Linear Vs. Nonlinear Reading: Examining the Dimensionality of PIRLS 2021

Purya Baghaei (corresponding author)

Research and Analysis Unit, IEA Hamburg, Hamburg, Germany

Email: purya.baghaei@iea-hamburg.de

Rolf Strietholt

Research and Analysis Unit, IEA Hamburg, Hamburg, Germany

Stefan Johansson

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

Andrés Christianssen

Research and Analysis Unit, IEA Hamburg, Hamburg, Germany

Abstract

In this study, the psychometric separability of linear sequential reading and nonlinear hypertext reading was examined using digitalPIRLS and ePIRLS 2021 data. Three 2-PL item response theory models were fitted to the data. A unidimensional model where all the items loaded on a single reading dimension, a correlated 2-factor model with linear and nonlinear reading as two separate but correlated factors, and a bifactor (S-1) model with nonlinear items loading on a specific factor and linear reading items as the reference facet were fitted to the data. Information criteria showed that the bifactor (S-1) model has the best fit across all the countries, suggesting that a dominant general factor of reading along with a specific nonlinear reading factor account for the variation in students' responses to the test items. An examination of the ECVs showed that the general reading factor explains approximately 90 percent of the item variability, while the specific nonlinear factor accounts for about 10 percent. The reliability coefficients of the specific factors were low. We conclude that although linear and nonlinear reading assessments measure similar constructs, nonlinear reading introduces additional variance that is not captured by linear reading assessments. The implications of this finding, as well as potential reasons for the discrepancy, are discussed.

Keywords: linear reading, nonlinear reading, single text comprehension, multiple documents comprehension, dimensionality

Introduction

The Progress in International Reading Literacy Study (PIRLS) is an international assessment conducted by the International Association for the Evaluation of Educational Achievement (IEA) that measures reading literacy among fourth-grade students. Results from PIRLS 2021 are widely anticipated by educators, researchers, and policymakers globally, particularly due to the importance placed on trends in countries' performance over time. Policymakers often closely monitor these trends to assess whether their country's educational performance has improved or declined relative to other nations. Many countries have previously received significant attention due to fluctuations in PIRLS reading scores over successive cycles (Mullis et al., 2017).

However, interpreting trends in PIRLS results requires caution, as methodological differences between assessment cycles can influence the outcomes. Changes in sampling procedures, student engagement, item formats, or administration contexts might lead to apparent shifts in national performance. These methodological factors should be carefully considered before drawing definitive conclusions about a country's educational performance based solely on PIRLS trends.

PIRLS 2021 marked a significant transition as it was the first PIRLS cycle that allowed countries to participate digitally. While traditional paper-based testing was still an option, most countries adopted digitalPIRLS, an innovative, fully digital assessment. This shift reflects a growing trend in educational testing towards digital formats. A digital PIRLS assessment increases operational efficiency and provides a visually attractive environment that motivates students (Mullis & Martin, 2019). By transitioning to a digital format, digitalPIRLS enables researchers and educators to gather more detailed data on how students interact with text in an online environment.

The ePIRLS, first introduced in 2016, is an extension of PIRLS that assesses online reading comprehension skills. Unlike digitalPIRLS, which is a full replacement for the traditional paper-based test, ePIRLS is a separate component that specifically evaluates how well students can navigate and comprehend digital informational texts. Key features of ePIRLS include simulated internet-like environments, where students read online articles, interpret multimedia elements, and answer questions. Tasks that reflect real-world digital literacy skills, such as evaluating hyperlinks, understanding embedded media, filtering pop-ups, and synthesizing online content are included in ePIRLS. While both digitalPIRLS and ePIRLS leverage technology to enhance reading assessment, digitalPIRLS serves as a direct modernization of PIRLS, whereas ePIRLS provides additional insights into students' ability to navigate and comprehend digital content. Table 1 below highlights the key differences and similarities between digitalPIRLS and ePIRLS.

Table 1

Comparison of digitalPIRLS and ePIRLS

	digitalPIRLS	ePIRLS
Launch year	2021	2016
Format	Fully digital version of PIRLS replacing the paper-based PIRLS	A supplemental online reading assessment
Relation to PIRLS	Main assessment	A separate but complementary assessment
Text types	Literary and acquire and use information	Only acquire and use information
Structure	Linear	Nonlinear

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, 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.

