CJPE: *Cokroaminoto Journal of Primary Education*
https://e-journal.my.id/cjpe
Volume 9 | Nomor 4 | 2026
e-ISSN: 2654-6434 dan p-ISSN: 2654-6426

Improving Reading Literacy through Diagnostic Assessment-Based Differentiated Instruction: An Analysis of Three AKM Cognitive Levels among Fifth-Grade Primary School Students

Harni [1*], Riska [2], Harmita Sari [3], Masnur [4]

Corespondensi Author
[1, 2, 3, 4] Pedagogi, Universitas Muhammadiyah Palopo, Indonesia

Email:
harniferry8138@gmail.com
nurspd812@guru.sd.belajar.id
harmitasari93@gmail.com
masnur1985@gmail.com

Keywords:
Reading Literacy; Differentiated Instruction; Diagnostic Assessment; AKM Cognitive Levels; Primary School.

Abstract. *This study examined reading literacy gains following diagnostic assessment-based differentiated instruction and identified achievement patterns across three cognitive levels of Indonesia's Minimum Competency Assessment (AKM). A pre-experimental one-group pretest–posttest design was employed with fifth-grade students at SDN 53 Sawerigading, Palopo. Twenty students completed the initial assessment, and paired analyses included 18 students with complete data. Instruments comprised two reading test forms and a diagnostic questionnaire. Analyses compared all 20 items with an 18-item scoring scope that excluded two items according to a decision recorded in the research documents. Data were analysed using descriptive statistics, paired-samples t-tests, Wilcoxon signed-rank tests for the 18-item scope, Cohen's dz, and normalized gain. For 20 items, mean scores increased from 69.09 to 84.47, with a mean individual normalized gain of 0.498. For 18 items, scores increased from 74.13 to 82.86, with a normalized gain of 0.337. Both differences were statistically significant, and both mean normalized gains were classified as moderate. Cognitive-level analysis within the 18-item scope showed the largest improvement in evaluating and reflecting, from 50.00% to 72.69%, while two indicators involving tabular information remained unchanged. The direction of improvement was consistent, although its magnitude depended on the scoring scope. Disaggregated analysis helped identify instructional priorities. The absence of a comparison group prevents causal attribution.*

This work is licensed under a Creative Commons Attribution 4.0 International License

Introduction

Reading literacy is a fundamental ability that enables students to understand information, build knowledge, and apply what they read to solve problems in everyday life. In primary education, reading literacy supports language development and conceptual understanding across various school subjects. Students need to understand explanations, follow instructions, connect information, and justify their answers. Therefore, successful reading instruction should be assessed based on students' ability to construct meaning and use information appropriately. Developing reading

comprehension requires multiple instructional components because word recognition alone does not guarantee adequate understanding (Duke et al., 2021).

Reading comprehension involves interconnected cognitive processes that may develop at different rates. Contemporary reading research extends beyond a simple distinction between word recognition and language comprehension by considering the complex processes involved in successful reading (Duke & Cartwright, 2021). The complexity of reading comprehension also highlights the need to connect research findings with instructional practices (Elleman & Oslund, 2019). In primary education, these perspectives support the assessment of different reading abilities while acknowledging their interrelationships. Students who read fluently must also demonstrate their ability to understand relationships between ideas, draw inferences, and formulate responses supported by textual evidence. Therefore, reading tasks should align with instructional objectives to identify students' abilities and determine where additional support is needed.

Strengthening reading literacy remains an important concern in Indonesian education. The results of the Programme for International Student Assessment (PISA) 2022 indicated that Indonesian students' reading achievement was below the Organisation for Economic Co-operation and Development (OECD) average (OECD, 2023). Although PISA participants are approximately 15 years old and do not directly represent primary school students, these findings provide important context for understanding national literacy challenges. They also reinforce the importance of developing reading comprehension and reasoning skills from the early stages of education. These abilities need to be strengthened before students encounter increasingly complex, lengthy, and diverse texts at subsequent educational levels. Primary education plays an important role in preparing students to meet these demands through age-appropriate activities that encourage them to understand, compare, and apply information.

Indonesia's Minimum Competency Assessment, known as Asesmen Kompetensi Minimum (AKM), provides a framework for assessing reading literacy through three cognitive levels: locating information, interpreting and integrating information, and evaluating and reflecting. Locating information involves retrieving explicitly stated information from a text. Interpreting and integrating information requires students to understand relationships between pieces of information and draw inferences. Evaluating and reflecting involves assessing textual content and connecting it with knowledge and experiences beyond the text (Pusat Asesmen dan Pembelajaran, 2020). This framework enables reading achievement to be examined according to different cognitive demands and helps identify abilities requiring targeted instructional support. It also provides a structured basis for classroom assessment without reducing all aspects of reading comprehension to a single overall score.

Distinguishing between cognitive processes has important implications for instructional planning. Students who can identify a character's name or retrieve factual details may still struggle to explain relationships between events, evaluate a character's decisions, or develop evidence-based opinions. When these abilities are summarized in a single score, differences in students' learning needs may be overlooked. Two students with identical overall scores may demonstrate different strengths and weaknesses across reading indicators. Similarly, improvements following instruction may occur in some indicators while others remain unchanged. Therefore, analyzing achievement by cognitive level is necessary to determine which reading abilities have developed and which require further instructional attention.

These needs are particularly relevant to fifth-grade students, whose reading activities increasingly support knowledge acquisition across subjects. Students encounter narratives, explanatory texts, instructions, and information presented alongside tables and illustrations. Such reading activities may require them to identify evidence, compare information, connect ideas, and communicate their understanding. Differences in vocabulary knowledge, reading experience, and the ability to understand relationships within texts can influence students' performance. Teachers therefore need specific information about students' learning needs to provide appropriate challenges and support for those experiencing difficulties with particular comprehension components. This information also helps determine how learning progress should be assessed, especially when overall achievement scores conceal weaknesses in specific reading abilities.

Differentiated instruction provides a framework for responding to differences in students' learning needs. In reading instruction, differentiation may involve adjustments to content, learning processes, and learning products, supported by assessment information that guides instructional decisions. A review of 28 empirical studies identified various forms of differentiated instruction in primary education, including adjustments to instructional content and processes and the strengthening of explicit reading instruction (Achmad et al., 2024). Differentiation therefore involves decisions about what students learn, how they receive instructional support, and how they demonstrate their understanding. Variations in learning activities are pedagogically meaningful when they correspond to students' learning needs and the competencies being developed. Nevertheless, common learning objectives should be maintained to ensure that different instructional activities remain academically purposeful and contribute to comparable expectations for reading comprehension.

