Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Cognitive Structure of Text Comprehension: A Psychometric Study of PIRLS-Format Items Among Primary School Students in Uzbekistan

Abduvali Ismailov
National Institute of Pedagogical Skills named after A.Avloni
Sardor Palvannazirov
National Institute of Pedagogical Skills named after A.Avloni
Kunduzkhon Ruzieva
ruziyeva.kunduzkhan@gmail.com
National Institute of Pedagogical Skills named after A.Avloni
Nilufar Mambetova
National Institute of Pedagogical Skills named after A.Avloni
Zubayda Murotalieva
National Institute of Pedagogical Skills named after A.Avloni

Research Article

Keywords: reading literacy, PIRLS, text comprehension, cognitive processes, language of instruction, psychometric analysis, Uzbekistan

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11565185/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Cognitive Structure of Text Comprehension: A Psychometric Study of PIRLS-Format Items Among Primary School Students in Uzbekistan

Abstract

Background. The cognitive-process model used in the Progress in International Reading Literacy Study (PIRLS) has rarely been tested in multilingual, non-Anglophone education systems. This study examined how the four PIRLS reading processes relate to one another among fourth-grade students in Tashkent, Uzbekistan, whether students perform differently across these processes, and whether these differences vary by language of instruction.

Methods. In a cross-sectional design, students completed two literary texts on a digital assessment platform; after exclusion of incomplete protocols, 1,835 entered the analysis of the first text and 2,231 the second. Each text carried 14 dichotomously and polytomously scored PIRLS-format items covering retrieval, straightforward inference, interpretation and evaluation. Analyses comprised descriptive and reliability statistics, exploratory factor analysis, repeated measures analysis of variance on a common proportion-of-maximum metric, and comparisons between Uzbek-medium and Russian-medium students.

Results. Internal consistency was acceptable for the full 14-item scale of each text ($\alpha = .733$ and .740) but low for the four process subscales ($\alpha = .298$–$.597$), so process-level findings are provisional. Three factors emerged in each text, with a basic component combining retrieval and inference items, but the remaining factors differed between texts. Cognitive process exerted a large within-subject effect ($\eta^2_p = .678$ and .758): performance fell significantly at every step from retrieval to evaluation in the first text, whereas in the second retrieval and inference did not differ and evaluation exceeded interpretation ($d_z = 0.937$), so the hierarchy is not monotone with cognitive demand. Russian-medium students outperformed Uzbek-medium students on every measure, most markedly for interpretation in the first text (d =

1.144); a significant Language × Process interaction (η^2_p = .129 and .078) showed the gap to be smallest for basic processes and largest for higher-order ones.

Conclusions. The findings suggest a broad distinction between basic and higher-order comprehension rather than a consistent four-step hierarchy. The variation in process structure across texts and the larger language-group differences for higher-order processes highlight the importance of examining cognitive-process measures within specific linguistic contexts. Because process-level reliability was limited and language-group comparisons were based on observed scores without measurement-invariance testing, the findings should be interpreted as evidence of differences in assessment performance rather than established differences in underlying reading ability.

Keywords

reading literacy; PIRLS; text comprehension; cognitive processes; language of instruction; psychometric analysis; Uzbekistan

Introduction

Reading literacy is treated in contemporary international research as a key indicator of educational quality and of students' functional preparedness for further learning. Within international large-scale assessment programs, it is defined as the ability to understand written texts, use their content, reflect on it, and apply the information obtained to achieve personal and social goals (OECD, 2019). Because these programs report reading achievement on a common metric across dozens of education systems, they have become a principal source of comparative evidence on how effectively school systems develop this competence (Mullis et al., 2023).

The conceptual framework of the present study is grounded in the model of reading literacy assessment developed within the Progress in International Reading Literacy Study (PIRLS). In the PIRLS 2021 assessment framework, text comprehension encompasses four cognitive processes: focusing on and retrieving explicitly stated information, making straightforward inferences, interpreting and integrating ideas and information, and evaluating and critiquing content and textual elements (Mullis & Martin, 2019). These four processes are carried forward unchanged into the framework for the current assessment cycle (von Davier & Kennedy, 2024), so the process model examined here remains the operational definition of comprehension in the study. The sequence reflects a progressive increase in the complexity of cognitive information processing, from working with explicit textual material to deep semantic reconstruction and reflective evaluation.

From a theoretical perspective, the PIRLS model aligns with cognitive theories of text comprehension. In the constructivist–integrative model proposed by Kintsch (1998), comprehension is conceptualized as the construction of a mental representation of content formed through the interaction between textual information and the reader's prior knowledge, and it entails a transition from a surface-level textual representation to a coherent semantic structure. Research in the cognitive psychology of reading likewise emphasizes the role of inferential processing, vocabulary knowledge, and metacognitive control in enabling deep interpretation of texts (Cain & Oakhill, 2011; Perfetti et al., 2005), and component models of comprehension describe these skills as hierarchically and dynamically related rather than as a single undifferentiated ability (Kim, 2020). The PIRLS model is therefore employed here as a theoretically grounded framework for the empirical examination of the structural organization of the cognitive processes underlying text comprehension among students at the final stage of primary education.

Given the age of the respondents, the developmental characteristics of younger schoolchildren also require consideration. Fourth-grade students (approximately 9–10 years old) are situated within the stage of concrete operational thinking described by Piaget (1952), which is characterized by the formation of logical structures of thought that still rely on concrete material and on visually meaningful representations of information. Operations such as classification and reversibility develop actively at this age, whereas the capacity for abstract reasoning and hypothetical-deductive analysis is still maturing. From the perspective of the cultural-historical theory of Vygotsky (1978), the development of higher mental functions at primary school age is associated with the internalization of external means of activity and with the formation of voluntary regulation and reflective control; in the context of text comprehension, this is manifested in a gradual transition from the superficial reproduction of information to the conscious construction of semantic relationships and the integration of content.

Contemporary research on cognitive development indicates that by the age of 9–10 the mechanisms of inferential processing and of mental model construction become substantially more advanced (Kintsch, 1998; Perfetti et al., 2005). Processes involving deep interpretation, critical evaluation, and metacognitive monitoring nevertheless remain more resource-intensive and depend on the development of executive functions and lexical-semantic knowledge, and meta-analytic evidence indicates that executive functions are systematically associated with reading outcomes in children (Zhou et al., 2026). Neuroimaging evidence further indicates that the association cortices, including the prefrontal regions that support voluntary control and analytical information processing, follow a protracted maturational trajectory extending well beyond this age (Sydnor et al., 2021). Taken together, these age-related characteristics suggest that an empirically identified hierarchy of cognitive processes may be interpreted not only as a feature of task structure but also as a reflection of

developmental patterns of thinking in primary school children, so that higher success rates on retrieval and straightforward inference than on interpretative and evaluative operations would be consistent with the normative trajectory of cognitive development at ages 9–10.

