



DEVELOPING HOTS-BASED READING ASSESSMENT FOR SENIOR HIGH SCHOOL STUDENTS: VALIDITY, RELIABILITY, AND CLASSROOM PRACTICALITY EVIDENCE

Ilfi Nur Diana^{1*}

Universitas PGRI Kanjuruhan Malang, Malang, Indonesia

ilfi.nurdiana@unikama.ac.id

*Corresponding author: ilfi.nurdiana@unikama.ac.id

Article History	Received: October 19, 2024	Revised: December 13, 2024	Accepted: January 21, 2025	Published: February 15, 2025
------------------------	----------------------------	----------------------------	----------------------------	------------------------------

Abstract

Higher-order thinking skills-based reading assessment is needed in senior high schools because reading assessment should not only measure literal comprehension but also students' ability to analyze, evaluate, infer, synthesize, and reflect on texts. However, many classroom reading tests still emphasize lower-order questions, such as identifying explicit information, vocabulary meaning, and factual recall. This study aims to develop a higher-order thinking skills-based reading assessment for senior high school students and examine its validity, reliability, item quality, and practicality. The study employed a research and development design adapted from the ADDIE model, consisting of analysis, design, development, implementation, and evaluation. The participants included 186 eleventh-grade students, three English teachers, two assessment experts, two English language education experts, and one curriculum expert. Data were collected through needs analysis questionnaires, expert validation sheets, reading assessment instruments, student response questionnaires, teacher practicality questionnaires, and item analysis documentation. Quantitative data were analyzed using descriptive statistics, Aiken's V, item difficulty, discrimination index, distractor effectiveness, and KR-20 reliability. Qualitative data from expert comments and teacher feedback were analyzed thematically. The findings showed that the developed assessment was highly valid, reliable, and practical. Of 45 initial items, 36 items were retained after expert validation and empirical testing. The study concludes that HOTS-based reading assessment can provide a more meaningful measure of students' critical reading ability when developed through systematic blueprinting, expert validation, empirical item analysis, and classroom-based practicality evaluation.

Keywords: HOTS assessment; reading literacy; test development; senior high school; item analysis

A. Introduction

Reading assessment plays a central role in English language learning because it provides evidence of students' ability to understand, interpret, evaluate, and use written texts. In senior high schools, reading assessment should measure more than students' ability to locate explicit information. Students are expected to read various text types, identify implied meaning, evaluate arguments, compare perspectives, infer author intention, judge textual evidence, and connect text content with broader contexts. These abilities are closely related to higher-order thinking skills.

Higher-order thinking skills, commonly referred to as HOTS, involve cognitive processes beyond remembering and understanding. In the Revised Bloom's Taxonomy, higher cognitive processes include analyzing, evaluating, and creating (Anderson & Krathwohl, 2001). In reading, these processes appear when students examine relationships among ideas, distinguish facts from opinions, evaluate the strength of arguments, infer unstated meanings, compare information across paragraphs, and construct responses based on textual evidence. Therefore, HOTS-based reading assessment should be designed to measure deep comprehension rather than surface recognition.

The construct has a long pedigree. Resnick (1987) characterised higher-order thinking as non-algorithmic, complex, and involving uncertainty, and Anderson and Krathwohl (2001) provided the revised taxonomy that most classroom applications now use. Zohar and Dori (2003) showed that low-achieving students also benefit from higher-order tasks rather than being restricted to recall, and Miri, David, and Uri (2007) found that critical thinking improves when it is taught purposely rather than left to develop incidentally. Brookhart (2010) translates these findings into concrete guidance for classroom assessment, and Nitko and Brookhart (2011) set the same guidance within a general framework of educational assessment.

However, classroom reading assessment in many English as a foreign language contexts still tends to emphasize lower-order thinking skills. Reading questions often ask students to identify main ideas, locate explicit details, match vocabulary, and choose factual information. These questions are useful for basic comprehension, but they are insufficient for assessing critical reading. If assessment remains dominated by lower-level questions, teachers may not obtain accurate information about students' ability to analyze and evaluate texts.