Adjustments to instructional content may include selecting appropriate texts, providing vocabulary explanations, and incorporating relevant visual support. Adjustments to learning processes may involve guided reading, paired discussions, prompting questions, or independent activities with varying levels of assistance. Learning products may include oral responses, story maps, or written assignments. However, these instructional choices must remain aligned with clearly defined reading objectives. When an objective requires students to support their answers with textual evidence, each assessment format should allow them to demonstrate their ability to identify evidence and explain their reasoning. The diversity of approaches and practices associated with differentiated instruction highlights the importance of clearly defining how differentiation is understood and implemented in classroom settings (Gheyssens et al., 2022). Clear reading objectives connect instructional support with assessment decisions.

Diagnostic assessment provides a foundation for determining appropriate instructional adjustments. Indonesian learning and assessment guidelines emphasize the importance of considering students' circumstances when planning instruction and assessment (Anggraena et al., 2022). In this study, an initial test was used to map students' reading achievement, while a questionnaire gathered information about their perceived difficulties, reading interests, and activity preferences. These instruments served distinct but complementary purposes. Responses indicating preferences for individual or group activities were interpreted as preferences at the time of data collection. Decisions about instructional support also considered students' achievement and the demands of specific reading tasks. Students' preferences were not treated as fixed ability categories or as evidence of permanent learning styles.

Assessment data need to be translated into concrete instructional decisions. A review of reading interventions for struggling readers examined approaches involving data-based decision-making and the use of assessment information to guide instructional support (Filderman et al., 2018). In classroom practice, identifying students' needs should be followed by a clear explanation of the assistance provided and the abilities targeted. Otherwise, assessment may remain merely descriptive without informing instructional improvement. Diagnosis, instructional design, and achievement evaluation should therefore form a traceable sequence. Establishing these connections is important for interpreting students' progress at the end of instruction and distinguishing observed score differences from evidence concerning the instructional processes associated with those differences.

Implementing differentiated instruction requires teachers to make complex instructional decisions. Differentiation is a complex teaching skill that involves responding appropriately to students' varied learning needs (Van Geel et al., 2019). In classroom research, this complexity requires clear descriptions of identified needs, the support provided, and the conditions under which instruction is implemented. Merely naming an instructional approach does not sufficiently explain students' learning experiences. Similarly, assigning students to different groups does not necessarily demonstrate that instruction has been adjusted to their needs. Research should describe the content of learning activities, the types of assistance provided, and the relationships between tasks and instructional objectives. Clear reporting helps other teachers understand how an approach can be implemented in different contexts and determine which elements are relevant to their own classrooms.

Previous evidence indicates that the benefits of differentiated instruction depend on its implementation and educational context. A synthesis of 21 studies in primary education identified benefits associated with several differentiated instructional practices, including computer-supported systems and differentiation implemented within broader educational programs (Deunk et al., 2018). These findings suggest that differentiated instruction should not be treated as a uniform intervention. The use of the same terminology does not necessarily indicate equivalence in instructional content, intensity, or the level of support provided. Classroom research therefore needs to describe students' initial achievement and the outcomes measured so that findings can be compared appropriately with previous evidence. Differences in students' baseline achievement also influence the potential for measurable improvement and should be considered when interpreting subsequent changes in scores.

More specifically, a meta-analysis of differentiated literacy instruction reported a positive weighted mean effect, with a Hedges' g of 0.13 (Puzio et al., 2020). This finding supports the potential value of differentiated literacy instruction while indicating the need for proportionate interpretations of its benefits. A review of differentiated instruction in secondary education also demonstrated diversity in implementation practices and research conditions (Smale-Jacobse et al., 2019). Although evidence from secondary education provides a useful comparative perspective, it cannot directly establish the extent of improvement among fifth-grade students. Research in primary education is therefore necessary to examine reading abilities in relation to students' specific learning needs and developmental demands. Clearly defining the scope of reading outcomes also ensures that comparisons with the broader literature remain meaningful.

Research on reading comprehension provides further justification for selecting specific forms of instructional support. Studies have examined inference instruction,

classroom reading-strategy interventions, and instruction on the structure of expository texts (Elleman, 2017; Okkinga et al., 2018; Pyle et al., 2017). These studies highlight the importance of explicitly defining the content of comprehension instruction and connecting differentiated support with specific reading competencies. Students may receive different forms of assistance in identifying causal relationships, drawing inferences, or organizing information. Nevertheless, instructional activities and assessments should make the targeted abilities and evidence of development explicit. Strengthening reading literacy therefore requires carefully selected instructional support and clearly defined learning objectives rather than relying solely on variations in activity formats as evidence of differentiation.

Against this background, the research need concerns the relationship between diagnostic information, differentiated instruction, and students' achievement patterns across different cognitive processes. Previous studies provide evidence of the potential benefits and varied implementation of differentiated instruction, but they do not directly establish the pattern of improvement across the three AKM cognitive levels in the classroom examined in this study. The research gap is therefore analytical and contextual: a more detailed account is needed to identify which reading abilities improve following instruction and which continue to require attention. This focus extends the interpretation of learning outcomes beyond an overall score toward a more specific achievement profile that can inform instructional priorities. It does not assume that improvement will occur uniformly across all reading processes.

The research setting at SDN 53 Sawerigading in Palopo provides an empirical basis for this investigation. Baseline data within the scope of the 18-item assessment indicated relatively high achievement in locating information, whereas evaluating and reflecting required further instructional support. The questionnaire also identified difficulties involving vocabulary, recalling textual content, and identifying main ideas. Students expressed varied preferences regarding reading materials, peer activities, and the formats through which they demonstrated their understanding. These findings suggest that classroom learning needs cannot be adequately explained by a mean score alone. Instruction should consider students' achievement profiles and supporting diagnostic information, while evaluation should reveal development in the competencies identified at baseline. The resulting diagnostic profile provides a basis for interpreting learning outcomes in relation to students' initial abilities.

Analyzing achievement by cognitive level also makes it possible to examine learning outcomes when students demonstrate high initial achievement. Indicators that students have already largely mastered provide less room for measurable improvement than indicators with lower baseline scores. Consequently, small gains at one cognitive level and larger gains at another should be interpreted in relation to students' starting points. Indicators that remain unchanged or decline also provide important information for subsequent instructional planning. Instructional success should therefore be interpreted by examining patterns of change, their educational significance, and the relevant statistical results. This approach supports targeted recommendations, such as strengthening students' ability to use textual evidence or comprehend information presented in tables. The quality and scope of the assessment also remain important for ensuring that comparisons are based on clearly defined measurements.