Identifying the hierarchy of cognitive processes underlying text comprehension and establishing their psychometric characteristics enables a more precise understanding of the structure of reading literacy and supports the development of targeted instructional approaches for its improvement (McNamara & Magliano, 2009; Shanahan & Shanahan, 2008). Contemporary requirements for educational measurement additionally necessitate rigorous psychometric validation of assessment instruments, including the analysis of internal consistency, item discrimination, and the structural organization of scales (Rutkowski et al., 2014). The application of PIRLS-format items within the education system of Uzbekistan therefore requires a thorough evaluation of their measurement quality and of their suitability for the constructs they are intended to capture.

Despite the extensive research literature that international large-scale assessments of reading have generated (Stiff et al., 2023), limited attention has been paid to the empirical validation of the cognitive structure of text comprehension in specific national and linguistic contexts. Studies that examine the comparability of PIRLS instruments across languages of instruction remain concentrated in a small number of settings (Mthimkhulu et al., 2024; Roux et al., 2022), and psychometric evidence on how the PIRLS cognitive framework operates within multilingual education systems is largely absent. The gap is consequential for Central Asian systems, which have begun to align national reading literacy policy with international assessment frameworks (Sarmurzin et al., 2025) but for which evidence on the measurement properties of PIRLS-format items remains scarce. The present study addresses this gap by providing an empirical analysis of the structure and hierarchy of cognitive processes in text

comprehension among primary school students in Uzbekistan, thereby contributing to the refinement of cognitive models of reading literacy and to the context-sensitive adaptation of international assessment frameworks.

Accordingly, the aim of the study is threefold: to identify the structural organization of the cognitive processes involved in the comprehension of literary texts among students of general secondary schools in Tashkent, Uzbekistan, on the basis of a psychometric analysis of PIRLS-format items; to establish the hierarchy of performance across these processes; and to examine whether performance differs by language of instruction and whether any such differences vary with the level of cognitive demand.

In accordance with this aim, the following hypotheses were formulated:

H1 (Structural hypothesis). Cognitive processes underlying text comprehension form an empirically distinguishable latent structure, identifiable through factor analysis, which may not fully correspond to the theoretical four-component model of PIRLS.

H2 (Hierarchical hypothesis). The level of task performance differs statistically significantly depending on the type of cognitive process: tasks at the basic level (information retrieval and making straightforward inferences) are performed more successfully than tasks requiring interpretation and evaluative-analytical processing of text.

H3 (Language-factor hypothesis). There are statistically significant differences in task performance between students with different languages of instruction, and these differences are more pronounced in tasks involving higher levels of cognitive processing.

The remainder of the paper is organized as follows. The Method section describes the sample, the instruments, and the analytic procedures. The Results section reports the descriptive and

reliability statistics, the factor structure obtained for each text, the comparison of performance across the four cognitive processes, and the differences associated with the language of instruction. The Discussion interprets these findings in relation to the three hypotheses, considers their implications for the use of PIRLS-format instruments in multilingual education systems, and sets out the limitations of the study.

Methods

Participants

Participants were fourth-grade students attending general secondary schools in Tashkent, Uzbekistan. The study included students from seven administrative districts for Text 1 and eight districts for Text 2. Approximately 2,200 fourth-grade students participated in the assessment overall.

All students completed two literary texts in PIRLS format, referred to here as Text 1 and Text 2. The data were obtained from the assessment platform of the A. Avloniy National Institute of Pedagogical Skills. The same student population completed both texts, although the number of students included in the analysis differed because only complete response protocols were retained for each text. The final analytic samples comprised 1,835 students for Text 1 and 2,231 students for Text 2.

Both Uzbek- and Russian-medium students were represented. In the Text 1 sample, 567 students (30.9%) were from Uzbek-medium schools and 1,268 (69.1%) from Russian-medium schools. In the Text 2 sample, 800 students (35.9%) were from Uzbek-medium schools and 1,431 (64.1%) from Russian-medium schools.

The Text 1 sample covered seven districts: Bektemir, Chilanzar, Mirabad, Olmazor, Shaykhantakhur, Uchtepa, and Yunusabad. The Text 2 sample included these districts as well as Mirzo-Ulugbek. Students attended 11 general secondary schools across the represented districts.

The sampling was conducted through schools and intact classes rather than through individual student selection. Because school and class identifiers were not available in the analytical dataset, the clustered sampling structure could not be incorporated into the statistical analyses. This limitation is considered when interpreting the results.

Ethics approval and consent to participate

The assessment was carried out under the written authorisation of the national ministry responsible for school education and was administered through the participating general secondary schools as part of system-level educational assessment; institutional names are withheld here for anonymised review and appear in full in the title-page file. Parents were informed in advance by the participating schools about the assessment and their children's participation in the study. Participation was voluntary, no personally identifying information was retained in the analytical dataset, and all responses were analysed in anonymised form.

Instruments

The diagnostic instrument consisted of two literary texts, each accompanied by a set of items constructed in accordance with the PIRLS assessment framework (Mullis & Martin, 2019) and distributed across its four cognitive processes: focus on and retrieve explicitly stated information (retrieval); make straightforward inferences (inference); interpret and integrate ideas and information (interpretation); and evaluate and critique content and textual elements (evaluation). Each text carried 14 items. Items were scored either dichotomously (0–1) or

polytomously (0–2; 0–3), the polytomous format allowing credit for partial understanding. Text 1 contains three polytomously scored items (two scored 0–2 and one scored 0–3) and Text 2 contains two (one scored 0–2 and one scored 0–3); this is why 14 items yield a maximum attainable total score of 18 in Text 1 and 17 in Text 2. The assessment was delivered on a national digital assessment platform in Uzbek and in Russian, each student receiving the version corresponding to their language of instruction. The number of items, their maximum scores, and their allocation to the four cognitive processes are given in Table 1.

Table 1. *Item Composition and Maximum Scores of the Four PIRLS Cognitive Processes in Each Text*

Text	Cognitive process	Items	Item maximum scores	Number of items (k)	Maximum score for the process
Text 1	Retrieval	V6, V8, V11, V16	1, 1, 1, 1	4	4
Text 1	Inference	V7, V10, V12, V14, V19	1, 1, 1, 1, 1	5	5
Text 1	Interpretation	V9, V13, V15, V18	1, 2, 1, 3	4	7
Text 1	Evaluation	V17	2	1	2
Text 1	**Total**			**14**	**18**
Text 2	Retrieval	V7, V10, V15	1, 1, 1	3	3

Text 2	Inference	V6, V8, V12, V14	1, 1, 1, 1	4	4
Text 2	Interpretation	V9, V11, V16, V17	1, 2, 1, 3	4	7
Text 2	Evaluation	V13, V18, V19	1, 1, 1	3	3
Text 2	**Total**			**14**	**17**

Note. Item labels are the identifiers used throughout this article, including in the item-level factor-analytic tables; the same labels denote different items in the two texts, because the two texts are distinct forms. Item maximum scores are listed in the same order as the items. The maximum score for a process is the sum of the maximum scores of its items. The evaluation measure of Text 1 consists of a single item.