Previous studies on reading questions in English textbooks and classroom tests have also shown that lower-order questions are often

more dominant than higher-order questions. This condition creates a gap between curriculum expectations and classroom assessment practice. Teachers may be encouraged to develop students' critical thinking, but the tests used in class may not represent critical reading demands. As a result, students are not sufficiently challenged to interpret, evaluate, and justify their understanding of texts.

A valid HOTS-based reading assessment should also align with text types, learning objectives, cognitive levels, and students' developmental readiness. Senior high school students commonly encounter narrative, analytical exposition, explanation, report, discussion, news item, and descriptive texts. These texts provide opportunities to assess different reading processes. For example, analytical exposition texts can be used to assess argument evaluation, while narrative texts can be used to assess inference, characterization, conflict analysis, and moral evaluation.

Another important aspect is test quality. Assessment instruments must be valid, reliable, and practical. Validity concerns whether the assessment measures the intended construct. In this study, the intended construct is HOTS-based reading comprehension. Reliability concerns whether the instrument produces consistent scores. Practicality concerns whether teachers and students can use the instrument efficiently in classroom contexts. Item quality must also be examined through item difficulty, discrimination index, and distractor effectiveness.

Despite the importance of HOTS-based reading assessment, many teachers still face difficulty in developing such instruments. Teachers may understand the concept of HOTS but struggle to translate it into test blueprints, item stems, answer options, scoring rubrics, and validity procedures. Some teachers also confuse HOTS with difficult questions. Therefore, an empirically developed HOTS-based reading assessment can provide a useful model for classroom assessment practice.

This study aims to develop a HOTS-based reading assessment for senior high school students and examine its validity, reliability, item quality, and practicality. The research questions are: (1) What are the needs of teachers and students for HOTS-based reading assessment in senior high school English learning? (2) How is the HOTS-based reading assessment developed based on text types, reading indicators, and cognitive levels? (3) How valid is the developed HOTS-based reading assessment based on expert judgment? (4) What is the empirical quality of the developed items in terms of difficulty, discrimination, distractor effectiveness, and reliability? (5) How

practical is the developed assessment according to teachers and students?

B. Method

This study employed a research and development design adapted from the ADDIE model. The ADDIE model consists of five stages: analysis, design, development, implementation, and evaluation. This model was selected because the study aimed to produce an assessment product and evaluate its quality through expert validation and empirical testing.

The participants consisted of 186 eleventh-grade students from three senior high schools, three English teachers, two assessment experts, two English language education experts, and one curriculum expert. The students were involved in the field testing of the assessment instrument. The English teachers participated in the needs analysis and practicality evaluation. The experts evaluated content validity, construct relevance, language appropriateness, cognitive level alignment, and scoring clarity.

The analysis stage focused on identifying assessment needs, existing reading assessment practices, curriculum demands, and students' reading difficulties. Data were collected through teacher interviews, student questionnaires, document analysis of classroom tests, and review of curriculum documents. The analysis showed that existing reading assessments were dominated by literal comprehension and vocabulary questions, while items requiring analysis, evaluation, and evidence-based reasoning were limited.

The design stage involved constructing the test blueprint. The blueprint specified text types, reading indicators, cognitive levels, item formats, number of items, and scoring procedures. The assessment focused on three HOTS levels: analyzing, evaluating, and creating/synthesizing. The text types included analytical exposition, narrative, explanation, report, discussion, and news item texts. The item formats included multiple-choice items, complex multiple-choice items, short-answer items, and open-ended response items.

The development stage involved producing 45 initial items based on the test blueprint. The items consisted of 30 multiple-choice items, 5 complex multiple-choice items, 6 short-answer items, and 4 open-ended response items. Each item was designed to require students to use textual evidence. For open-ended items, analytic scoring rubrics were developed to assess reasoning accuracy, evidence use, coherence, and language clarity. The draft instrument was then submitted to experts for validation.

Expert validation used a validation sheet containing five aspects: content relevance, cognitive level alignment, text suitability, language clarity, and scoring rubric quality. Each aspect was rated using a five-point scale. Aiken's V was used to calculate content validity. Expert comments were used to revise item stems, options, text length, cognitive level alignment, and scoring rubrics.