This study aims to examine fifth-grade students' reading literacy gains following diagnostic assessment-based differentiated instruction, both overall and across the three AKM cognitive levels. Specifically, it compares students' pretest and posttest achievement,

identifies differences in improvement patterns across cognitive levels, and examines indicators that increase, remain unchanged, or decline. The study is expected to contribute classroom-based evidence that clarifies students' literacy development needs through an analysis of measured achievement. Given its pre-experimental one-group design, the findings describe improvements observed during the instructional period rather than establishing definitive causal effects. Interpretation takes into account students' baseline abilities, measurement characteristics, and differences in scoring scope so that the resulting recommendations remain proportionate to the strength of the available evidence.

Method

Research Design, Setting, and Participants

The study used a quantitative approach with a pre-experimental one-group pretest–posttest design. One group completed an initial assessment, received diagnostic assessment-based differentiated instruction, and completed a final assessment. The same participants were compared to identify changes in achievement, rather than comparing classes with different baseline characteristics. This design corresponds to the available data from two measurement occasions. There was no random assignment or comparison group. Results are therefore interpreted as improvement observed during the study period, while considering other factors that may have accompanied instruction (Creswell & Creswell, 2023).

The study took place in September 2026 at SDN 53 Sawerigading, Palopo, South Sulawesi, Indonesia. The school is a partner in the Teacher Professional Education programme, Pendidikan Profesi Guru (PPG), organized by Universitas Muhammadiyah Palopo as a teacher education institution, Lembaga Pendidikan Tenaga Kependidikan (LPTK). This partnership provides the institutional context of the research site. It does not establish that all data collection formed part of PPG activities, because the administrative relationship between the research activities and the programme has not been documented in detail. Participants comprised one fifth-grade class of 20 students purposively selected as the instructional setting. All participants had pretest records and questionnaire responses. Posttest records were available for 18 students, who formed the sample for paired analyses. Students S06 and S20 had no posttest sheets in the available files. The data readiness document states that these students did not participate in the final assessment, but detailed reasons are unavailable. Participant reporting is based on the availability of measurement records, avoiding assumptions about the reasons for non-participation.

Anonymous codes S01–S20 were used consistently in the test matrix and diagnostic profiles. Data from the two students without posttests were retained to describe the initial assessment but excluded from paired testing. Unavailable scores were marked as missing and were not replaced with zero, the class mean, or estimated values. Diagnostic profiles therefore covered 20 participants, whereas estimates of change covered 18 matched pairs. The relevant denominator is specified in the tables to make percentages and statistical results traceable.

Instruments, Item Mapping, and Scoring

Reading literacy was measured using two test forms designed around the three AKM cognitive levels. Form A was used for the pretest and Form B for the posttest. Each contained 20 items, four stimulus texts, corresponding question formats, and a maximum

raw score of 39. Item pairs were mapped using indicators in the test specification. The corresponding design supports content comparability but is not treated as evidence that the forms' difficulty was empirically equivalent. Differences in texts and response demands remain relevant to interpreting score changes.

Form A included the story "Kapurung untuk Nenek" (Kapurung for Grandmother), the fable "Tedong dan Burung Jalak" (Tedong and the Starling), the explanatory text "Sagu, Tepung dari Batang Pohon" (Sago, Flour from a Tree Trunk), and the text with a table "Jumat Tanpa Plastik" (Plastic-Free Friday). Form B included "Dompet di Pantai Labombo" (A Wallet at Labombo Beach), "Bunga Matahari dan Lebah Kecil" (The Sunflower and the Little Bee), "Kakao, dari Buah Menjadi Cokelat" (Cocoa, from Fruit to Chocolate), and "Jumat Gemar Membaca" (Friday Reading Activities). The first two stimuli in each form were literary texts, and the remaining two were informational texts. Paired items assessed corresponding indicators despite differences in stimulus content.

Table 1. Item distribution and maximum raw scores under two scoring scopes

Cognitive level	20-item scope: items	Maximum score	18-item scope: items	Maximum score
Locating information	6	6	5	5
Interpreting and integrating	9	18	9	18
Evaluating and reflecting	5	15	4	12
Total	20	39	18	35

Multiple-choice and short-answer items were scored 0–1. Complex multiple-choice, matching, and true–false items were scored 0–2 according to the answer key and partial-credit rubric. For complex multiple-choice items, maximum credit required the correct selections specified in the key, with partial credit determined by the rubric. Matching and true–false scores reflected the accuracy of pairs or judgments. Blank responses received zero. Multiple marks, unclear entries, and ambiguous responses were recorded as distinct data statuses so that their scoring decisions were not conflated with missing test sheets. Constructed responses were scored 0–3 using task-specific rubrics. Assessment components included response accuracy, reasoning, evidence, and connections with experience where required. For the moral-message task, the maximum score required an appropriate message supported by story evidence. For a proposal based on tabular information, it required a concrete, relevant proposal supported by data. Because task demands differed, score interpretations followed the respective rubrics. Aggregate scores were not used to conclude that students had fulfilled every response component across all indicators.

Documented scoring was undertaken by one scorer. Complete content validation results, a second independent scoring procedure, and empirical form-equivalence testing were unavailable. Aiken's V and inter-rater agreement coefficients are therefore not reported. Principles of content validity review inform the discussion of instrument evaluation needs, without implying that an index was calculated or expert review was conducted (Almanasreh et al., 2019). The original item-level records for the two items excluded from one scoring scope were retained so that the consequences of the scoring decision could be examined. The diagnostic questionnaire provided information about preferred reading materials, ways of undertaking reading activities, preferences for working individually or with peers, preferred products, and perceived difficulties. It was used descriptively alongside the initial test. Preferences were not treated as fixed abilities or evidence of particular learning styles. Academic needs were considered in relation to

cognitive-level scores and task objectives. No follow-up questionnaire was used to infer changes in interests or preferences during instruction.

Assessment Procedures and Differentiated Instruction

Form A and the questionnaire were administered on 21 September 2026. The test mapped baseline achievement, while the questionnaire provided supporting information from students' perspectives. Results were organized into individual profiles connecting participant codes, pretest scores, subscores, and relevant responses. These profiles allowed differences in needs to be examined without relying on one total score. In interpretation, students' reported difficulties were distinguished from abilities demonstrated in test tasks. Differentiated instruction was reported to have taken place between the pretest and posttest. Recorded forms included adjustments to text difficulty and visual support for content; guided reading, paired or group discussion, and independent reading for process; and options for oral storytelling, story maps, summaries, or written responses for product. Reading objectives provided a common instructional direction, while forms of support were reported as varied. This description reflects activities recorded in the research documentation and does not establish an instructional dosage that has not been documented.