Scoring and composite construction

The total score for each text is the sum of the scores on its 14 items and thus ranges from 0 to 18 in Text 1 and from 0 to 17 in Text 2.

To make the four processes comparable, process-level scores were calculated using a common 0–1 scale. Since the items within a process had different maximum possible scores, each item score was converted into a proportion of its maximum score. The resulting item-level scores were then averaged to obtain the composite score for each process. Thus, a score of 1 represents the maximum possible performance for a process, while 0 represents the minimum. This approach also ensures that items with different scoring ranges contribute equally to the process-level composite rather than being weighted according to their maximum scores.

All analyses reported in this article use the common 0–1 metric described above. This scoring approach differs from the unweighted mean of raw item scores used in the original analysis of

these data. For processes consisting entirely of dichotomous items, the two approaches produce equivalent scores apart from scale. For processes containing items with different maximum scores, however, the common-metric approach accounts for differences in item scoring ranges before calculating the composite. The results reported below are therefore based consistently on the common-metric scoring procedure.

Polytomous items (0–2 and 0–3) were scored automatically by the assessment platform using predefined scoring rules. No human rating or manual coding was involved.

Data Source and Collection

The data analysed in this study were obtained from the assessment platform operated by the A. Avloniy National Institute of Pedagogical Skills. The platform was used to administer PIRLS-based reading tasks to fourth-grade students and to record their responses and scores. The dataset used in the present analysis included students who completed the assessment in either Uzbek or Russian, corresponding to their language of instruction.

Statistical analysis

Analyses were conducted separately for the two texts. Because the items differed in their maximum scores, all process-level analyses were based on composites expressed as proportions of the maximum attainable score. Item-level scores were first divided by their respective maximum scores and then averaged within each cognitive process, giving each item equal weight.

Descriptive statistics were calculated for the total score and for each of the four process composites. Item facility ((p_i)) was used to describe item difficulty. For dichotomous items, this was the proportion of students answering correctly; for polytomous items, it was the mean score expressed as a proportion of the item's maximum possible score. Full-scale and process-level internal consistency were assessed using Cronbach's alpha. No reliability

coefficient was calculated for the Text 1 evaluation measure because it consisted of a single item.

The dimensional structure of the 14 items was examined separately for each text using exploratory factor analysis. Principal axis factoring with direct Oblimin rotation was used because the cognitive processes were expected to be related. The suitability of the data for factor analysis was assessed using the Kaiser–Meyer–Olkin (KMO) measure and Bartlett's test of sphericity. Factors with initial eigenvalues greater than 1 were retained. Because an oblique rotation was used, both pattern coefficients and factor correlations were examined. Loadings of |.30| or greater were considered salient.

Differences in performance across the four cognitive processes were examined using repeated-measures analysis of variance (ANOVA), with cognitive process as the within-subjects factor. The analysis was conducted separately for each text. Mauchly's test was used to assess the assumption of sphericity. As this assumption was not met for either text, multivariate results based on Pillai's Trace were used. Partial eta squared ((\eta_p^2)) was reported as the effect-size measure.

Following the omnibus test, pairwise comparisons between the four processes were conducted using estimated marginal means. Bonferroni adjustment was applied to the six comparisons within each text, resulting in an adjusted significance level of .0083. Within-subject effect sizes were expressed using Cohen's (d_z).

Differences between Uzbek-medium and Russian-medium students were examined separately for each text using independent-samples (t) tests for the total score and the four process composites. Levene's test was used to assess equality of variances, with Welch's test applied when this assumption was violated. Cohen's (d) was calculated using the pooled standard

deviation and reported with 95% confidence intervals. The direction of the group difference was determined from the group means, with absolute values of (d) reported in the tables.

To examine whether the language-group difference varied across the four cognitive processes, a mixed-design ANOVA was conducted separately for each text. Language of instruction (Uzbek-medium vs. Russian-medium) was specified as the between-subjects factor and cognitive process as the within-subjects factor. The language × cognitive process interaction provided the direct test of whether the magnitude of the language-group difference differed across processes. Multivariate results are reported using Pillai's Trace, together with the corresponding (F), degrees of freedom, (p), and partial eta squared.

All statistical tests were two-tailed and evaluated at (\alpha = .05, except for the Bonferroni-adjusted pairwise comparisons described above. Given the large sample sizes, effect sizes and their confidence intervals were considered alongside statistical significance when interpreting the substantive importance of the findings.

The original analyses of the assessment data were conducted using IBM SPSS Statistics 22. For the present study, analyses involving the rescored process composites were recomputed from the raw data using Python. Descriptive statistics, reliability estimates, and factor-analysis adequacy measures that were unaffected by the change in scoring metric were checked against the corresponding original results.

Results

Analytic Samples and Score Metric

The study examined the cognitive structure of literary text comprehension among fourth-grade students in general secondary schools in Tashkent on the basis of PIRLS-format tasks.

Both texts were administered to the same cohort, but because only fully completed protocols were retained for each text, the two analytic samples differ in size and in composition. Text 1 comprised 1,835 students, of whom 567 (30.9%) were taught in Uzbek and 1,268 (69.1%) in Russian; Text 2 comprised 2,231 students, of whom 800 (35.9%) were taught in Uzbek and 1,431 (64.1%) in Russian. Comparisons between the two texts are therefore comparisons between parallel datasets drawn from a single cohort, whereas comparisons among the four cognitive processes within a text are based on fully matched data.

Descriptive Statistics and Item Difficulty

Descriptive statistics, full-scale internal consistency and the sampling-adequacy indices for both texts are reported in Table 2. For Text 1 the mean total score was 8.98 of a maximum attainable score of 18 (SD = 3.80); for Text 2 it was 10.36 of a maximum attainable score of 17 (SD = 3.15). Because the two texts differ in the maximum attainable score, their raw total-score means are not directly comparable. The analysis of item difficulty indices showed substantial variability across items in both texts.

Table 2. *Descriptive Statistics, Internal Consistency and Sampling Adequacy for the Two Texts*

Indicator	Text 1	Text 2
n	1,835	2,231
Number of items	14	14
Maximum attainable score	18	17
M (total score)	8.98	10.36
SD	3.80	3.15

Cronbach's α (full scale)	.733	.740
Kaiser–Meyer–Olkin measure	.853	.854
Bartlett's test of sphericity, $\chi^2(91)$	3,733.53	3,485.21
Bartlett's test, p	< .001	< .001

Note. Total scores are sums of item scores across the 14 items of each text. Both texts were administered to the same cohort; the two analytic samples differ because only fully completed protocols were retained for each text.

Internal Consistency

Internal consistency of the full 14-item scale was comparable across the two texts (Text 1: Cronbach's α = .733; Text 2: α = .740). Reliability was also estimated separately for the four process subscales, which serve as the dependent variables in every process-level analysis reported below. These coefficients are given in Table 3.