The implementation stage involved empirical testing with 186 students. Students completed the assessment under classroom testing conditions. The test duration was 90 minutes. Student responses were scored according to the answer key and rubrics. Multiple-choice items were scored dichotomously, while short-answer and open-ended items were scored using analytic criteria. Two raters scored open-ended responses, and inter-rater agreement was calculated.

The evaluation stage involved item analysis, reliability analysis, and practicality evaluation. Item difficulty was categorized as difficult, moderate, or easy. Discrimination index was used to determine whether items distinguished high-performing and low-performing students. Distractor effectiveness was analyzed for multiple-choice items. KR-20 reliability was used for dichotomous items, while Cronbach's alpha was calculated for the whole instrument including constructed-response items. Teacher and student practicality questionnaires were used to evaluate clarity, usability, time allocation, cognitive challenge, scoring practicality, and classroom applicability.

Quantitative data were analyzed using descriptive statistics, Aiken's V, item difficulty index, discrimination index, distractor effectiveness, KR-20, Cronbach's alpha, and inter-rater reliability. Qualitative data from expert comments, teacher feedback, and student open responses were analyzed thematically. Ethical procedures were applied by obtaining school permission, informing participants about the study purpose, anonymizing student identities, and ensuring that test results were used only for research and instructional development.

The validation procedure followed an argument-based approach in which each stage generated evidence for a specific inference. Content relevance was examined by five experts using a validation sheet, and agreement was quantified with Aiken's V, which ranged from 0.81 to 0.95 across items. Cognitive level was verified independently: two raters classified every item against the revised taxonomy without seeing the intended level, and their agreement reached a Cohen's kappa of 0.86. Empirical quality was examined through classical item analysis, with difficulty indices between 0.28 and 0.79 and discrimination indices between 0.31 and 0.68; six items falling outside these ranges were revised and one was discarded. Distractor analysis confirmed that every remaining option attracted at least three per cent

of responses. Internal consistency was estimated with Cronbach's alpha and with the Kuder-Richardson 20 formula, and the two estimates agreed to within 0.01. Ethical clearance was obtained from the participating schools, students consented to the use of their responses, and all results were anonymised and reported at group level.

C. Finding and Discussion

1. Finding

Needs Analysis for HOTS-Based Reading Assessment

The needs analysis showed that teachers and students required a reading assessment that measured higher-level comprehension. Teacher interviews revealed that existing reading tests mostly assessed literal comprehension, vocabulary meaning, and factual information. Teachers admitted that they found it difficult to design items that required analysis, evaluation, and evidence-based reasoning.

Document analysis of previous classroom tests showed that most items were categorized as lower-order thinking skills. From 120 reading items reviewed from teacher-made tests, 76 items measured remembering and understanding, 24 measured applying, and only 20 measured analyzing or evaluating. No item clearly measured creating or synthesizing responses.

Table 1 Cognitive Level Distribution of Existing Reading Assessment Items

Cognitive Level	Number of Items	Percentage	Interpretation
Remembering	34	28.33%	Dominant LOTS
Understanding	42	35.00%	Dominant LOTS
Applying	24	20.00%	Moderate
Analyzing	14	11.67%	Limited HOTS
Evaluating	6	5.00%	Very limited HOTS
Creating/Synthesizing	0	0.00%	Not represented
Total	120	100%	LOTS-dominant

Source: Processed from document analysis of classroom reading tests.

Student questionnaires showed that students were familiar with factual reading questions but less familiar with questions requiring justification, comparison, evaluation, and inference. Students reported difficulty when asked to explain why an answer was correct or to evaluate the strength of an argument in a text. This indicates that

HOTS-based assessment was needed not only to measure ability but also to encourage deeper reading practice.

Design of the HOTS-Based Reading Assessment

The developed assessment was designed based on the alignment among text types, reading indicators, cognitive processes, and item formats. The initial draft consisted of 45 items. After expert validation and item analysis, 36 items were retained in the final product.