In previous assessments linear reading items have been administered on paper while nonlinear items have been delivered digitally, and it is still up to date unclear whether this structure is a factor reflecting an actual cognitive distinction between reading styles or merely an effect of different testing modes. To clarify this issue, the present study examines data from digitalPIRLS 2021 and ePIRLS 2021, where both linear and nonlinear reading tasks were administered in a computer-based format. By keeping the test mode constant, we aim to determine whether a nonlinear reading ability factor can still be identified. If such a factor emerges under these conditions, it would provide stronger evidence that linear and nonlinear reading require distinct cognitive abilities rather than being a byproduct of mode effects.

Linear and non-linear reading

Reading comprehension is a core academic skill essential for knowledge acquisition, critical thinking, and lifelong learning (National Institute of Child Health and Human Development,

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

2000). Traditionally, reading has been associated with a linear format, where readers follow a fixed sequence from beginning to end. Linear reading follows a sequential structure, where readers progress step by step through the text. This is the typical format for printed books, newspapers, and traditional academic materials, where comprehension relies on memory, contextual understanding, and deep processing of content in a fixed order (Grammatikopoulou et al., 2025; Jeong & Gweon, 2021). However, the increasing digitization of education and everyday life has led to the rise of nonlinear reading, which allows users to navigate freely through hyperlinked and multimedia content. Nonlinear reading is nonsequential and necessities moving from one text to another, filtering out irrelevant texts and integrating information from several resources to build a coherent picture.

Some researchers argue that nonlinear reading demands additional cognitive load due to the need for navigation strategies, self-regulation, and integration of multiple sources of information (Zumbach & Mohraz, 2007). Unlike linear texts, which guide readers through a fixed structure, hypertext environments require active decision-making regarding information pathways, increasing cognitive demand (Shneiderman, 1989). Moreover, nonlinear reading often includes multimodal elements such as images, videos, and hyperlinks, which may either enhance comprehension or create distractions, depending on the reader's ability to filter and synthesize relevant content (Kress, 2003; Coiro, 2011).

Other researchers have referred to the same issue as *single text comprehension* (STC) and *multiple document comprehension* (MDC, Bråten & Strømsø, 2010a, 2010b; Mahlow et al. 2020). They argue that the nature of reading and the demands it places on individuals have evolved due to digital advancements. With the vast accessibility of online information, individuals must integrate and assess content from multiple sources. This ability, known as MDC, involves understanding, organizing, and synthesizing information from various texts on the same topic. MDC and STC are conceptually aligned with nonlinear and linear reading,

respectively. When reading multiple documents, readers encounter multiple sources with overlapping, complementary, or contradictory content, requiring them to discern differences and create a coherent understanding. Research shows that many students struggle with comprehending multiple documents (Britt & Rouet, 2012), but interventions can enhance this skill, and it tends to improve as students progress through their studies (Schoor et al., 2020; von der Mühlen et al., 2016).

Researchers argue that reding multiple texts is more complex than reading a single text, as it requires readers to compare, integrate, and evaluate information across multiple sources (Britt & Rouet, 2012; Goldman, 2004). Research suggests that strategies used in reading a single text, such as identifying important information and monitoring comprehension, also apply to reading multiple texts but in more extensive ways (Afflerbach & Cho, 2009). MDC differs from STC since multiple documents can either complement or contradict one another, requiring readers to reconcile conflicting information (Stadtler & Bromme, 2014). Furthermore, in MDC, readers get involved in text evaluation based on source credibility, considering factors like author expertise, publication context, and intended audience. Source evaluation becomes particularly important when texts present conflicting information (Wineburg, 1991).

In the past, reading comprehension models focused primarily on STC, which involves deriving meaning from a single source (Kintsch, 1988; Trabasso et al., 1989; Graesser et al., 1994; Zwaan et al., 1995). However, the documents model framework introduced in the late 1990s (Perfetti et al., 1999) addressed the additional complexities of MDC by incorporating two key components: the integrated situation model (which combines information from multiple texts) and the intertext model (which includes source-related information such as author and publication details). Several subsequent models have built upon this framework to examine different aspects of document comprehension.