Complete information on the number of meetings, session dates, actual duration, instructors, attendance, group allocation, and individual products was unavailable. The seven-day interval is not interpreted as evidence of seven sessions or daily instruction. Questionnaire preferences are also not treated as proof that each student received the selected activity format. The study does not test relationships between exposure to specific support and individual improvement because the necessary data are unavailable. Form B was administered on 28 September 2026 and matched with the pretest using participant codes. A planned testing time of 70 minutes was specified, but actual duration was not documented. The two forms were administered on different occasions; supervision and assistance during testing could not be established in detail from the files. No delayed assessment was available to evaluate retention. Analysis consequently focused on differences in achievement between the two available measurement occasions.

Data Management and Analytical Strategy

Data checks covered participant codes, score ranges, response statuses, subscore sums, and consistency of totals. Blank responses, ambiguous marks, and unavailable sheets were treated as different statuses. Reanalysis used Python 3.12.14, NumPy 2.3.5, and SciPy 1.17.0. Calculations used raw scores before rounding. Table values were rounded after statistics were calculated, ensuring that rounding individual scores did not affect category assignments or test results. Two scoring scopes were presented alongside each other. The full instrument comprised 20 items with a maximum raw score of 39. The 18-item scope excluded items 12 and 15 from both forms and used a maximum of 35. The documented reasons for exclusion concerned time units and interpretation of the pretest instructions. Because these reasons do not sufficiently establish item invalidity, both sets of results were retained. The scoring scope was not selected according to which produced greater improvement or a larger effect size. The comparison examined how estimates depended on the choice of included items.

Descriptive statistics included means, medians, standard deviations, ranges, and differences. Questionnaire profiles were reported as frequencies and percentages of 20 responses. Overall changes in both scopes were examined with paired-samples t-tests at

the 0.05 significance level and reported with 95% confidence intervals. Within the 18-item scope, the Shapiro–Wilk test examined paired score differences, and the Wilcoxon signed-rank test provided a complementary analysis. Distributional assessment focused on within-pair differences rather than using separate pretest and posttest normality checks as the sole basis for test selection (Mishra et al., 2019).

Cohen's dz expresses mean change relative to the standard deviation of paired differences. It was interpreted as standardized within-group change, rather than an effect relative to a control group. Interpretation considered context and data variability, as emphasized by Funder and Ozer (2019) and Schäfer and Schwarz (2019). Individual normalized gain (N-gain) describes change relative to the remaining room for improvement and was summarized by its mean. Operational categories were low for values < 0.30, moderate for values ≥ 0.30 and < 0.70, and high for values ≥ 0.70. The methodological discussion by Coletta and Steinert (2020) was considered while recognizing that their measurement context involved physics concept inventories.

Cognitive-level and indicator analyses used the 18-item scope. Achievement was expressed relative to the maximum for each level, while contributions to total score change used the overall maximum of 35. Level-specific normalized gains were calculated from group means and distinguished from the mean of individual normalized gains. Normalized gain was not calculated when baseline achievement was 100%. Cognitive-level Wilcoxon tests were exploratory and were not adjusted for multiple testing. Changes in selected indicators were described without separate hypothesis tests for each item. This reporting avoids treating numerous indicators as independent demonstrations of instructional effectiveness.

Identity protection and ethical context

The research took place at a partner school for the PPG programme delivered by Universitas Muhammadiyah Palopo in its capacity as an LPTK. Data were presented using anonymous codes, without student names or personal identifiers in the tables. Main findings were reported in aggregate, while codes maintained correspondence between paired measurements. Coding supported analytical traceability without disclosing participant identities in the article. The status of research permission, parental or guardian consent, student assent, and ethical review could not be established from the available documentation. The school partnership describes an institutional relationship and is not treated as ethics committee approval.

Results and Discussion

Results

Data Completeness and Diagnostic Profiles

All 20 students had pretest records and diagnostic questionnaire responses. Posttest records were available for 18 students, or 90% of the original participants. Baseline profiles used all responses, whereas measurement comparisons used complete pairs. The two students without posttests each had a pretest score of 74.29 in the 18-item scope, equal to the median among students included in the analysis. This similarity indicates their baseline position but does not explain missing posttests or establish the absence of missing-data bias.

Table 2. Diagnostic Profiles of 20 Students Before Instruction

Aspect	Choice or Response	N	%
Reading materials	Stories or fairy tales	7	35
	Informational reading	4	20
	Comics	7	35
	News	2	10
Work arrangement	Individual	4	20
	Pairs	7	35
	Groups	9	45
Product	Oral storytelling	6	30
	Summary	4	20
	Drawing or story map	10	50
Difficulty	Vocabulary	5	25
	Recall of content	5	25
	Main ideas	5	25
	None reported	5	25

Stories or fairy tales and comics were the most frequently preferred reading materials, each selected by seven students. Four preferred informational reading, and two preferred news. Sixteen selected paired or group activities, while four preferred working alone. Half preferred drawings or story maps as products. These choices describe baseline responses and do not demonstrate participation in particular activities or receipt of the preferred product options during instruction. Reported difficulties concerned vocabulary, recall of content, and identification of main ideas, with five students in each category. Five reported no difficulty. Reporting no difficulty was not equated with maximum test mastery. Within the 18-item scope, all initial participants achieved the maximum in locating information, whereas their evaluating and reflecting subscores were uniformly 6 out of 12. The questionnaire therefore provided variation in information that was not visible in the uniform baseline evaluation subscores.

Overall Achievement Under Two Scoring Scopes

Table 3. Descriptive Score Statistics Under Two Scoring Scopes (n = 18)

Scope	Occasion	Mean	Median	SD	Min.–max.
20 items	Pretest	69.09	69.23	2.24	61.54–71.79
20 items	Posttest	84.47	84.62	3.11	79.49–89.74
18 items	Pretest	74.13	74.29	2.49	65.71–77.14
18 items	Posttest	82.86	82.86	3.53	77.14–88.57

Across all 20 items, the mean score increased from 69.09 to 84.47. In the 18-item scope, it increased from 74.13 to 82.86. All 18 participants had positive differences under both scopes. Table 3 presents descriptive results, and Table 4 compares estimates and statistical tests.