None of the four subscales reached α = .60 in either text. In Text 1 the coefficients were α = .415 for Retrieve (k = 4), α = .597 for Inference (k = 5) and α = .501 for Interpret (k = 4); the Text 1 Evaluate measure consists of a single item, so no internal-consistency coefficient is estimable for it. In Text 2 the coefficients were α = .298 for Retrieve (k = 3), α = .422 for Inference (k = 4), α = .573 for Interpret (k = 4) and α = .409 for Evaluate (k = 3).

Table 3. *Internal Consistency of the Full Scale and of the Four Process Subscales*

Process	Text 1 k	Text 1 α	Text 2 k	Text 2 α
Retrieve	4	.415	3	.298
Inference	5	.597	4	.422
Interpret	4	.501	4	.573
Evaluate	1	not estimable	3	.409

Full scale	14	.733	14	.740

Note. k = number of items. Retrieve = focus on and retrieve explicitly stated information; Inference = make straightforward inferences; Interpret = interpret and integrate ideas and information; Evaluate = evaluate and critique content and textual elements. The Text 1 Evaluate measure comprises one item, for which an internal-consistency coefficient cannot be computed.

Factor Structure of Text Comprehension (H1)

Text 1. The suitability of the data for factor analysis was supported by a Kaiser–Meyer–Olkin measure of .853 and a significant Bartlett's test of sphericity, $\chi^2(91) = 3{,}733.53$, $p < .001$. Principal axis factoring with direct Oblimin rotation yielded three factors with eigenvalues greater than 1. The initial eigenvalues of these three factors were 3.553, 1.358 and 1.166, accounting cumulatively for 43.4% of the total variance. The pattern matrix is reported in Table 4.

Seven items defined the first factor: three retrieval items (V8, V11, V16) and four inference items (V10, V12, V14, V19). The second factor was defined by two interpretative items (V9, V15), the single evaluative item (V17) and one inference item (V7). The third factor was defined exclusively by the two polytomously scored interpretative items (V13, V18). One retrieval item (V6) did not reach a loading of |.30| on any factor. Factor correlations were F1–F2 = .340, F1–F3 = .345 and F2–F3 = .164 (range .16–.35).

Table 4. *Pattern Matrix for the 14 Items of Text 1 (Principal Axis Factoring, Direct Oblimin Rotation, n = 1,835)*

Item	PIRLS process	F1	F2	F3	h^2	Salient loading on
V6	Retrieve	.227	−.286	.227	.155	none
V7	Inference	.073	**.327**	.046	.138	F2
V8	Retrieve	**.431**	.025	−.099	.173	F1

V9	Interpret	.196	**.464**	.065	.339	F2
V10	Inference	**.567**	.035	.000	.336	F1
V11	Retrieve	**.574**	.046	.056	.376	F1
V12	Inference	**.629**	−.048	.003	.379	F1
V13	Interpret	−.116	.058	**.515**	.246	F3
V14	Inference	**.448**	.043	−.053	.201	F1
V15	Interpret	.143	**.486**	.085	.333	F2
V16	Retrieve	**.360**	.119	.055	.192	F1
V17	Evaluate	.103	**.421**	.141	.267	F2
V18	Interpret	.116	.250	**.534**	.467	F3
V19	Inference	**.533**	.072	.058	.342	F1

Note. All pattern coefficients are reported; none are suppressed. Loadings with an absolute value of .30 or greater are shown in bold and were treated as salient. Factor correlations: F1–F2 = .340, F1–F3 = .345, F2–F3 = .164. h^2 = extraction communality, derived from the archived pattern and structure matrices as the sum over factors of their products, an identity that holds exactly for an oblique solution. Kaiser–Meyer–Olkin measure = .853; Bartlett's test of sphericity, $\chi^2(91) = 3{,}733.53$, $p < .001$.

Text 2. Sampling adequacy was likewise supported (Kaiser–Meyer–Olkin measure = .854; Bartlett's test of sphericity, $\chi^2(91) = 3{,}485.21$, $p < .001$). Principal axis factoring with direct Oblimin rotation again yielded three factors. The initial eigenvalues were 3.277, 1.221 and 1.077, accounting cumulatively for 39.8% of the total variance. The pattern matrix is reported in Table 5.

The first factor was defined by two retrieval items (V7, V10) and three inference items (V6, V8, V14). The second factor was defined predominantly by interpretative items (V9, V11, V16, V17) but also carried one inference item (V12), one retrieval item (V15) and one

evaluative item (V19). The third factor was defined by two evaluative items (V13, V18). Every item loaded at |.30| or above on at least one factor. Factor correlations were F1–F2 = .568, F1–F3 = .340 and F2–F3 = .356 (range .34–.57).

Table 5. *Pattern Matrix for the 14 Items of Text 2 (Principal Axis Factoring, Direct Oblimin Rotation, n = 2,231)*

Item	PIRLS process	F1	F2	F3	h^2	Salient loading on
V6	Inference	**.348**	.084	.017	.128	F1
V7	Retrieve	**.421**	−.009	−.121	.192	F1
V8	Inference	**.425**	.047	.102	.193	F1
V9	Interpret	.006	**.375**	.100	.151	F2
V10	Retrieve	**.396**	.029	.121	.172	F1
V11	Interpret	.117	**.578**	−.138	.366	F2
V12	Inference	.007	**.396**	.023	.157	F2
V13	Evaluate	.234	−.024	**.379**	.199	F3
V14	Inference	**.344**	.113	.079	.137	F1
V15	Retrieve	−.046	**.476**	.127	.245	F2
V16	Interpret	−.034	**.482**	.013	.233	F2
V17	Interpret	−.052	**.546**	.042	.302	F2
V18	Evaluate	.022	.069	**.531**	.287	F3
V19	Evaluate	.063	**.447**	.050	.206	F2

Note. The original output of this analysis was not preserved, and the solution reported here was recomputed with the same procedure (principal axis factoring with direct Oblimin rotation); applied to Text 1, that procedure reproduces the archived Text 1 pattern matrix to a mean absolute difference of .021. All pattern coefficients are

reported; none are suppressed. Loadings with an absolute value of .30 or greater are shown in bold and were treated as salient. h^2 = extraction communality. Factor correlations: F1–F2 = .568, F1–F3 = .340, F2–F3 = .356. Kaiser–Meyer–Olkin measure = .854; Bartlett's test of sphericity, $\chi^2(91) = 3{,}485.21$, $p < .001$.

Comparison across the two texts. Two features of the solution were obtained in both texts: the number of factors (three) and the existence of a basic component on which retrieval and inference items predominantly load. The composition of the remaining factors was not the same. In Text 1 the single evaluative item loaded with the interpretative items, and the third factor was defined purely by the two polytomously scored interpretative items; in Text 2 the third factor is evaluative. The basic component itself is also less clean in Text 2, where one inference item (V12), one retrieval item (V15) and one evaluative item (V19) load away from their intended process, onto the interpretative factor. Factor correlations differ substantially between the two texts (Text 1: .16–.35; Text 2: .34–.57). The two solutions therefore constitute a partial replication and support the stability of the basic component; they do not show that the three-factor configuration as a whole was reproduced.