Despite the theoretical advance, the precise nature of MDC and its relationship with STC remains unclear. Given the shift in digitization and the availability of online resources, an important question in reading research is whether linear (single text) reading and nonlinear (multiple documents) reading constitute distinct cognitive constructs or whether there is a single reading ability that accounts for all types of reading.

This study contributes to the ongoing discussion of reading comprehension assessment in the digital era, with implications for both educational practice and large-scale literacy assessments. If nonlinear reading represents a unique construct, reading instruction may need to adapt to help students develop navigation, integration, and evaluation skills specific to digital texts.

Method

Data and Material

Publicly available data from TIMSS and PIRLS study center were used for the present study. All the 26 countries who took the digital PIRLS and ePIRLS 2021 were included in the analyses. The data for Belgium (Flemish) (*n*=5114), Chinese Taipei (*n*=5555), Czech Republic (*n*=6621), Germany (*n*=4611), Denmark (*n*=4821), Spain (*n*=8551), Finland (*n*=7018), Croatia (*n*=3937), Hungary (*n*=5312), Israel (*n*=4890), Italy (*n*=5440), Kazakhstan (*n*=7023), Lithuania (*n*=4623), Malta (*n*=3030), Norway (*n*=5382), New Zealand (*n*=5557), Portugal (*n*=6111), Qatar (*n*=5258), Russian Federation (*n*=5217), Saudi Arabia (*n*=4778), Singapore (*n*=6719), Slovak Republic (*n*=4841), Slovenia (*n*=5110), Sweden (*n*=5175), and the United Arab Emirates (*n*=27448), were analyzed. A pooled analysis with all the countries combined was also run.

In PIRLS 2021, the reading tasks or passages were presented in a rotated booklet design with each booklet containing two tasks. In countries participating in digitalPIRLS,

students were also given ePIRLS, which included either two ePIRLS tasks or one digitalPIRLS informational passage followed by an ePIRLS task. The ePIRLS assessment for 2021 included five tasks delivered via computer or tablet. Students involved in digitalPIRLS received one of three types of booklets: a standard booklet with two digitalPIRLS passages, an ePIRLS booklet containing two ePIRLS tasks, or a hybrid booklet featuring one digitalPIRLS informational passage followed by an ePIRLS task. The PIRLS assessment was conducted over two 40-minute sessions, each dedicated to a passage or task, with a brief break in between, and concluded with a 30-minute session for the student questionnaires.

DigitalPIRLS 2021 contained 142 literary items based on 9 passages and 226 informational items based on 14 passages (9 digitalPIRLS passages and 5 ePIRLS passages). To avoid the conflation of method and construct, we excluded all the literary experience items and focused only on the acquire and use information items. Since the countries who took the paper PIRLS did not take the ePIRLS, they were not included in Study 2. The informational section of the assessment analyzed in this study contained 91 ePIRLS nonlinear items and 135 digitalPIRLS linear items.

Analyses

Following the design in Authors et al. (in review), a similar set of analyses were run to examine the dimensionality of the combination of digitalPIRLS and ePIRLS items. The analyses were run separately within each country. A final pooled analysis with all the countries combined was also run. The following models were estimated using the *mirt* package (Chalmers, 2012) in R (R Core Team, 2023).

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

assumed (Figure 1, Panel A). This model posits that both digital and electronic items (linear and nonlinear) can be explained with a single dimension of general reading comprehension.

2. A correlated two-factor model where linear (digital) items loaded on one factor and nonlinear (electronic) 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 includes a general factor alongside an orthogonal specific factor. In this model, digital linear items load exclusively on the general factor, while nonlinear ePIRLS items load on both the general and specific factors (Figure 1, Panel C). This structure assumes that general reading comprehension underlies both linear and nonlinear reading but that reading nonlinear texts requires additional distinct skills.

