabilities ranging from .36 to .58, indicating that a portion of the variance in test performance is not solely explained by general reading ability but also requires specific skills. In the same vein, we found that computer-based nonlinear items are more difficult than paper-based linear items in some countries and these differences amount to .50-.70 logit in some countries. This could be an indication that the specific factorof nonlinear reading could still be a relevant factor that affects performance. While these differences exist in some countries but not in others is subject of future investigations.

The better fit of the bifactor (S-1) model, the emergence of a group factor with reliabilities as high as .50, and the differences in mean item difficulty parameters all suggest that multidimensionality in these data cannot be ignored at least in some countries. Reflecting on all the information and evidence would help us see the nuances and the complexity of a construct and avoid falling in the trap of simplicity and reductionism which potentially leads to oversimplified or incomplete theories. These findings are in line with Mahlow et al. (2020) who compared the dimensionality of what they referred to as STC (single text comprehension) and MDC (multiple documents comprehension). Their study showed that while single text reading and multiple texts reading are separable, they are highly correlated with a latent correlation coefficient of .84. They concluded that STC and MDC are distinct constructs, suggesting that single-text and multiple-text reading scenarios place different demands on readers. However, the strong correlation between these constructs suggests that the core skills underlying the two types of readings are the same. The better fit of the bifactor (S-1) model in this study supports Mahlow et al.'s conclusion.

Limitations

It is important to exercise caution when interpreting the findings of this study, as they must be considered within the constraints imposed by both test mode and trait factor. As previously outlined, nonlinear reading items were administered via computer, whereas linear reading items were presented in a traditional paper-and-pencil format. Consequently, the factor identified as the "nonlinear reading factor" may, in fact, reflect a test mode effect linked to participants' familiarity with computers and computer-based assessments rather than an inherent characteristic of nonlinear reading itself. In other words, the observed multidimensionality may stem from differences in test administration modes (paper vs. computer) rather than genuine distinctions between the cognitive processes involved in linear and nonlinear reading. Given the structure of the data, it is not possible to definitively determine the precise nature of this factor. Future research employing a consistent test mode while varying test constructs (linear vs. nonlinear) is necessary to establish whether nonlinear reading constitutes a distinct construct from traditional linear reading.

Implications of the findings

The findings of the current study have several theoretical and practical implications. It showed that linear and nonlinear reading tests measure an essentially unidimensional construct but the nonlinear assessment introduces a notable secondary factor at least in some countries. The findings of the present study promote more refined and accurate reading theories that reflect the complexity of the construct, which can inform both research and practice. It also highlights that learners may need targeted interventions for reading nonlinear texts and helping educators design more focused instructional strategies. It also implies that reading assessments should measure the ability to read nonlinear texts. That is, a more precise test development that can offer richer, more diagnostic information about test-takers' abilities is required.

Our findings provide important insights into the role of general reading competence in PIRLS assessments. We observe that general reading ability is the dominant factor in measuring reading comprehension in both paper and digital PIRLS. This suggests that for many research questions, such as those investigating the determinants of reading proficiency, it may not be critical whether PIRLS assessments are administered on paper or digitally. However, we caution against an uncritical merging of concepts when interpreting results across formats.

Certain covariates of reading competence may interact differently with paper-based and digital assessments. A possible example is familiarity with digital texts. Research suggests that using digital devices for schoolwork, rather than leisure, is linked to better reading performance in both formats (Grammatikopoulou et al., 2025). Moreover, children who engage with digital devices for educational purposes may perform better on digital reading than paper-based reading. This suggests that while general reading ability remains the primary factor, digital reading assessments may introduce additional influences that should not be overlooked.

It is also important to acknowledge an alternative interpretation of our findings. While we argue that nonlinear digital reading introduces additional skills related to navigation and synthesis, it could equally be the case that linear paper-based reading places greater demands on sustained memory, concentration, and the integration of longer continuous texts. In this sense, the distinction we observe might not solely arise from digital reading requiring "extra" skills, but from both modes tapping into partly different sets of cognitive resources. Nevertheless, the empirical outcome remains the same: despite the dominance of a general reading factor, linear and nonlinear assessments each capture unique aspects of comprehension that warrant careful consideration in research and educational practice.

Finally, we advise caution when comparing reading scores over time, particularly between older paper-based PIRLS cycles and newer digitally administered PIRLS cycles. Even small differences in performance across cycles could be partially attributed to text linearity rather than genuine shifts in reading ability. In PIRLS 2016, the computer-based ePIRLS test components were administered for the first time alongside the traditional paper-based assessments. However, the results were reported on separate scales, meaning that students received one score based on the paper-based PIRLS test and another based on the digital ePIRLS test. At the country level, significant differences emerged in students' performances between the two assessments, indicating that they measured different constructs. Despite this, PIRLS 2021 combined the linear test items from PIRLS with the non-linear tasks from ePIRLS into a single test score. However, in 2021, the linear PIRLS items were presented to students on computer screens, ensuring that the mode of presentation was consistent across all test components. Future research should explore these potential text effects in greater depth to ensure that longitudinal comparisons remain valid.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19*, 716–723.