These results address the structural hypothesis (H1), which is partially supported: three factors and a basic component were obtained in both texts, but the composition of the remaining factors differs between texts and three Text 2 items load off-process. Both solutions are exploratory.

Performance Across the Four Cognitive Processes (H2)

Mean performance on each of the four processes, expressed as a proportion of the maximum attainable score, is reported with standard deviations and 95% confidence intervals in Table 6.

Table 6. *Performance on the Four PIRLS Cognitive Processes, Expressed as a Proportion of the Maximum Attainable Score*

Process	Text 1 M	Text 1 SD	Text 1 95% CI	Text 2 M	Text 2 SD	Text 2 95% CI

Retrieve	.743	.263	[.731, .755]	.813	.251	[.802, .823]
Inference	.603	.296	[.590, .617]	.811	.233	[.801, .821]
Interpret	.332	.272	[.320, .345]	.374	.258	[.363, .384]
Evaluate	.298	.350	[.282, .314]	.679	.315	[.666, .692]

Note. Text 1 n = 1,835; Text 2 n = 2,231. Each composite is the mean of its constituent items after every item has been divided by its own maximum score, so that all four measures are bounded by 0 and 1 and every item contributes equally. CI = confidence interval. The Text 1 Evaluate measure comprises one item.

A repeated measures analysis of variance was conducted separately for each text. Mauchly's test indicated that the assumption of sphericity was violated in both texts (Text 1: $W = .362$, $\chi^2(5) = 1{,}859.75$, $p < .001$; Text 2: $W = .786$, $\chi^2(5) = 535.43$, $p < .001$), and multivariate tests are therefore reported throughout. The effect of cognitive process was significant in both texts: for Text 1, Pillai's Trace = .678, $F(3, 1832) = 1{,}284.43$, $p < .001$, $\eta^2p = .678$; for Text 2, Pillai's Trace = .758, $F(3, 2228) = 2{,}328.31$, $p < .001$, $\eta^2p = .758$.

All six pairwise comparisons among the four processes were examined within each text on the estimated marginal means, with a Bonferroni correction for the family of six comparisons (per-comparison criterion $\alpha = .05/6 = .0083$). The results are reported in Table 7.

Table 7. *Bonferroni-Corrected Pairwise Comparisons Among the Four Cognitive Processes*

Text	Comparison	M difference	t	df	p (corrected)	d_z	Significant
1	Retrieve – Inference	.140	21.82	1834	< .001	0.509	Yes
1	Retrieve – Interpret	.411	57.31	1834	< .001	1.338	Yes
1	Retrieve – Evaluate	.445	48.90	1834	< .001	1.142	Yes
1	Inference – Interpret	.271	38.46	1834	< .001	0.898	Yes

1	Inference – Evaluate	.306	33.44	1834	< .001	0.781	Yes
1	Interpret – Evaluate	.035	4.25	1834	< .001	0.099	Yes
2	Retrieve – Inference	.002	0.33	2230	1.000	0.007	No
2	Retrieve – Interpret	.439	71.71	2230	< .001	1.518	Yes
2	Retrieve – Evaluate	.134	19.52	2230	< .001	0.413	Yes
2	Inference – Interpret	.437	76.83	2230	< .001	1.627	Yes
2	Inference – Evaluate	.132	20.29	2230	< .001	0.430	Yes
2	Evaluate – Interpret	.305	44.24	2230	< .001	0.937	Yes

Note. Text 1 n = 1,835; Text 2 n = 2,231. Comparisons are of estimated marginal means on the proportion-of-maximum metric. p (corrected) = probability after Bonferroni correction for six comparisons within each text; the per-comparison criterion was $\alpha = .0083$. d_z = within-subject Cohen's d. Each comparison is labeled so that the mean difference is positive; in the final row of Text 2 the higher mean is that of Evaluate.

In Text 1, all six comparisons were significant, yielding the ordering Retrieve > Inference > Interpret > Evaluate, in which every step is significant. The smallest of these differences was between interpretation and evaluation (M difference = .035, d_z = 0.099); all other contrasts had d_z values between 0.509 and 1.338.

In Text 2, retrieval and inference did not differ (M difference = .002, t = 0.33, p = 1.000, d_z = 0.007) and the remaining five comparisons were significant, yielding the ordering Retrieve ≈ Inference > Evaluate > Interpret, with interpretation the lowest-performing process.

The ordering is therefore not monotone with cognitive demand. In Text 2, evaluation significantly exceeded interpretation (M difference = .305, t = 44.24, df = 2230, p < .001, d_z = 0.937), so that the process treated as the most demanding in the PIRLS framework was performed more successfully than interpretation, and interpretation rather than evaluation was the process on which performance was lowest.

These results address the hierarchical hypothesis (H2), which is supported with one qualification. The basic processes significantly exceeded the higher-order processes in both texts, with all corresponding d_z values at 0.41 or above. The qualification is that the ordering is not monotone: in Text 2 evaluation significantly exceeded interpretation (d_z = 0.937), and retrieval and inference did not differ from each other.

Differences by Language of Instruction (H3)

Performance was compared between Uzbek-medium and Russian-medium students on the total score and on each of the four process measures, separately for each text. The results are reported in Table 8 (Text 1) and Table 9 (Text 2).

Table 8. *Differences by Language of Instruction, Text 1*

Measure	M (Uz)	SD (Uz)	M (Ru)	SD (Ru)	Test	t	df	p	d	95% CI for d
Total score	6.96	3.00	9.89	3.78	Welch	−17.80	1349.88	< .001	0.825	[0.722, 0.927]
Retrieve	.715	.278	.755	.255	Welch	−2.97	1009.71	.003	0.155	[0.056, 0.254]
Inference	.529	.281	.636	.296	Pooled	−7.29	1833	< .001	0.368	[0.269, 0.468]
Interpret	.142	.178	.417	.264	Welch	−26.15	1555.01	< .001	1.144	[1.038, 1.249]
Evaluate	.156	.313	.361	.348	Welch	−12.50	1199.96	< .001	0.607	[0.506, 0.708]

Note. Uzbek-medium n = 567; Russian-medium n = 1,268. The four process measures are proportions of the maximum attainable score; the total score is a sum with a maximum of 18. Welch = Welch's test, reported where

Levene's test was significant; Pooled = pooled-variance t test, which is why the degrees of freedom differ across rows. Cohen's d was computed with the pooled standard deviation and is reported as an absolute magnitude; the negative t values indicate higher mean performance among Russian-medium students on every measure. CI = confidence interval.