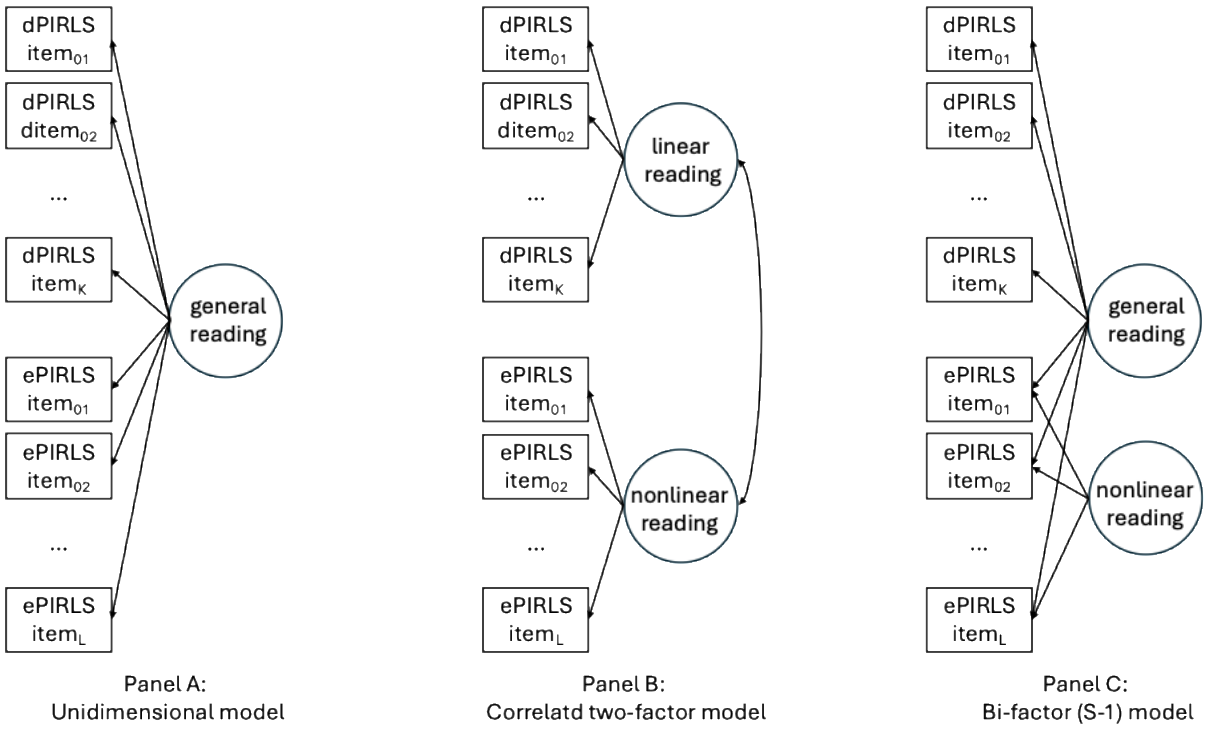

Figure 1.
Candidate Models

Results

Table 2 shows the information criteria for the relevant models across the countries. Information criteria show that the bifactor (S-1) model has the best fit across all the countries. There are two exceptions though. In Malta and in the pooled data, the Bayesian Information Criterion (BIC) indicates that the 2-factor model has the best fit. Nevertheless, the Akaike Information Criterion (AIC) suggest that the bifactor (S-1) has the best fit in all the countries and in the pooled data.

Table 2.