Aparicio, X., Belaïd, S., Baccino, T., & Megalakaki, O. (2022). Reading-comprehension performances of expository and narrative texts on Interactive-Whiteboards and Paper: evidence from 5th grade children. *Educational Technology Research and Development*, *70*(4), 1151–1168. https://doi.org/10.1007/s11423-022-10121-z

Baghaei, P. (2016). Modeling multidimensionality in foreign language comprehension tests: An Iranian example. In Aryadoust, V., & Fox, J. (Eds.), *Trends in language assessment research and practice: The view from the Middle East and the Pacific Rim* (pp.47–66). Cambridge Scholars.

Bi, X., Smith, B. A., & Zhai, S. (2012). Multilingual touchscreen keyboard design and optimization. *Human–Computer Interaction*, *27*(4), 352-382.

Billington, J., Dowrick, C., Hamer, A., Robinson, J., and Williams, C. (2010). An Investigation Into the Therapeutic Benefits of Reading in Relation to Depression and Well-Being. Liverpool Health Inequalities Research Institute.

Bonifay, W. E., Reise, S. P., Scheines, R., & Meijer, R. R. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal, 22*(4), 504–516. https://doi.org/10.1080/10705511.2014.93859

Cartelli, A. (Ed.). (2012). *Current trends and future practices for digital literacy and competence*. IGI Global.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06.

Chen, S.-F. (2017). Modeling the influences of upper-elementary school students' digital reading literacy, socioeconomic factors, and self-regulated learning strategies. *Research in Science & Technological Education*, *35*(3), 330–348. https://doi.org/10.1080/02635143.2017.1314958

Cho, B.-Y., Hwang, H., & Jang, B. G. (2021). Predicting fourth grade digital Reading comprehension: A secondary data analysis of (e)PIRLS 2016. *International Journal of Educational Research*, *105*, 101696. https://doi.org/10.1016/j.ijer.2020.101696

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading*, *42*(2), 288–325. https://doi.org/10.1111/1467-9817.12269

Coiro, J. (2003). Exploring Literacy on the Internet: Reading Comprehension on the Internet: Expanding Our Understanding of Reading Comprehension to Encompass New Literacies. *The Reading Teacher*, *56*(5), 458–464.

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

Cunningham, A. E., & Stanovich, K. E. (2007). What Reading Does for the Mind. *Estudios públicos*, *108*, 207–228.

Daane M. C., Campbell J. R., Grigg W. S., Goodman M. J., Oranje A. (2005). *The nation's report card: fourth-grade students reading aloud. NAEP 2002 special study of oral reading. U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics*. Washington, DC: U.S. Government Printing Office.

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. https://doi.org/10.1016/j.edurev.2018.09.003

Dueber, D. M., & Toland, M. D. (2023). A bifactor approach to subscore assessment. *Psychological Methods, 28*(1), 222–241. https://doi.org/10.1037/met0000459

Egert, F., Cordes, A.-K., & Hartig, F. (2022). Can e-books foster child language? Meta-analysis on the effectiveness of e-book interventions in early childhood education and care. *Educational Research Review*, *37*, 100472. https://doi.org/10.1016/j.edurev.2022.100472

Eid, M., Geiser, C., Koch, T., & Heene, M. (2017). Anomalous results in G-factor models: Explanations and alternatives. *Psychological Methods, 22*(3), 541–562. https://doi.org/10.1037/met0000083

Empson, R. (2014). Education giant Blackboard buys MyEdu to help refresh its brand and reanimate its user experience. *TechCrunch*. Retrieved October 01, 2024 from https://techcrunch.com/2014/01/16/education-giant-blackboard-buys-myedu-to-help-refresh-its-brand-and-reanimate-its-user-experience/

Furenes, M. I., Kucirkova, N., & Bus, A. G. (2021). A Comparison of Children's Reading on Paper Versus Screen: A Meta-Analysis. *Review of Educational Research*, *91*(4), 483–517. https://doi.org/10.3102/0034654321998074

Grammatikopoulou, E., Johansson, S., & Rosén, M. (2025). Paper-based and digital reading in 14 countries: exploring cross-country variation in mode effects. *Educational Review*, 1–19. https://doi.org/10.1080/00131911.2025.2452236

Hahnel, C., Goldhammer, F., Naumann, J., & Kröhne, 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. https://doi.org/10.1016/j.chb.2015.09.042

Hsiao, K.-L., & Chen, C.-C. (2015). How do we inspire children to learn with e-readers? *Library Hi Tech*, *33*(4), 584–596.

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

Kaman, S., & Ertem, I. S. (2018). The effect of digital texts on primary students' comprehension, fluency, and attitude. *Eurasian Journal of Educational Research*, *2018*(76), 147–164.

Kidd, D. C., & Castano, E. (2013). Reading Literary Fiction Improves Theory of Mind. *Science (American Association for the Advancement of Science)*, *342*(6156), 377–380. https://doi.org/10.1126/science.1239918

KMK (2016) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland (2016). *Strategie der Kultusministerkonferenz "Bildung in der digitalen Welt"* (Resolution of the KMK from December 8, 2016, as amended on December 7, 2017.).