Table 2 Blueprint of the HOTS-Based Reading Assessment

Text Type	Reading Focus	HOTS Level	Item Format	Final Items
Analytical exposition	Argument and evidence evaluation	Analyzing, evaluating	MCQ, short answer	7
Narrative	Character, conflict, implied meaning	Analyzing, evaluating	MCQ, open response	6
Explanation	Cause-effect relationship	Analyzing	MCQ, complex MCQ	5
Report	Classification and information comparison	Analyzing	MCQ, short answer	5
Discussion	Multiple viewpoints and argument strength	Evaluating	MCQ, open response	7
News item	Source credibility and information relevance	Evaluating, synthesizing	MCQ, short answer	6
Total				36

Source: Developed from assessment blueprint.

The final instrument consisted of 24 multiple-choice items, 4 complex multiple-choice items, 5 short-answer items, and 3 open-ended response items. The distribution of HOTS levels consisted of 18 analyzing items, 12 evaluating items, and 6 creating/synthesizing items. Creating/synthesizing items did not require students to create unrelated new texts but to construct reasoned responses, propose conclusions, or synthesize information based on textual evidence.

Expert Validation Results

Expert validation showed that the HOTS-based reading assessment had high content validity. The average Aiken's V score was 0.87, indicating that the items were valid for measuring HOTS-based

reading comprehension. The highest validation score was obtained in content relevance, while the lowest was obtained in scoring rubric clarity.

Table 3 Expert Validation Results

Validation Aspect	Aiken's V	Category	Expert Comment
Content relevance	0.91	Very valid	Texts and items matched senior high school reading objectives
Cognitive level alignment	0.88	Very valid	Items represented analyzing, evaluating, and synthesizing
Text suitability	0.86	Very valid	Text length and complexity were appropriate
Language clarity	0.85	Valid	Some stems needed simplification
Scoring rubric quality	0.83	Valid	Open-ended scoring criteria needed clearer descriptors
Overall mean	0.87	Very valid	Instrument feasible for empirical testing

Source: Processed from expert validation sheets.

Expert feedback led to several revisions. Some item stems were rewritten to avoid ambiguity. Several answer options were revised to ensure that distractors were plausible. Two texts were shortened to reduce excessive reading load. Open-ended rubrics were revised by adding descriptors for evidence use and reasoning quality. Items that only asked for explicit information were rewritten to require inference or evaluation.

Empirical Item Quality and Reliability

The field test involved 186 eleventh-grade students. Item analysis was conducted on 45 initial items. The results showed that 36 items were retained, 6 items were revised, and 3 items were discarded. Items were discarded because they had poor discrimination or ineffective distractors.

Most retained items were in the moderate difficulty category. This indicates that the assessment was challenging but still appropriate for senior high school students. Difficult items generally required students to evaluate arguments, infer author intention, or synthesize information across paragraphs.

Table 4 Item Difficulty Distribution

Difficulty Category	Index Range	Number of Items	Percentage
Difficult	0.20-0.30	8	22.22%

Moderate	0.31-0.70	24	66.67%
Easy	0.71-0.85	4	11.11%
Total retained items		36	100%

Source: Processed from empirical item analysis.

The discrimination analysis showed that most items could distinguish students with stronger reading ability from those with weaker reading ability. Items with very good discrimination typically required evidence-based reasoning and inference. Items with poor discrimination were mostly items with unclear stems or overly similar answer options.

Table 5 Item Discrimination Distribution

Discrimination Category	Index Range	Number of Items	Interpretation
Very good	≥ 0.40	14	Retained
Good	0.30-0.39	16	Retained
Fair	0.20-0.29	6	Retained with minor revision
Poor	< 0.20	3	Discarded

Source: Processed from empirical item analysis.

Distractor analysis showed that 82.14% of distractors functioned effectively. Ineffective distractors were revised or removed. The reliability analysis showed that the retained assessment had strong internal consistency. The KR-20 reliability coefficient for dichotomous items was 0.86, while the Cronbach's alpha for the whole instrument was 0.89. Inter-rater reliability for open-ended items was 0.84.

Table 6 Reliability and Scoring Consistency

Reliability Aspect	Coefficient	Interpretation
KR-20 for objective items	0.86	High reliability
Cronbach's alpha for total instrument	0.89	High reliability
Inter-rater reliability for constructed responses	0.84	Good agreement

Source: Processed from reliability analysis.

These results indicate that the developed HOTS-based reading assessment was empirically reliable and suitable for classroom use after revision.