Information Criteria for the Three Models by Country

	Unidimensional Model		Correlated 2-Factor Model		Bifactor (S-1) Model	
Country	AIC	BIC	AIC	BIC	AIC	BIC
Pooled	271,261	274,560	271,148	274,455	270,977	275,250
ARE	1,749,156	1,753,300	1,748,764	1,752,918	1,746,019	1,751,386
BFL	1,603,213	1,607,302	1,602,624	1,606,723	1,597,014	1,602,310
CZE	2,223,437	2,227,682	2,222,183	2,226,437	2,213,169	2,218,665
DEU	14,338,938	14,344,029	14,329,696	14,334,798	14,280,587	14,287,180
DNK	1,212,728	1,216,713	1,212,113	1,216,107	1,207,027	1,212,188
ESP	9,641,183	9,646,080	9,640,184	9,645,092	9,604,663	9,611,005
FIN	1,246,075	1,250,048	1,245,269	1,249,251	1,241,422	1,246,578
HRV	753,684	757,448	753,362	757,134	748,423	753,298
HUN	1,715,166	1,719,293	1,714,085	1,718,221	1,708,802	1,714,147
ISR	2,459,109	2,463,409	2,458,555	2,462,865	2,450,337	2,455,906
ITA	10,352,057	10,356,984	10,351,053	10,355,991	10,313,857	10,320,238
KAZ	7,222,192	7,226,934	7,219,379	7,224,132	7,196,011	7,202,152
LTU	583,920.6	587,575.1	583,671.6	587,334.4	581,310.5	586,043.3
MLT	96,151	98,987	96,121	98,963	95,967	99,640
NOR	1,283,622	1,287,630	1,282,805	1,286,821	1,278,852	1,284,042
NZL	1,306,288	1,310,304	1,305,727	1,309,752	1,302,370	1,307,571
PRT	1,986,837	1,991,023	1,985,875	1,990,071	1,978,665	1,984,086
QAT	497,309	500,894	497,156	500,748	495,481	500,123
RUS	35,209,469	35,214,925	35,194,633	35,200,100	35,083,676	35,090,749
SAU	6,193,227	6,197,936	6,189,639	6,194,358	6,164,611	6,170,709
SGP	749,744	753,526	749,256	753,046	747,697	752,595
SVK	1,126,287	1,130,207	1,126,059	1,129,988	1,117,591	1,122,664
SVN	452,426	455,964	452,199	455,745	450,418	455,000

SWE	2,491,711	2,496,007	2,489,328	2,493,634	2,480,118	2,485,686
TWN	3,690,755	3,695,217	3,687,733	3,692,205	3,674,884	3,680,668

Table 3 shows the explained common variances (ECV) and factor reliabilities across countries. Examination of the ECVs shows that the ECVs for the general factor range from .79–.91. ECV values greater than .70 are high and indicate a dominant general factor (Rodriguez et al., 2016). The ECVs for the specific factors range from .05 to .21, indicating that while the explained variance is small, it is not negligible and suggests some differentiation in the construct (Liaw et al., 2025). The reliability of the specific factor across countries and the pooled data ranged from .16 to .29 which are very small. This is different from Authors et al. (in review) (Study 1) where the reliabilities of the specific factors ranged from .38 to .58. The latent correlations between the two factors of linear and nonlinear reading in the multidimensional model range from .79 to .93.

Table 3.

Reliabilities, Factor Correlations, and Explained Common Variances for the Three Models by Country

	Unidimensional Model	Correlated 2-Factor Model			Bifactor (S-1) Model			
Country	R_G	R_L	R_N	Cor_{N*L}	R_G	R_N	ECV_G	ECV_N
Pooled	0.82	0.77	0.69	0.87	0.79	0.19	0.90	0.10
ARE	0.86	0.83	0.79	0.94	0.85	0.16	0.95	0.05
BFL	0.77	0.71	0.63	0.85	0.75	0.26	0.85	0.15
CZE	0.78	0.72	0.63	0.84	0.72	0.29	0.79	0.21
DEU	0.80	0.74	0.65	0.85	0.73	0.29	0.81	0.19
DNK	0.78	0.72	0.64	0.85	0.75	0.25	0.86	0.14
ESP	0.77	0.74	0.70	0.93	0.76	0.23	0.89	0.11
FIN	0.80	0.74	0.65	0.85	0.75	0.27	0.88	0.12
HRV	0.77	0.72	0.64	0.87	0.75	0.26	0.85	0.15
HUN	0.81	0.75	0.66	0.85	0.75	0.27	0.83	0.17
ISR	0.80	0.77	0.72	0.92	0.79	0.24	0.91	0.09
ITA	0.76	0.73	0.68	0.92	0.75	0.22	0.91	0.09
KAZ	0.79	0.74	0.66	0.86	0.77	0.24	0.88	0.12

LTU	0.77	0.72	0.64	0.86	0.76	0.26	0.91	0.09
MLT	0.81	0.77	0.70	0.89	0.79	0.28	0.88	0.12
NOR	0.79	0.73	0.64	0.84	0.75	0.25	0.85	0.15
NZL	0.82	0.78	0.71	0.89	0.79	0.24	0.88	0.12
PRT	0.78	0.72	0.63	0.84	0.77	0.23	0.90	0.10
QAT	0.82	0.78	0.72	0.91	0.80	0.22	0.91	0.09
RUS	0.78	0.72	0.65	0.87	0.72	0.27	0.85	0.15
SAU	0.79	0.74	0.65	0.84	0.76	0.22	0.83	0.17
SGP	0.81	0.76	0.69	0.88	0.76	0.25	0.88	0.12
SVK	0.79	0.75	0.69	0.89	0.79	0.24	0.90	0.10
SVN	0.78	0.72	0.63	0.84	0.76	0.24	0.88	0.12
SWE	0.80	0.73	0.63	0.82	0.74	0.29	0.79	0.21
TWN	0.77	0.69	0.58	0.79	0.71	0.28	0.81	0.19