Table 4. Comparison of paired analyses

Measure	20 items	18 items
Mean difference (points)	15.38	8.73
SD of differences	2.92	3.32
95% CI for mean difference	13.93–16.84	7.08–10.38
t (df = 17)	22.38	11.16
p, paired-samples t-test	< 0.001	< 0.001
Cohen's dz	5.27	2.63
Mean individual normalized gain	0.498	0.337
Students with increased scores	18	18

For 20 items, the mean increase was 15.38 points, with a 95% confidence interval of 13.93–16.84. For 18 items, it was 8.73 points, with an interval of 7.08–10.38. Both paired-

samples t-tests were significant at $p < 0.001$. In the 18-item scope, the Shapiro–Wilk test of paired differences yielded p approximately 0.051, while the Wilcoxon test also yielded $p < 0.001$. The complementary analysis was consistent with the direction and conclusion of the paired-samples test for that scope. Mean individual normalized gains were 0.498 for 20 items and 0.337 for 18 items. Both were moderate, although individual distributions differed. All participants were in the moderate category for 20 items. For 18 items, six were low and 12 were moderate. No participant was in the high category under either scope. Categories were determined from raw scores; student S10 had an exact normalized gain of 0.30 in the 18-item scope and was therefore classified as moderate.

Table 5. Distribution of individual normalized-gain categories (n = 18)

Category	20 items: n	20 items: %	18 items: n	18 items: %
Low (< 0.30)	0	0.00	6	33.33
Moderate (0.30–< 0.70)	18	100.00	12	66.67
High (≥ 0.70)	0	0.00	0	0.00

Table 5 presents the distribution of individual normalized-gain categories among 18 students based on two assessment scopes, consisting of 20 items and 18 items. In the 20-item assessment, all 18 students (100.00%) achieved a moderate normalized-gain category, with no students classified in the low or high categories. In contrast, the results based on the 18-item assessment showed that 12 students (66.67%) achieved moderate normalized gain, while 6 students (33.33%) were classified in the low category. No students achieved a high normalized-gain category in either assessment scope. These findings indicate that the distribution of individual normalized-gain categories differs depending on the number of assessment items included in the analysis. The 20-item assessment showed a consistent moderate-gain classification across all students, whereas the 18-item assessment revealed greater variation in individual learning gains. This difference highlights the importance of clearly specifying the assessment scope when interpreting normalized-gain results and reporting students' reading literacy improvement.

Achievement across the three cognitive levels in the 18-item scope

Cognitive-level analysis showed uneven development. Locating information began at maximum achievement, interpreting and integrating at 83.02%, and evaluating and reflecting at 50.00%. Each percentage used its level-specific maximum and was not summed with other level percentages to calculate the total score.

Table 6. Achievement by cognitive level within the 18-item scope

Cognitive level	Pretest (%)	Posttest (%)	Difference (pp)	Wilcoxon p
Locating information	100.00	96.67	−3.33	0.180
Interpreting and integrating	83.02	85.80	2.78	0.059
Evaluating and reflecting	50.00	72.69	22.69	< 0.001

For locating information, 16 participants maintained their initial scores and two declined. The mean subscore decreased from 5.00 to 4.83 out of 5. Final achievement remained high at 96.67%, and the difference was not statistically significant. Normalized gain was not calculated because all participants had already reached the maximum at baseline within this scope. This level showed no improvement despite the increase in the overall score. Interpreting and integrating increased from a mean subscore of 14.94 to 15.44 out of 18. Eleven participants improved, three remained unchanged, and four declined. The increase of 2.78 percentage points corresponded to a group normalized gain of 0.164. The Wilcoxon test yielded $p = 0.059$, which did not meet the 0.05 criterion.

Differences in individual directions indicate that the mean increase was not experienced by every participant at this level.

Evaluating and reflecting increased from a mean subscore of 6.00 to 8.72 out of 12. Seventeen participants improved and one remained unchanged; none declined. The increase of 22.69 percentage points corresponded to a group normalized gain of 0.454, with $p < 0.001$. On the overall scale based on a maximum raw score of 35, this level contributed +7.78 points, interpreting and integrating +1.43 points, and locating information −0.48 points. These contributions sum to the total increase of 8.73 points, demonstrating the predominance of evaluating and reflecting on a common contribution scale.

Changes in selected indicators and score patterns

Indicator analysis examined evaluation, reflection, and tabular-reading tasks. Each item percentage was calculated relative to its maximum score, rather than as the percentage of students who fully mastered the indicator. This distinction matters because some tasks awarded partial credit. For example, achievement of 66.67% could result from every participant receiving 2 out of 3, rather than two-thirds providing fully correct responses.

Table 7. Changes in selected indicators within the 18-item scope

Item	Indicator	Pretest	Posttest	Difference
5	Evaluating a character's decision	33.33 (%)	70.37 (%)	37.04 (pp)
9	Explaining a moral message using story evidence	33.33 (%)	90.74 (%)	57.41 (pp)
10	Reflecting on experience and formulating a personal stance	66.67 (%)	62.96 (%)	−3.70 (pp)
17	Checking statements against tabular information	50.00 (%)	50.00 (%)	0.00 (pp)
20	Proposing relevant actions based on table data	66.67 (%)	66.67 (%)	0.00 (pp)

The largest increase occurred in explaining a moral message using story evidence, at 57.41 percentage points. Sixteen participants improved and two maintained their scores. On the posttest, 15 received the maximum score of 3. Evaluating a character's decision improved for every participant: the initial score of 1 changed to 2 for 16 students and to 3 for two. Despite this improvement, most responses remained below the maximum. Reflection on experience and formulation of a personal stance declined slightly. Sixteen participants maintained a score of 2 and two declined to 1. For checking statements against tabular information, all maintained 1 out of 2. For proposing actions, all maintained 2 out of 3. According to the rubric, a proposal scored 2 when it was concrete and relevant but lacked supporting table data. Aggregate improvement in evaluation therefore did not extend to all of its constituent indicators. At pretest, 13 of the 18 participants had identical score patterns across all 20 items. Several constructed-response items also had the same score for every participant. This is uniformity in the score matrix; the data do not establish that the wording of written responses was identical. Uniformity is consequently reported as a measurement characteristic without attributing it to assistance, copying, or another unverified process.