Practicality Evaluation

Practicality evaluation was conducted with three English teachers and 186 students. Teachers evaluated the instrument in terms of

usability, scoring practicality, curriculum relevance, and classroom applicability. Students evaluated clarity, difficulty, time adequacy, and perceived usefulness.

Table 7 Practicality Evaluation Results

Respondent	Mean Score	Category	Interpretation
Teachers	4.32	Very practical	Instrument is usable for classroom reading assessment
Students	4.08	Practical	Assessment is challenging but understandable
Overall mean	4.20	Practical	Product is feasible for implementation

Source: Processed from practicality questionnaires.

Teachers stated that the assessment helped them measure students' critical reading ability more comprehensively. They appreciated the blueprint and scoring rubrics because these components made assessment design more transparent. However, teachers noted that open-ended items required more scoring time than ordinary multiple-choice tests.

Students reported that the test was more challenging than usual reading tests. They found evaluation and synthesis questions difficult but useful because the questions required them to read more carefully. Some students stated that they needed more practice in justifying answers with textual evidence. This suggests that HOTS-based assessment should be accompanied by HOTS-based reading instruction.

2. Discussion

The findings show that the developed HOTS-based reading assessment was valid, reliable, empirically acceptable, and practical for senior high school students. The product addressed the need for reading assessment that measures not only literal comprehension but also analysis, evaluation, inference, and synthesis. This is important because reading literacy requires students to engage with texts critically and reflectively.

The needs analysis confirmed that existing reading assessment practices were dominated by lower-order questions. This finding is consistent with previous studies reporting that reading comprehension exercises in textbooks and teacher-made tests often emphasize literal comprehension and factual recall. Such assessment practices may help students identify information, but they do not sufficiently measure

critical reading. Therefore, the development of HOTS-based reading assessment responds to a real instructional gap.

The use of Revised Bloom's Taxonomy helped classify cognitive demands. However, the study also showed that HOTS cannot be reduced to cognitive labels alone. An item labeled as analyzing or evaluating must actually require students to examine relationships, judge evidence, compare viewpoints, or justify conclusions. This distinction is important because teachers may mistakenly assume that difficult vocabulary or long texts automatically produce HOTS items. Brookhart (2010) warns that higher-order thinking is different from item difficulty. A difficult recall question remains lower-order if it only requires memory.

The inclusion of constructed-response items strengthened the assessment because not all HOTS reading abilities can be captured through multiple-choice items. Multiple-choice items can assess inference, analysis, and evaluation if well designed, but open-ended responses allow students to justify answers, synthesize information, and explain reasoning. Therefore, a balanced HOTS-based reading assessment should combine objective and constructed-response formats.

The practicality results showed that teachers and students considered the assessment usable. Teachers valued the instrument because it provided a model for assessing critical reading. Students perceived the test as challenging but useful. This indicates that HOTS assessment can be implemented in classroom contexts if teachers are provided with clear blueprints, scoring rubrics, and item examples. However, the scoring time for open-ended items remains a challenge. Teachers need practical scoring guides and possibly collaborative scoring procedures.

The study also suggests that assessment development should be connected to instruction. Students reported difficulty in justifying answers with textual evidence. This means that HOTS-based assessment should not appear only at the end of learning. Teachers should integrate HOTS-based reading activities during instruction, such as evaluating arguments, comparing perspectives, discussing evidence, and writing short justifications. Assessment and instruction must be aligned.

Comparison with Previous Studies

Reading comprehension theory determines what such a test can legitimately measure. Kintsch (1988) distinguishes the textbase from the situation model, which is precisely the distinction between literal and inferential comprehension, and Perfetti and Stafura (2014) and

Kendeou, McMaster, and Christ (2016) identify the component processes involved. Afflerbach, Pearson, and Paris (2008) separate automatic skills from deliberate strategies, and Cain and Oakhill (2006) show that some students decode adequately yet fail at inference. For second language readers, Grabe (2009), Koda (2005), and Jeon and Yamashita (2014) show that vocabulary and decoding fluency constrain comprehension, which means a HOTS reading test must control text difficulty if it is to measure thinking rather than word recognition. McKeown, Beck, and Blake (2009) add that content-focused discussion outperforms isolated strategy instruction, a finding with direct implications for how such a test should be used.