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

Table 4 shows the means of item difficulty parameters (based on the Rasch model) for the two sets of items across countries and the pooled data. Assuming that the two sets of items have randomly equivalent difficulty except for the reading type, it is expected to observe some differences between the means. The mean differences across the countries range from .10 (United Arab Emirates) to .54 (Singapore) which are statistically significant for six countries (p<.05). This suggests that the cognitive processes and demands behind the two sets of items are comparable.

Table 4.

Mean Item Difficulty Parameters for Linear and Nonlinear Items by Country

Country	$Mean_N$	$Mean_L$	Mean Difference	t	p
Pooled	-0.20	-0.48	0.28	-1.84	0.07
ARE	0.17	0.08	0.10	-0.68	0.50
BFL	0.06	-0.17	0.23	-1.34	0.18
CZE	-0.43	-0.67	0.23	-1.42	0.16
DEU	-0.04	-0.46	0.42	-2.37	0.02
DNK	-0.32	-0.63	0.31	-1.80	0.07
ESP	-0.14	-0.40	0.26	-1.50	0.14
FIN	-0.57	-0.83	0.26	-1.44	0.15
HRV	-0.56	-0.88	0.32	-1.92	0.06
HUN	-0.41	-0.62	0.21	-1.26	0.21

ISR	0.06	-0.24	0.30	-1.90	0.06
ITA	-0.36	-0.65	0.29	-1.77	0.08
KAZ	0.18	0.01	0.17	-1.04	0.30
LTU	-0.41	-0.91	0.50	-2.90	0.00
MLT	0.00	-0.26	0.25	-1.64	0.10
NOR	-0.40	-0.67	0.27	-1.58	0.12
NZL	-0.12	-0.37	0.25	-1.51	0.13
PRT	-0.16	-0.32	0.16	-0.99	0.32
QAT	0.32	0.11	0.22	-1.44	0.15
RUS	-0.87	-1.08	0.21	-1.31	0.19
SAU	0.72	0.60	0.12	-0.77	0.44
SGP	-0.88	-1.42	0.54	-3.37	0.00
SVK	-0.28	-0.48	0.20	-1.25	0.21
SVN	0.12	-0.41	0.53	-3.00	0.00
SWE	-0.42	-0.80	0.38	-2.23	0.03
TWN	-0.34	-0.85	0.51	-2.77	0.01

Note: $Mean_N$= mean of nonlinear/electronic items; $Mean_L$= mean of linear/digital items

Discussion

Nonlinear reading is characterized by hypertext and interactivity, allowing readers to navigate flexibly between different sections, jump across links, and integrate information from multiple sources (Zumbach & Mohraz, 2007). This type of reading is prevalent in digital environments, where websites, multimedia articles, and interactive textbooks present content in a non-sequential manner. Hypertext has been noted for mirroring the associative nature of human memory, offering a dynamic, personalized reading experience (Shneiderman, 1989).

Nonlinear reading may introduce additional cognitive demands. Readers must engage in self-directed navigation, decide which links to follow, and synthesize fragmented pieces of information into a coherent understanding. This requires higher levels of metacognition, as individuals must constantly monitor their comprehension and adjust their reading path accordingly (Coiro et al., 2008).