Table 9. *Differences by Language of Instruction, Text 2*

Measure	M (Uz)	SD (Uz)	M (Ru)	SD (Ru)	Test	t	df	p	d	95% CI for d
Total score	9.54	3.15	10.81	3.05	Pooled	−9.28	2229	< .001	0.410	[0.322, 0.497]
Retrieve	.794	.255	.823	.248	Welch	−2.64	1618.82	.008	0.117	[0.031, 0.204]
Inference	.794	.241	.820	.228	Pooled	−2.59	2229	.010	0.114	[0.028, 0.201]
Interpret	.312	.242	.408	.261	Welch	−8.72	1758.31	< .001	0.377	[0.290, 0.464]
Evaluate	.555	.316	.747	.293	Pooled	−14.45	2229	< .001	0.638	[0.549, 0.726]

Note. Uzbek-medium n = 800; Russian-medium n = 1,431. The four process measures are proportions of the maximum attainable score; the total score is a sum with a maximum of 17. Welch = Welch's test, reported where Levene's test was significant; Pooled = pooled-variance t test, which is why the degrees of freedom differ across rows. Cohen's d was computed with the pooled standard deviation and is reported as an absolute magnitude; the negative t values indicate higher mean performance among Russian-medium students on every measure. CI = confidence interval.

Differences were significant on the total score and on all four process measures in both texts. Within each text the standardized differences were smallest for the basic processes and largest for the higher-order processes: in Text 1 they rose from d = 0.155 for retrieval and d = 0.368 for inference to d = 0.607 for evaluation and d = 1.144 for interpretation, and in Text 2 from d = 0.117 for retrieval and d = 0.114 for inference to d = 0.377 for interpretation and d = 0.638 for evaluation.

Whether the size of the language difference varies across cognitive processes was tested directly by means of the Language × Process interaction in each text; the results are reported in Table 10. The interaction was significant in both texts: for Text 1, Pillai's Trace = .129, $F(3, 1831) = 90.28$, $p < .001$, $\eta^2p = .129$; for Text 2, Pillai's Trace = .078, $F(3, 2227) = 62.95$, $p < .001$, $\eta^2p = .078$. Averaged over the four processes, the language main effect was also significant in both texts (Text 1: $t = -16.85$, $p < .001$; Text 2: $t = -10.17$, $p < .001$).

Table 10. *Language × Process Interaction in Each Text*

Text	Effect	Pillai's Trace	F	df	p	η^2p
1	Language × Process	.129	90.28	3, 1831	< .001	.129
2	Language × Process	.078	62.95	3, 2227	< .001	.078

Note. Text 1 n = 1,835; Text 2 n = 2,231. Multivariate tests are reported, as for the within-subject effect of cognitive process. Language of instruction has two levels (Uzbek-medium, Russian-medium) and cognitive process four levels, measured on the proportion-of-maximum metric.

These results address the language-factor hypothesis (H3), which is supported: performance differed significantly by language of instruction on every measure in both texts, and the Language × Process interaction was significant in both ($\eta^2p = .129$ and .078), showing that the size of the language difference is not the same across the four processes. Measurement invariance across the Uzbek and Russian versions of the instrument was not established, so all comparisons reported in this subsection are comparisons of observed scores.

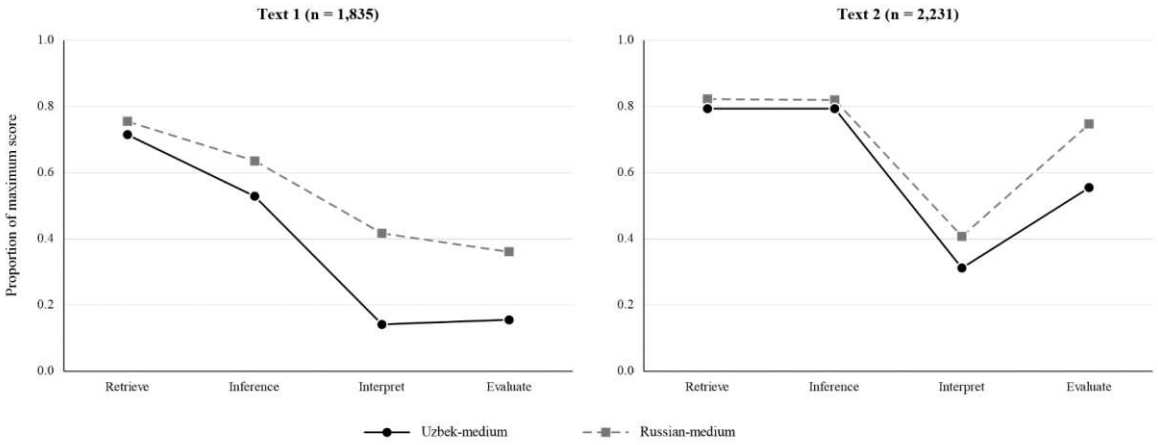


Fig. 1 Performance on the four PIRLS cognitive processes by language of instruction. Values are proportions of the maximum attainable score for each process, as reported in Tables 8 and 9. Text 1 n = 1,835 (567 Uzbek-medium, 1,268 Russian-medium); Text 2 n = 2,231 (800 Uzbek-medium, 1,431 Russian-medium)

Discussion

This study examined how the four PIRLS cognitive processes are represented among fourth-grade students in Tashkent and whether performance across these processes varies by language of instruction. Overall, the findings provide partial support for the PIRLS process framework, while also showing differences in how the processes are organised and performed across the two texts.

Cognitive structure of the reading processes

The first hypothesis was only partly supported. Exploratory factor analysis produced three factors in both texts rather than four clearly separated dimensions. Retrieval and straightforward inference consistently converged on the same factor, while the composition of the remaining factors differed between the texts. The resulting structure can be interpreted as a distinction between a basic component involving retrieval and straightforward inference and higher-order processing involving interpretation and evaluation. This interpretation is consistent with cognitive accounts that distinguish relatively surface-level processing from

processes requiring knowledge integration, situation-model construction, and metacognitive control (Kintsch, 1998; Perfetti et al., 2005).

The finding also needs to be considered in relation to the way PIRLS conceptualises and operationalises reading. Although the framework distinguishes four comprehension processes (Mullis & Martin, 2019), previous analyses of large-scale reading assessment data have often identified a dominant general factor alongside more specific dimensions (Martin et al., 2017; Rijmen, 2011; Foorman et al., 2015; Pokropek et al., 2022; von Davier et al., 2023; Baghaei et al., 2025). Research also suggests that the dimensional structure of reading may vary among second-language learners depending on proficiency (Min et al., 2022). The present findings are therefore broadly consistent with the view that theoretically distinct processes may not always emerge as fully independent empirical dimensions.

Performance across cognitive processes

The second hypothesis was supported, although the results did not show a uniform four-step hierarchy. In both texts, retrieval and straightforward inference were performed better than the higher-order processes, with large overall differences across the four processes. This broad distinction is compatible with developmental accounts suggesting that higher-order comprehension draws more strongly on linguistic, executive, and metacognitive resources (Kim, 2020; Vettori et al., 2024).