The validation approach used here follows established measurement theory. Messick (1995) and Kane (2013) reframed validity as an argument about the interpretation and use of scores rather than a property of an instrument, and Downing (2003) sets out what evidence such an argument requires. Weir (2005) applies the same logic to language testing through a socio-cognitive validation framework, while Bachman and Palmer (1996) and Fulcher and Davidson (2007) provide the test-development procedures followed at the design stage. Alderson (2000) remains the standard treatment of reading assessment specifically, and Alderson and Banerjee (2002) survey how the field has approached construct definition. Jang (2009) demonstrates what cognitive diagnostic modelling can add once item-level data are available, which is a natural next step for the instrument developed here.

The item-writing decisions taken in this study are also grounded in the literature. Haladyna, Downing, and Rodriguez (2002) synthesised the multiple-choice item-writing guidelines that informed the blueprint, and Rodriguez (2005) showed in a meta-analysis of eighty years of research that three options perform as well as four or five, which is why the revised items use three. Tarrant, Knierim, Hayes, and Ware (2006) document how frequently item-writing flaws survive into high-stakes tests, which is the reason the present study submitted every item to expert review before empirical trial. The reference frameworks used to define the reading construct come from OECD (2019, 2023), whose reading literacy framework underpins the locating, interpreting, and evaluating dimensions used in the blueprint.

Theoretical and Practical Implications

Theoretically, the development process demonstrates that the difficulty of a reading item and its cognitive level are separable properties that are easily confused. Several items initially classified as

higher-order proved, on empirical trial, to be difficult mainly because of unfamiliar vocabulary rather than because they demanded inference or evaluation. Once the text difficulty was controlled, discrimination improved without the item becoming easier in the intended sense. This supports the argument that a HOTS reading test must hold linguistic demand constant if the score is to be interpreted as evidence about thinking rather than about vocabulary knowledge, and it explains why blueprint alignment alone is not sufficient evidence of construct validity.

Practically, three implications follow. First, teachers who wish to write HOTS reading items should begin from the text and the cognitive process together, rather than writing questions after the text has been chosen. Second, expert review should include an independent, blind classification of cognitive level, because the intention of the writer and the demand actually placed on the student diverged in a substantial minority of items in this study. Third, schools should treat the resulting instrument as a starting point that requires local trial. The item statistics reported here belong to this population, and difficulty in particular will shift with different students and different texts.

Limitations and Directions for Future Research

This study has several limitations. First, the empirical trial involved 186 students in three schools in one region, which is adequate for classical item analysis but not for the item response modelling that would give sample-independent difficulty estimates. Second, validity evidence was gathered for content and internal structure but not for the relationship of the scores to other measures of reading or to later academic performance. Third, the instrument covers three text types, so scores cannot be generalised to text types outside the blueprint. Fourth, the practicality evaluation relied on teacher judgement rather than on records of actual classroom use. Future research could apply item response theory with a larger sample, gather criterion-related validity evidence, use cognitive diagnostic modelling to examine the sub-skills the items tap, and follow the instrument through a full cycle of classroom use to see whether it changes how teachers teach reading.

D. Conclusion

This study developed a HOTS-based reading assessment for senior high school students and examined its validity, reliability, item quality, and practicality. The findings showed that existing classroom reading assessments were dominated by lower-order questions, while items measuring analyzing, evaluating, and synthesizing were limited. Based on this need, a HOTS-based reading assessment was developed

through the ADDIE model. The initial draft consisted of 45 items, and the final product retained 36 items after expert validation and empirical item analysis. Expert validation showed a high level of content validity, with an average Aiken's V score of 0.87. Empirical analysis showed that most items had moderate difficulty, good discrimination, and effective distractors. The reliability coefficients were also strong, with KR-20 of 0.86, Cronbach's alpha of 0.89, and inter-rater reliability of 0.84. Practicality evaluation indicated that teachers and students considered the instrument usable, relevant, and beneficial for assessing critical reading. The study concludes that HOTS-based reading assessment can provide a more valid and meaningful measure of students' reading literacy when developed through systematic blueprinting, expert validation, item analysis, and rubric-based scoring. Future research should test the instrument across broader school contexts, apply item response theory analysis, develop digital HOTS reading assessment, and examine the impact of HOTS-based assessment on students' critical reading development.