In 2015, PISA introduced a significant change in its administration by moving from paper-based to computer-based assessments in most participating countries. As documented by

Jerrim et al. (2018), field trial data from Germany, Sweden, and Ireland revealed measurable differences in how students performed on identical items presented on screen versus on paper. Although the PISA contractors (ETS) attempted to account for and mitigate these effects, the comparability of PISA 2015 results with previous paper-based cycles remains somewhat uncertain. While Jerrim et al.'s investigation was constrained by reliance on field trial data, factors beyond reading were suggested as contributing to the observed effects, including differences in test-taking strategies and the logistical challenges of administering computer-based tests. These experiences with PISA 2015 offer valuable lessons for PIRLS 2021, which has similarly transitioned from paper to a digital format. Test administrators and researchers must consider how these changes might influence students' responses and should implement robust bridging or linking studies to preserve comparability with earlier cycles. Particular attention could also be paid to new item formats, the possibility of subgroup differences in computer familiarity, and how potential mode effects could influence background questionnaire data.

As reported in an earlier investigation using paper PIRLS 2016 and ePIRLS 2016, we found that a general reading factor accounted for most of the variance in reading performance, but a distinct nonlinear reading factor also emerged (Authors et al., under review). However, because the test modes were different (paper vs. computer), it remained unclear whether this factor was an inherent trait of nonlinear reading or an artifact of test mode effects. PIRLS 2021 marked a major milestone in international literacy assessment by introducing digitalPIRLS, a fully digital alternative to the traditional paper-based test. In contrast, ePIRLS remains a separate digital component designed to assess students' online reading comprehension skills. The present study, using digitalPIRLS 2021 and ePIRLS 2021, eliminates this confound by presenting both linear and nonlinear reading tasks on a computer. If a nonlinear reading factor is still observable under these conditions, it would provide

stronger evidence that nonlinear reading requires distinct cognitive abilities rather than being a result of differences in test mode.

In a conceptually similar study, Mahlow et al. (2020) used the same set of models, i.e., a unidimensional model, a correlated two factor model, and a bifactor (S-1) model to investigate the dimensionality of a combination of computer-based STC and MDC items. Their findings showed that the correlated two-factor model, where STC and MDC are modeled as two separate but correlated factors, has the best fit (latent correlation=.84). They concluded that although STC and MDC are highly correlated, they are separable constructs. This is different to our finding that the bifactor model turned out to be the best fitting model. The conclusions of the two studies diverge at this stage as the better fit of the bifactor (S-1) model in our study implies that the two types of reading share a common core, and nonlinear reading is just a factor which represents the peculiarities of this type of reading. However, the better fit of the correlated two-factor model in Mahlow et al.'s study implies that linear and nonlinear reading are two separate constructs which need to be attended to in their own rights. The reason of this divergence could be because Mahlow et al.'s study was with university students and ours was with 4th graders. The other reason might be that the teacher avatar used in ePIRLS to direct students to different sections of the text renders the ePIRLS tasks into, more or less, linear reading activities rather than true nonlinear reading where examinees have to search for the right texts and evaluate their relevance.

In this study, following Authors et al. (under review), we assumed that nonlinear reading requires not only reading ability but also an additional skill, such as the capacity to integrate multiple perspectives or maintain them in working memory. This assumption is reflected in the bifactor (S-1) model (Figure 1, Panel C). This model suggests that the ability to manage nonlinear reading is independent of general reading comprehension, whereas the two-factor model (Figure 1, Panel B) assumes that linear and nonlinear reading are related

but distinct skills that correlate. The first model (Figure 1, Panel A) posits that linear and nonlinear reading reflect a single construct. This model assumes a shared underlying factor for both types of comprehension.

Our findings showed that the bifactor (S-1) model has the best fit across all the countries. This suggests that a general factor of reading underlies the responses while a weaker specific factor of nonlinear reading is also present. In Authors et al. (under review), where the nonlinear items were presented on the computer and the linear items were delivered as paper-and-pencil items, a sharper distinction was observed between linear and nonlinear items. In contrast, in the present study, where all items were delivered digitally, the distinction between linear and nonlinear reading appears less pronounced.

Taken together, the findings from both studies highlight the influence of text structure on the structure of reading assessment. While the specific factor reliabilities in the earlier study ranged from .36 to .58—indicating some reliable variance for the nonlinear reading factor—in the current study, they range from .16 to .29, pointing to weaker evidence for a distinct nonlinear factor at the individual level. Nonetheless, the presence of even modest specific factor reliabilities suggests that linear and nonlinear reading, while sharing a common foundation, also capture different aspects of reading competence. This residual multidimensionality implies that the two item types should not be treated as interchangeable, even when delivered via the same digital medium.