Dependence of estimates on item inclusion

Comparison of scoring scopes showed consistent improvement direction but different magnitudes. On item 12, all participants had a pretest score of 0, while 17 had a posttest score of 1. On item 15, all had a pretest score of 1 and a posttest score of 3. Including these items added positive differences while changing the maximum raw score from 35 to 39. Because the denominator changed, scores from the two scopes represent

different transformations of raw scores and do not describe two participant groups. The change in the maximum also affected baseline scores and the room for improvement used in normalized gain. Mean increases of 15.38 and 8.73 are therefore not interchangeable in the narrative. Each result remains linked to its scoring scope. The cognitive-level analysis presented above applies to 18 items and is not automatically extended to 20 items. The analysis producing the larger effect was not considered more appropriate on that basis. Consistent direction is the finding that can be maintained, while magnitude remains dependent on instrument decisions. This dependence is also apparent in normalized-gain distributions and dz. Both scoring scopes are presented so that readers can directly assess the consequences of item exclusion.

Discussion

Meaning of overall improvement and comparison of outcome measures

The study showed score increases among all participants with complete data under both the full instrument and the 18-item assessment scope. This consistency supports the conclusion that final measurement scores were higher than initial scores. The confidence intervals also indicated positive mean differences under both assessment scopes. However, the substantial difference between the estimates requires careful consideration of how the outcomes are defined and calculated. Improvement describes changes in measurement results within the study group and does not directly establish the magnitude of an instructional effect relative to a comparison condition. This finding is directionally consistent with evidence of positive outcomes associated with differentiated literacy instruction reported (Puzio et al. 2020). Their meta-analysis supports an association between differentiated instruction and improved literacy achievement under the research conditions examined (Puzio et al., 2020). Nevertheless, comparisons across studies must account for differences in participants, instructional duration, teacher support, and outcome measures. The present study involved one class and two closely spaced assessment occasions. Therefore, the findings provide evidence about a specific classroom context but do not replace effect estimates derived from studies involving comparison groups or longer instructional periods.

The Cohen's \(d_z\) values of 2.63 for the 18-item assessment and 5.27 for the 20-item assessment cannot be directly compared with Hedges' \(g\) reported in the meta-analysis by Puzio et al. (2020). Cohen's \(d_z\) is calculated using the standard deviation of within-participant differences, meaning that large values can occur when improvements are relatively consistent and the standard deviation of the differences is small relative to their mean. Consequently, these large effect-size estimates do not establish that the instructional intervention was superior to other approaches. The importance of interpreting effect sizes in context is emphasized by Funder and Ozer (2019), while Schäfer and Schwarz (2019) discuss the variation and interpretation of effect-size estimates across research contexts. Normalized gain complements estimates of score change by considering the available room for improvement. Both assessment scopes produced moderate mean normalized gains, although the distributions of individual gains differed. Under the 18-item assessment scope, six participants were classified as having low normalized gains, whereas all participants were classified as having moderate normalized gains under the 20-item assessment scope. This difference demonstrates that normalized-gain classifications depend on baseline scores and the items included in the assessment. Coletta and Steinert (2020) provide a relevant methodological basis for understanding normalized gain as a measure of learning

improvement; however, the categories in the present study are used operationally and should not be interpreted as validated levels of reading-literacy progress without further supporting evidence.

In practical terms, statistical differences must be connected to the specific abilities that have changed. Overall scores may increase substantially when performance on several tasks shifts from low to high achievement, even when other abilities remain unchanged. Disaggregated assessment results therefore help teachers identify areas that require further instructional support. Accordingly, this study uses overall scores as an initial indication of improvement and subsequently examines cognitive levels and individual indicators to explain the educational significance of the observed changes. This approach reduces the risk of equating higher overall test performance with mastery of every aspect of reading comprehension.

Differences in the development of cognitive reading processes

The 18-item analysis identified evaluating and reflecting as the largest contributor to overall score improvement, with a contribution of +7.78 points on the overall score scale rather than a comparison based solely on percentages with different denominators. Interpreting and integrating contributed +1.43 points, whereas locating information contributed −0.48 points. These results indicate that the instructional period was followed by a pattern of change concentrated in particular cognitive tasks. This pattern helps explain how the overall score could increase despite a slight decline in locating information. For this cognitive level, maximum baseline scores limited students' opportunities to demonstrate further progress. However, the absence of an observed increase does not necessarily indicate an absence of learning experiences, as the assessment instrument provided limited room for additional score improvement. Nevertheless, coverage of five items alone does not establish mastery of all information-retrieval abilities. Future assessments should vary task demands, text length, and information locations to better distinguish students who have already achieved high performance levels. This recommendation concerns the quality of measurement and does not imply that previously obtained scores should be altered.

Interpreting and integrating demonstrated less consistent changes, with eleven students improving and four students experiencing declines, resulting in only a small increase in the mean score. The p-value of 0.059 indicates that the result did not meet the specified statistical significance criterion; however, it does not prove that no individual student developed this ability. Research on inference instruction highlights the importance of helping students connect textual cues with relevant background knowledge (Elleman, 2017). Similarly, research examining reading strategies emphasizes the need to specify the content and focus of comprehension instruction (Okkinga et al., 2018). These findings provide a theoretical basis for instructional recommendations but do not establish that any particular strategy was implemented with the participants in this study. Meanwhile, evaluating and reflecting began at a lower level, providing greater room for improvement.

Higher scores on explanations of moral messages and judgments of characters' decisions indicate improved responses in relation to the assessment rubric's requirements. However, such progress does not automatically demonstrate proficiency in all forms of text evaluation, including assessing credibility or using evidence across multiple sources. The Asesmen Kompetensi Minimum (AKM) framework provides a basis for mapping assessment indicators, while interpretations of students' abilities must

remain limited to the tasks actually measured (Pusat Asesmen dan Pembelajaran, 2020). Reading comprehension should be understood as involving interconnected abilities rather than a single, uniform skill (Duke et al., 2021; Duke & Cartwright, 2021; Elleman & Oslund, 2019). Accordingly, comprehension instruction needs to address multiple processes, as improvement in one component does not eliminate the learning needs associated with another. For this class, factual retrieval activities can be maintained, while additional reasoning opportunities should target cognitive components that remain comparatively weaker. Using evidence from specific assessment indicators to establish instructional priorities provides a more targeted approach than simply increasing the overall intensity of reading activities.