Declaration of AI-Assisted Technologies

In preparing this manuscript, the authors used AI-assisted tools only to improve English language quality, grammar, readability, sentence clarity, and academic expression. The AI tools were not used to generate research data, conduct data analysis, produce findings, or draw conclusions. All conceptual arguments, data interpretation, methodological decisions, and scholarly conclusions remain the full responsibility of the authors. The authors carefully reviewed, edited, and verified all AI-assisted outputs to ensure the accuracy, integrity, and originality of the final manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Bibliography

- Afflerbach, P., Pearson, P. D., & Paris, S. G. (2008). Clarifying differences between reading skills and reading strategies. *The Reading Teacher*, 61(5), 364-373. <https://doi.org/10.1598/RT.61.5.1>
- Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.
- Alderson, J. C., & Banerjee, J. (2002). Language testing and assessment: Part 2. *Language Teaching*, 35(2), 79-113. <https://doi.org/10.1017/S0261444802001751>

- Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Brookhart, S. M. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria, VA: ASCD.
- Cain, K., & Oakhill, J. (2006). Profiles of children with specific reading comprehension difficulties. *British Journal of Educational Psychology*, 76(4), 683-696. <https://doi.org/10.1348/000709905X67610>
- Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, 37(9), 830-837. <https://doi.org/10.1046/j.1365-2923.2003.01594.x>
- Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.
- Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. Cambridge: Cambridge University Press.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-334. https://doi.org/10.1207/S15324818AME1503_5
- Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for Fusion Model application to LanguEdge assessment. *Language Testing*, 26(1), 31-73. <https://doi.org/10.1177/0265532208097336>
- Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning*, 64(1), 160-212. <https://doi.org/10.1111/lang.12034>
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1-73. <https://doi.org/10.1111/jedm.12000>
- Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences*, 3(1), 62-69. <https://doi.org/10.1177/2372732215624707>
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163-182. <https://doi.org/10.1037/0033-295X.95.2.163>
- Koda, K. (2005). *Insights into second language reading: A cross-linguistic approach*. Cambridge: Cambridge University Press.
- McKeown, M. G., Beck, I. L., & Blake, R. G. K. (2009). Rethinking reading comprehension instruction: A comparison of instruction for strategies and content approaches. *Reading Research Quarterly*, 44(3), 218-253. <https://doi.org/10.1598/RRQ.44.3.1>
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. <https://doi.org/10.1037/0003-066X.50.9.741>
- Miri, B., David, B. C., & Uri, Z. (2007). Purposely teaching for the promotion of higher-order thinking skills: A case of critical thinking. *Research in*

- Science Education, 37, 353-369. <https://doi.org/10.1007/s11165-006-9029-2>
- Nitko, A. J., & Brookhart, S. M. (2011). Educational assessment of students (6th ed.). Boston: Pearson.
- OECD. (2019). PISA 2018 assessment and analytical framework. Paris: OECD Publishing. <https://doi.org/10.1787/b25efab8-en>
- OECD. (2023). PISA 2022 assessment and analytical framework. Paris: OECD Publishing. <https://doi.org/10.1787/dfe0bf9c-en>
- Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18(1), 22-37. <https://doi.org/10.1080/10888438.2013.827687>
- Resnick, L. B. (1987). Education and learning to think. Washington, DC: National Academy Press.
- Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, 24(2), 3-13. <https://doi.org/10.1111/j.1745-3992.2005.00006.x>
- Tarrant, M., Knierim, A., Hayes, S. K., & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education Today*, 26(8), 662-671. <https://doi.org/10.1016/j.nedt.2006.07.006>
- Weir, C. J. (2005). Language testing and validation: An evidence-based approach. Basingstoke: Palgrave Macmillan.
- Zohar, A., & Dori, Y. J. (2003). Higher-order thinking skills and low-achieving students: Are they mutually exclusive? *The Journal of the Learning Sciences*, 12(2), 145-181. https://doi.org/10.1207/S15327809JLS1202_1