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

Kress, G. (2003). *Literacy in the new media age*. Routledge. https://doi.org/10.4324/9780203299234

Leu, Jr., D. J., Kinzer, C. K., Coiro, J., & Cammack, D. W. (2004). Toward a Theory of New Literacies Emerging From the Internet and Other Information and Communication Technologies. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical Models and*

Processes of Reading (5th ed., Vol. 1, pp. 1570–1613). International Reading Association. https://doi.org/10.1598/0872075028.54

Liaw, Y.-L., Baghaei, P., Strietholt, R., Meinck, S., & Strello, A. (2025). Environmental knowledge: Conceptualization and measurement. In Isac, M. M., Sandoval-Hernández, A., & Sass, W. (Eds.), *Knowledge and willingness to act pro-environmentally* (pp. 31–56). Springer. https://doi.org/10.1007/978-3-031-76033-4_4

Madden, M., Lenhart, A., Duggan, M., Cortesi, S., & Gasser, U. (2013). *Teens and technology 2013*. Washington, DC: Pew Research Center's Internet & American Life Project. Retrieved September 25, 2024 from https://www.pewresearch.org/internet/2013/03/13/teens-and-technology-2013/.

Mahlow, N., Hahnel, C., Kroehne, U., Artelt, C., Goldhammer, F., & Schoor, C. (2020). More than (single) text comprehension? – On university students' understanding of multiple documents. *Frontiers in Psychology*, *11*, 562450. https://doi.org/10.3389/fpsyg.2020.562450

Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 Assessment Framework* (2nd ed.). Retrieved from Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2016/framework.html

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. Retrieved from Boston College, TIMSS & PIRLS International Study Center, website, from http://timssandpirls.bc.edu/pirls2016/international-results/

Mullis, I. V. S., & Prendergast, C. O. (2017). Developing the PIRLS 2016 achievement items. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in PIRLS 2016* (pp. 1.1–1.29). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-1.html

Murphy, P. K., Long, J. F., Holleran, T. A., & Esterly, E. (2003). Persuasion online or on paper: A new take on an old issue. *Learning and Instruction*, *13*(5), 511–532. https://doi.org/10.1016/S0959-4752(02)00041-5

National Institute of Child Health and Human Development (NICHD) (2000). *Report of the National Reading Panel. Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*. NIH Publication No. 00-4769. Washington, DC: U.S. Government Printing Office.

Organization for Economic Co-operation and Development (OECD). (2011). *PISA 2009 Results: Students online*. OECD Publishing. Retrieved from http://www.oecd-ilibrary.org/education/pisa-2009-results-students-on-line_9789264112995-en.

Quinn, H. (2014). *Bifactor models, explained common variance (ECV), and the usefulness of scores from unidimensional item response theory analyses*. Master's thesis. University of North Carolina at Chapel Hill. https://doi.org/10.17615/t6ff-a088

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement, 73*, 5–26. https://doi.org/10.1177/0013164412449831

Ronconi, A., Veronesi, V., Mason, L., Manzione, L., Florit, E., Anmarkrud, Ø., & Bråten, I. (2022). Effects of reading medium on the processing, comprehension, and calibration of adolescent readers. *Computers and Education*, *185*, 104520. https://doi.org/10.1016/j.compedu.2022.104520

Rosén, M., & Gustafsson, J.-E. (2016). Is computer availability at home causally related to reading achievement in grade 4? A longitudinal difference in differences approach to IEA data from 1991 to 2006. *Large-Scale Assessments in Education*, *4*(1), 5. https://doi.org/10.1186/s40536-016-0020-8

Savva, M., Higgins, S., & Beckmann, N. (2022). Meta-analysis examining the effects of electronic storybooks on language and literacy outcomes for children in grades Pre-K to grade 2. *Journal of Computer Assisted Learning*, *38*(2), 526–564. https://doi.org/10.1111/jcal.12623

Schwabe, A., Lind, F., Kosch, L., & Boomgaarden, H. G. (2022). No Negative Effects of Reading on Screen on Comprehension of Narrative Texts Compared to Print: A Meta-analysis. *Media Psychology*, *25*(6), 779–796. https://doi.org/10.1080/15213269.2022.2070216

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics* 6, 461–464.

Shneiderman, B. (1989). Reflections on authoring, editing and managing hypertext. In E. Barrett (Ed.), *The society of text: hypertext, hypermedia, and the social construction of information* (115–131). Cambridge, MA: MIT Press.

Shneiderman, B. "Reflections on authoring, editing, and managing hypertext, The society of text: hypertext, hypermedia, and the social construction of information." (1989).

Zuckerman, B., & Augustyn, M. (2011). Books and Reading: Evidence-Based Standard of Care Whose Time Has Come. *Academic Pediatrics*, *11*(1), 11–17. https://doi.org/10.1016/j.acap.2010.09.007

Zumbach, J., & Mohraz, M. (2008). Cognitive load in hypermedia reading comprehension: Influence of text type and linearity. *Computers in Human Behavior*, *24*(3), 875–887. https://doi.org/10.1016/j.chb.2007.02.015