Diagnostic assessment as a basis for differentiation decisions

The questionnaire provided more varied information than the initial evaluation subscores. Five participants reported vocabulary difficulties, five difficulties recalling content, and five difficulties identifying main ideas. Most preferred activities with peers, and half selected story maps. This information provides a basis for offering choices and assistance, but preferences alone cannot establish academic needs. A stronger diagnosis combines students' perceptions with what they demonstrate in reading tasks. Provide a basis for using achievement data to adjust interventions (Filderman et al., 2018). In this study, baseline data explain a pedagogical rationale for considering different support, but documentation is insufficient to examine the relationship between actual support and improvement. Design, implementation, and outcomes need to be distinguished. Questionnaire information provides a basis for planning, whereas test increases describe measured outcomes. The mechanism connecting them cannot be established without linked activity records or student work.

Situate assessment and content, process, and product differentiation within reading instruction (Achmad et al., 2024). Demonstrate variation in teachers' practices and understanding, while show the complexity of the teaching skills involved (Gheyssens et al., 2022 ; Van Geel et al. 2019). Together, these sources support accompanying the term differentiation with concrete decisions. Here, reported formats follow the records, and unverified session counts or individual allocations are not added. This keeps the explanation of the approach proportionate to the implementation evidence. Diagnostic recommendations can include contextual vocabulary support for students reporting word difficulties, information organization for those struggling with recall, and practice distinguishing main and supporting ideas for those reporting main-idea difficulties. Support can be reviewed using task responses rather than fixed labels. Grouping can also change according to the competency being addressed. Provide context for understanding differentiation within programmes and implementation conditions; group allocation alone does not explain the substance of support (Deunk et al., 2018).

Written products can clarify the relationship between understanding and reasoning. Examined programmes balancing reading and writing, providing a basis for considering summaries or written responses within literacy activities (Graham et al, 2018). Nevertheless, oral responses and story maps should also require evidence when the objective calls for it. Product choice should not lower the target competency. Assessment can distinguish content understanding from presentation quality so that product options do not introduce irrelevant differences in standards.

Use of evidence and comprehension of tabular information

The moral-message indicator showed the greatest increase, while judgment of a character's decision improved with most scores remaining below the maximum. These findings draw attention to the quality of response components. A student may provide an appropriate response without fulfilling the required combination of reasoning, specific evidence, and connections with experience. Interpretation needs to consider the rubric so that score increases are not assigned a meaning broader than the task warrants. Partially complete responses provide a basis for identifying components requiring further support. Two tabular-reading indicators showed no progress. Unchanged scores for checking statements suggest that the accuracy of connecting statements with data did not change in the assessed task. For proposed actions, a score of 2 indicated a relevant proposal without numerical support according to the rubric. This differs from an inability to generate ideas: students could propose actions but did not demonstrate the required data basis. This distinction matters because follow-up support should target evidence use rather than merely increasing the number of proposals.

Support can begin with examining a table's title, categories, units, and values, followed by matching each statement with the relevant row or column. When developing a justification, students can be asked to identify data, explain their meaning, and connect them with an action. This sequence is a recommendation derived from unchanged indicators. Provide context for teaching informational text structures, while support attention to strategies and background knowledge for struggling readers (Filderman et al., 2022; Pyle et al., 2017). Neither is treated as direct evidence that a specific table-reading exercise was effective in this study. Feedback can focus on components not yet fulfilled. Rather than providing only a score, teachers can explain that a proposal is relevant but lacks numerical evidence, then request a revision linking the data with the choice. Provide a basis for considering the role of feedback in learning. The present study did not measure the form or effect of feedback during instruction, so this recommendation is not treated as an established mechanism (Wisniewski et al., 2020). Further assessment should examine the quality of reasoning in new tasks rather than repetition of the same answer.

For students requiring support, prompting questions can focus on essential information. More prepared students can compare several alternatives using a table. The shared objective is to use information accurately to develop a reasoned response. Activities can combine literary and informational texts so that students learn that narrative and numerical evidence have different forms but must both be relevant to a response. Achievement should be checked through individual responses so that improvement is not inferred solely from group products.

Robustness of findings, measurement quality, and scope of contribution

Both scoring scopes produced positive gains, but their magnitude was sensitive to item exclusion. Consistent direction supports the descriptive conclusion of score improvement. However, consistency does not establish that all instrument decisions have been validated. Patterns on items 12 and 15 need to be considered alongside differences in stimuli and rubrics. Uniform pretest errors may reflect mastery or task demands, while posttest success may reflect development or form characteristics. Available data do not separate these possibilities. Emphasize procedures for assessing content validity. Here, the test specification and rubrics provide a basis for examining coverage but do not replace unavailable expert assessment results (Almanasreh et al., 2019). Equal item

counts and maximum scores also do not guarantee equal difficulty. The study therefore describes two forms with corresponding structures rather than empirically equivalent parallel forms. This distinction prevents the basis of comparison from being stated more strongly than the measurement evidence permits.

Uniform pretest score patterns add a reason for caution. Thirteen participants had identical patterns, but the matrix does not explain whether their written responses differed while receiving equivalent scores or whether recording issues were involved. Scoring by one person provides no estimate of independent agreement. Because original responses were unavailable for additional examination during manuscript preparation, interpretation follows the available matrix and rubrics. No claim is made about the cause of uniformity or uncalculated reliability estimates. Transparent reporting of data characteristics allows readers to assess the evidence. The seven-day interval describes short-term achievement. No delayed measurement establishes retention or transfer to other texts. Two posttests were unavailable, and detailed reasons were not documented. Baseline scores matching the median show only initial achievement positions, not that missingness was unrelated to outcomes. The single classroom also restricts contextual coverage. Findings are not generalized as estimates for all primary school students or all forms of differentiated instruction. Further studies with comparison groups are needed to strengthen attribution of improvement to instruction.

The study contributes an ability profile and transparent analytical decisions. Higher total scores occurred alongside unchanged indicators, small declines, and differences between scoring scopes. This information supports instructional priorities more specific than a single success category. The PPG partnership establishes an institutional relationship with the school but does not extend claims about implementation or ethics. The article provides classroom evidence of observed change, with recommendations directed toward evidence use, information integration, and improved documentation of instructional evaluation.