The ordering within the processes, however, differed between texts. In Text 1, performance declined progressively from retrieval to inference, interpretation, and evaluation. In Text 2, retrieval and inference did not differ, while evaluation was performed better than interpretation. Thus, the findings support a distinction between basic and higher-order

processing rather than a fixed progression in which performance decreases steadily with increasing cognitive demand. This is also consistent with the broader view that reading comprehension involves multiple interacting processes rather than a simple linear sequence (Cain & Oakhill, 2011).

The difference between the texts may partly reflect differences in the items used to represent each process. The number of items and their scoring formats were not identical across processes, and polytomous items were concentrated in the interpretation measure. The observed process differences should therefore be interpreted in relation to the particular assessment materials rather than assumed to represent fixed differences in the underlying cognitive processes.

Language of instruction

The third hypothesis was supported. Russian-medium students achieved higher scores than Uzbek-medium students on the total score and on each process measure in both texts. More importantly, the significant Language × Process interactions showed that the size of the difference varied across cognitive processes. The gap was smallest for the basic processes and larger for the higher-order processes, particularly interpretation in Text 1 and evaluation in Text 2.

One possible interpretation is that higher-order comprehension places greater demands on the academic language resources required to construct and evaluate meaning. This interpretation is consistent with Cummins's distinction between basic interpersonal communicative skills and cognitive academic language proficiency (Cummins, 1979, 2000), as well as research on cross-linguistic transfer in reading and the role of language proficiency in the dimensionality of second-language reading (Melby-Lervåg & Lervåg, 2011; Ke et al., 2021; Min et al.,

2022). However, this explanation cannot be established directly from the present data because language proficiency was not measured.

The language findings also require caution because measurement invariance and differential item functioning were not tested. Observed differences between language groups may reflect differences in comprehension as well as differences in how items function across languages, including possible effects of translation, adaptation, or cultural context (Rutkowski et al., 2014; Joo et al., 2022). Previous research in multilingual large-scale assessments has demonstrated the importance of examining such comparability through measurement invariance and DIF analyses (Mthimkhulu et al., 2024; Roux et al., 2022; Gökçe et al., 2021; Park et al., 2025). The present language-group results should therefore be viewed as differences in observed assessment performance rather than definitive evidence of differences in underlying reading ability. The interpretation is also consistent with concerns that conventional invariance criteria may behave differently in large-scale assessment samples (Rutkowski & Svetina, 2014).

Implications

The findings suggest that process-level results in multilingual large-scale assessments should be interpreted in relation to the specific assessment materials and language context. The convergence of retrieval and straightforward inference across the two texts provides some evidence for a basic component, while the variation in the higher-order factors indicates that the empirical structure of the processes may depend partly on the items used. The larger language-group differences observed for higher-order processes also highlight the need for careful translation, cultural review, and cross-linguistic validation of such items (International Test Commission, 2018; Gökçe et al., 2021; Roux et al., 2022; D'Agostino et al., 2025).

These findings are relevant to the broader ILSA literature because evidence from multilingual and less-represented education systems provides an opportunity to examine whether assessment frameworks developed in other contexts operate similarly across populations. Recent work involving PIRLS and other post-Soviet contexts highlights the value of such evidence for examining the generalisability of established assessment frameworks (Mullis et al., 2023; Stiff et al., 2023; Sarmurzin et al., 2025).

Limitations and future research

Several limitations should be considered. The study was conducted with students in Tashkent and used two literary texts, limiting the generalisability of the findings. The factor analysis was exploratory and therefore requires confirmation using a broader set of texts and items. The low internal consistency of the process subscales also limits the stability of process-level comparisons, and evaluation in Text 1 was represented by a single item.

In addition, language comparisons were based on observed scores without formal measurement-invariance or DIF testing, and the clustered sampling structure was not accounted for in the analyses. These limitations are particularly relevant when interpreting the language-group differences. Future research should use larger and more varied assessment materials, confirm the factor structure using confirmatory methods, examine item functioning across languages, and account for the sampling design when estimating standard errors.

Conclusions

This study examined the organisation and performance of the four PIRLS cognitive processes among fourth-grade students in Tashkent, with particular attention to language of instruction. The findings provide partial support for the four-process framework. Although four processes were theoretically distinguished, the exploratory factor analysis produced three factors in both

texts, with retrieval and straightforward inference consistently showing a close empirical relationship.

Performance also differed across the processes. Students generally performed better on retrieval and straightforward inference than on interpretation and evaluation, although the precise ordering of the processes varied between texts. This suggests that the processes may be better understood as reflecting a broader distinction between basic and higher-order comprehension rather than as a fixed four-level hierarchy.

Language of instruction was associated with performance across both texts. Russian-medium students achieved higher scores than Uzbek-medium students, and the significant Language × Process interactions showed that the size of this difference varied across processes. The larger differences observed for higher-order processes suggest that language may be particularly relevant when interpreting performance on more demanding comprehension tasks.

Overall, the findings highlight the importance of examining the empirical structure of cognitive-process measures within specific linguistic and educational contexts. They also indicate that process-level comparisons across language groups should be interpreted carefully, particularly when the comparability of measures across languages has not been formally established. Further research using a larger and more diverse set of texts and items, together with confirmatory and cross-linguistic measurement analyses, would help establish whether the patterns observed in Tashkent extend to other multilingual education settings.

List of abbreviations

ANOVA: analysis of variance; BICS: basic interpersonal communicative skills; CALP: cognitive academic language proficiency; CFA: confirmatory factor analysis; CI: confidence interval; DIF: differential item functioning; EFA: exploratory factor analysis; ILSA:

international large-scale assessment; IRT: item response theory; KMO: Kaiser–Meyer–Olkin; PAF: principal axis factoring; PIRLS: Progress in International Reading Literacy Study; SD: standard deviation.

Declarations

Availability of data and materials

The datasets generated and analysed during the current study are not publicly available because they contain assessment data on minors collected in identifiable school populations, but they are available from the corresponding author on reasonable request.

References

Baghaei, P., Grammatikopoulou, E., Johansson, S., & Strietholt, R. (2025). Comparing linear and nonlinear reading assessments in PIRLS 2016. *Studies in Educational Evaluation, 87*, Article 101513. https://doi.org/10.1016/j.stueduc.2025.101513

Cain, K., & Oakhill, J. (2011). Matthew effects in young readers: Reading comprehension and reading experience aid vocabulary development. *Journal of Learning Disabilities, 44*(5), 431–443. https://doi.org/10.1177/0022219411410042

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*(2), 222–251. https://doi.org/10.3102/00346543049002222

Cummins, J. (2000). *Language, power and pedagogy: Bilingual children in the crossfire*. Multilingual Matters.