Finally, our study underscores the importance of how reading comprehension is defined and measured. Although different reading assessments tend to correlate, both the administration mode (paper vs. digital) and the type of reading tasks (linear vs. nonlinear) matter. These findings have important implications for the comparability of test scores over time. Observed changes in PIRLS scores—or in their relationship with other variables such as

student gender—may reflect shifts in assessment design rather than actual changes in reading proficiency. If the goal is to measure change over time, the measure itself must remain consistent.

References

Afflerbach, P., & Cho, B.-Y. (2009). Identifying and describing constructively responsive comprehension strategies in new and traditional forms of reading. In S. E. Israel & G. G. Duffy (Eds.), *Handbook of research on reading comprehension* (pp. 69–90). Routledge.

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.

Bråten, I., & Strømsø, H. I. (2010a). Effects of task instruction and personal epistemology on the understanding of multiple texts about climate change. *Discourse Processes, 47*, 1–31. https://doi.org/10.1080/01638530902959646

Bråten, I., & Strømsø, H. I. (2010b). When law students read multiple documents about global warming: Examining the role of topic-specific beliefs about the nature of knowledge and knowing. *Instructional Science, 38*, 635–657. https://doi.org/10.1007/s11251-008-9091-4

Britt, M. A., & Rouet, J.-F. (2012). Learning with multiple documents: Component skills and their acquisition. In J. R. Kirby & M. J. Lawson (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 276–314). Cambridge University Press.

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*, 458–464.

Goldman, S. R. (2004). Cognitive aspects of constructing meaning through and across multiple texts. In N. Shuart-Faris & D. Bloome (Eds.), *Uses of intertextuality in*

classroom and educational research (pp. 313–347). Information Age.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*, 371–395. https://doi.org/10.1037/0033-295X.101.3.371

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

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

Jeong, Y. J., & Gweon, G. (2021). Advantages of print reading over screen reading: A comparison of visual patterns, reading performance, and reading attitudes across paper, computers, and tablets. *International Journal of Human–Computer Interaction, 37*(17), 1674–1684. https://doi.org/10.1080/10447318.2021.1908668

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163–182. https://doi.org/10.1037/0033-295X.95.2.163

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

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

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.). (2019). *PIRLS 2021 Assessment Frameworks*. TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/pirls2021/frameworks/

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). U.S. Government Printing Office.

Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Lawrence Erlbaum.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods, 21*(2), 137–150. https://doi.org/10.1037/met0000045

Schoor, C., Hahnel, C., Mahlow, N., Klagges, J., Kroehne, U., Goldhammer, F., & Zlatkin-Troitschanskaia, O. (2020). Multiple document comprehension of university students: Test development and relations to person and process characteristics. In O. Zlatkin-Troitschanskaia, H. A. Pant, M. Toepper, & C. Lautenbach (Eds.), *Student learning in German higher education – Innovative measurement approaches and research results* (pp. 221–240). Springer VS. https://doi.org/10.1007/978-3-658-27886-1_11

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). MIT Press.

Stadtler, M., & Bromme, R. (2014). The content–source integration model: A taxonomic

description of how readers comprehend conflicting scientific information. In D. N. Rapp & J. Braasch (Eds.), *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences* (pp. 379–402). MIT Press.

Trabasso, T., van den Broek, P., & Suh, S. Y. (1989). Logical necessity and transitivity of causal relations in stories. *Discourse Processes, 12*, 1–25. https://doi.org/10.1080/01638538909544717

von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E. M., & Berthold, K. (2016). The use of source-related strategies in evaluating multiple psychology texts: A student–scientist comparison. *Reading & Writing, 29*, 1677–1698. https://doi.org/10.1007/s11145-015-9601-0

Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83*, 73–87. https://doi.org/10.1037/0022-0663.83.1.73

Wry, E., & Mullis, I. V. S. (2023). Developing the PIRLS 2021 achievement instruments. In M. von Davier, I. V. S. Mullis, B. Fishbein, & P. Foy (Eds.), *Methods and Procedures: PIRLS 2021 Technical Report* (pp. 1.1-1.24). Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2101.kb7549

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

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science, 6*, 292–297. https://doi.org/10.1111/j.1467-9280.1995.tb00513.x