Conclusion

Reading literacy scores increased among the 18 fifth-grade students with complete pretest and posttest data following diagnostic assessment-based differentiated instruction. Improvement was found under both scoring scopes. Across all 20 items, mean scores increased from 69.09 to 84.47, with a mean individual normalized gain of 0.498. Within the 18-item scope, scores increased from 74.13 to 82.86, with a normalized gain of 0.337. Both mean gains were classified as moderate. Analysis of the three cognitive levels within the 18-item scope showed the greatest improvement in evaluating and reflecting, particularly explaining moral messages using story evidence and judging characters' decisions. Interpreting and integrating improved slightly, while locating information remained high with a limited decline. Two tabular-reading indicators did not improve. The study contributes an achievement profile that clarifies priorities for strengthening literacy. Teachers are encouraged to use diagnostic information to design targeted support, especially for integrating information and using evidence in responses. Consistent improvement direction across scoring scopes does not remove the dependence of magnitude on scoring decisions. Given the one-group design, untested form equivalence, and limited process documentation, the findings do not establish causality. Further research should strengthen measurement quality, implementation documentation, comparison conditions, and assessment of sustained achievement.

References

Achmad, W. K. S., Rachman, S. A., Aras, L., & Amran, M. (2024). Differentiated instruction in reading in elementary schools: A systematic review. *International Journal of Evaluation and Research in Education,* 13(3), 1997–2005. https://doi.org/10.11591/ijere.v13i3.27134

Almanasreh, E., Moles, R., & Chen, T. F. (2019). Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy*, 15(2), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

Anggraena, Y., Ginanto, D., Felicia, N., Andiarti, A., Herutami, I., Alhapip, L., Iswoyo, S., Hartini, Y., & Mahardika, R. L. (2022). Learning and Assessment Guidelines: Early Childhood Education, Primary and Secondary Education. *Education Standards, Curriculum, and Assessment Agency*..

Coletta, V. P., & Steinert, J. J. (2020). Why normalized gain should continue to be used in analyzing preinstruction and postinstruction scores on concept inventories. *Physical Review Physics Education Research,* 16(1), 010108. https://doi.org/10.1103/PhysRevPhysEducRes.16.010108

Creswell, J. W., & Creswell, J. D. (2023). *Research design: Qualitative, quantitative, and mixed methods approaches* (6th ed.). SAGE.

Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation practices in primary education. *Educational Research Review*, 24, 31–54. https://doi.org/10.1016/j.edurev.2018.02.002

Duke, N. K., & Cartwright, K. B. (2021). The science of reading progresses: Communicating advances beyond the simple view of reading. *Reading Research Quarterly*, 56(S1), S25–S44. https://doi.org/10.1002/rrq.411

Duke, N. K., Ward, A. E., & Pearson, P. D. (2021). The science of reading comprehension instruction. *The Reading Teacher,* 74(6), 663–672. https://doi.org/10.1002/trtr.1993

Elleman, A. M. (2017). Examining the impact of inference instruction on the literal and inferential comprehension of skilled and less skilled readers: A meta-analytic review. *Journal of Educational Psychology,* 109(6), 761–781. https://doi.org/10.1037/edu0000180

Elleman, A. M., & Oslund, E. L. (2019). Reading comprehension research: Implications for practice and policy. *Policy Insights from the Behavioral and Brain Sciences,* 6(1), 3–11. https://doi.org/10.1177/2372732218816339

Filderman, M. J., Austin, C. R., Boucher, A. N., O'Donnell, K., & Swanson, E. A. (2022). A meta-analysis of the effects of reading comprehension interventions on the reading comprehension outcomes of struggling readers in third through 12th grades. *Exceptional Children,* 88(2), 163–184. https://doi.org/10.1177/00144029211050860

Filderman, M. J., Toste, J. R., Didion, L. A., Peng, P., & Clemens, N. H. (2018). Data-based decision making in reading interventions: A synthesis and meta-analysis of the effects for struggling readers. *The Journal of Special Education,* 52(3), 174–187. https://doi.org/10.1177/0022466918790001

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, 2(2), 156–168. https://doi.org/10.1177/2515245919847202

Gheyssens, E., Coubergs, C., Griful-Freixenet, J., Engels, N., & Struyven, K. (2022). Differentiated instruction: The diversity of teachers' philosophy and praxis to adapt teaching to students' interests, readiness and learning profiles. *International Journal of Inclusive Education,* 26(14), 1383–1400. https://doi.org/10.1080/13603116.2020.1812739

Graham, S., Liu, X., Aitken, A., Ng, C., Bartlett, B., Harris, K. R., & Holzapfel, J. (2018). Effectiveness of literacy programs balancing reading and writing instruction: A meta-analysis. *Reading Research Quarterly,* 53(3), 279–304. https://doi.org/10.1002/rrq.194

Mishra, P., Pandey, C. M., Singh, U., Gupta, A., Sahu, C., & Keshri, A. (2019). Descriptive statistics and normality tests for statistical data. *Annals of Cardiac Anaesthesia*, 22(1), 67–72. https://doi.org/10.4103/aca.ACA_157_18

OECD. (2023). PISA 2022 results (Volume I): The state of learning and equity in education. *OECD Publishing.* https://doi.org/10.1787/53f23881-en

Okkinga, M., van Steensel, R., van Gelderen, A. J. S., van Schooten, E., Sleegers, P. J. C., & Arends, L. R. (2018). Effectiveness of reading-strategy interventions in whole classrooms: A meta-analysis. *Educational Psychology Review*, 30(4), 1215–1239. https://doi.org/10.1007/s10648-018-9445-7

Pusat Asesmen dan Pembelajaran. (2020). *Minimum Competency Assessment (AKM) and Its Implications for Learning.* Center for Assessment and Learning, Research and Development and Book Agency, Ministry of Education and Culture.

Puzio, K., Colby, G. T., & Algeo-Nichols, D. (2020). Differentiated literacy instruction: Boondoggle or best practice? Review of Educational Research, 90(4), 459–498. https://doi.org/10.3102/0034654320933536

Pyle, N., Vasquez, A. C., Lignugaris/Kraft, B., Gillam, S. L., Reutzel, D. R., Olszewski, A., Segura, H., Hartzheim, D., Laing, W., & Pyle, D. (2017). Effects of expository text structure interventions on comprehension: A meta-analysis. *Reading Research Quarterly*, 52(4), 469–501. https://doi.org/10.1002/rrq.179

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between sub-disciplines and the impact of potential biases. *Frontiers in Psychology, 10*, 813. https://doi.org/10.3389/fpsyg.2019.00813

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology,* 10, 2366. https://doi.org/10.3389/fpsyg.2019.02366

Van Geel, M., Keuning, T., Frèrejean, J., Dolmans, D., van Merriënboer, J., & Visscher, A. J. (2019). Capturing the complexity of differentiated instruction. *School Effectiveness and School Improvement,* 30(1), 51–67. https://doi.org/10.1080/09243453.2018.1539013

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. https://doi.org/10.3389/fpsyg.2019.03087