D'Agostino, T. J., Liu, C., Soares, F., Conaghan, B., Brunette, T., Nakamura, P., Donovan, C., & Pentimonti, J. M. (2025). Seeing emerging comprehenders: Refining reading comprehension assessment in international contexts. *International Journal of Educational Development, 117*, Article 103373. https://doi.org/10.1016/j.ijedudev.2025.103373

Foorman, B. R., Koon, S., Petscher, Y., Mitchell, A., & Truckenmiller, A. (2015). Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. *Journal of Educational Psychology, 107*(3), 884–899. https://doi.org/10.1037/edu0000026

Gökçe, S., Berberoğlu, G., Wells, C. S., & Sireci, S. G. (2021). Linguistic distance and translation differential item functioning on Trends in International Mathematics and Science Study mathematics assessment items. *Journal of Psychoeducational Assessment, 39*(6), 728–745. https://doi.org/10.1177/07342829211010537

International Test Commission. (2018). ITC guidelines for translating and adapting tests (second edition). *International Journal of Testing, 18*(2), 101–134. https://doi.org/10.1080/15305058.2017.1398166

Joo, S., Ali, U., Robin, F., & Shin, H. J. (2022). Impact of differential item functioning on group score reporting in the context of large-scale assessments. *Large-scale Assessments in Education, 10*, Article 18. https://doi.org/10.1186/s40536-022-00135-7

Ke, S., Miller, R. T., Zhang, D., & Koda, K. (2021). Crosslinguistic sharing of morphological awareness in biliteracy development: A systematic review and meta-analysis of correlation coefficients. *Language Learning, 71*(1), 8–54. https://doi.org/10.1111/lang.12429

Kim, Y.-S. G. (2020). Hierarchical and dynamic relations of language and cognitive skills to reading comprehension: Testing the direct and indirect effects model of reading (DIER). *Journal of Educational Psychology, 112*(4), 667–684. https://doi.org/10.1037/edu0000407

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

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2017). *Methods and procedures in PIRLS 2016*. Boston College, TIMSS & PIRLS International Study Center.

McNamara, D. S., & Magliano, J. (2009). Toward a comprehensive model of comprehension. In B. Ross (Ed.), *Psychology of learning and motivation* (Vol. 51, pp. 297–384). Academic Press. https://doi.org/10.1016/S0079-7421(09)51009-2

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading, 34*(1), 114–135. https://doi.org/10.1111/j.1467-9817.2010.01477.x

Min, S., Bishop, K., & Cook, H. G. (2022). Reading is a multidimensional construct at child-L2-English-literacy onset, but comprises fewer dimensions over time: Evidence from multidimensional IRT analysis. *Language Testing, 39*(2), 265–288. https://doi.org/10.1177/02655322211045296

Mthimkhulu, S., Roux, K., & Mihai, M. (2024). Investigating measurement invariance in PIRLS 2021 across English and isiZulu language groups. *Reading & Writing, 15*(1), Article a455. https://doi.org/10.4102/rw.v15i1.455

Mullis, I. V. S., & Martin, M. O. (2019). *PIRLS 2021 assessment frameworks*. Boston College, TIMSS & PIRLS International Study Center.

Mullis, I. V. S., von Davier, M., Foy, P., Fishbein, B., Reynolds, K. A., & Wry, E. (2023). *PIRLS 2021 international results in reading*. Boston College, TIMSS & PIRLS International Study Center. https://doi.org/10.6017/lse.tpisc.tr2103.kb5342

OECD. (2019). *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

Park, J. Y., Joo, S., Li, Z., & Yoon, H. (2025). Measurement invariance for multilingual learners using item response and response time in PISA 2018. *Educational Measurement: Issues and Practice, 44*(1), 55–65. https://doi.org/10.1111/emip.12640

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Blackwell Publishing. https://doi.org/10.1002/9780470757642.ch13

Piaget, J. (1952). *The origins of intelligence in children*. International Universities Press.

Pokropek, A., Marks, G. N., Borgonovi, F., Koc, P., & Greiff, S. (2022). General or specific abilities? Evidence from 33 countries participating in the PISA assessments. *Intelligence, 92*, Article 101653. https://doi.org/10.1016/j.intell.2022.101653

Rijmen, F. (2011). Hierarchical factor item response theory models for PIRLS: Capturing clustering effects at multiple levels. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments, 4*, 59–74.

Roux, K., van Staden, S., & Pretorius, E. J. (2022). Investigating the differential item functioning of a PIRLS Literacy 2016 text across three languages. *Journal of Education*, (87), 1–21. https://doi.org/10.17159/2520-9868/i87a07

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement, 74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Rutkowski, L., von Davier, M., & Rutkowski, D. (Eds.). (2014). *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis*. Chapman & Hall/CRC.

Sarmurzin, Y., Kerimbekova, B., Toleubaeva, K., Zhunusova, M., Amanzhol, N., & Zhumagulov, A. (2025). "Forced us to adopt alternative pedagogical approaches": Kazakhstani reading literacy improvement initiatives. *Research in Comparative and International Education, 20*(3), 397–421. https://doi.org/10.1177/17454999251342142

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review, 78*(1), 40–59. https://doi.org/10.17763/haer.78.1.v62444321p602101

Stiff, J., Lenkeit, J., Hopfenbeck, T. N., Kayton, H. L., & McGrane, J. A. (2023). Research engagement in the Progress in International Reading Literacy Study: A systematic review.

Educational Research Review, 40, Article 100547.
https://doi.org/10.1016/j.edurev.2023.100547

Sydnor, V. J., Larsen, B., Bassett, D. S., Alexander-Bloch, A., Fair, D. A., Liston, C., Mackey, A. P., Milham, M. P., Pines, A., Roalf, D. R., Seidlitz, J., Xu, T., Raznahan, A., & Satterthwaite, T. D. (2021). Neurodevelopment of the association cortices: Patterns, mechanisms, and implications for psychopathology. *Neuron, 109*(18), 2820–2846.
https://doi.org/10.1016/j.neuron.2021.06.016

Vettori, G., Casado-Ledesma, L., Tesone, S., & Tarchi, C. (2024). Key language, cognitive and higher-order skills for L2 reading comprehension of expository texts in English as foreign language students: A systematic review. *Reading and Writing, 37*(9), 2481–2519.
https://doi.org/10.1007/s11145-023-10479-3

von Davier, M., & Kennedy, A. (Eds.). (2024). *PIRLS 2026 assessment frameworks*. Boston College, TIMSS & PIRLS International Study Center.
https://doi.org/10.6017/lse.tpisc.tr2103.kb4199

von Davier, M., Mullis, I. V. S., Fishbein, B., & Foy, P. (Eds.). (2023). *Methods and procedures: PIRLS 2021 technical report*. Boston College, TIMSS & PIRLS International Study Center. https://pirls2021.org/methods

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Zhou, T., Nguyen, T. A., Maba, A. P., Thomas, D., Kim, Y.-S. G., & Edwards, E. J. (2026). Executive functions and reading in children: A systematic review and meta-analysis.

703 *Educational Psychology Review, 38*(1), Article 67. https://doi.org/10.1007/s10648-026-
704 10160-5