International Journal of Language and Literary Studies
Volume 8, Issue 6, 2026
Homepage : http://ijlls.org/index.php/ijlls

Evaluating Reading-Test Validity in Cambridge First: A Qualitative Item Analysis

Ivana Popovikj & Tijana Popovikj
Corresponding Author Email: ivana.pretty.91@gmail.com

DOI: https://doi.org/10.36892/ijlls.v8i6.2858

APA Citation: Popovikj, I., & Popovikj, T. (2026). Evaluating Reading-Test Validity in Cambridge First: A Qualitative Item Analysis. *International Journal of Language and Literary Studies*, *8*(6).514-522. https://doi.org/10.36892/ijlls.v8i6.2858

Received:
02/09/2026

Accepted:
20/11/2026

Keywords:
Reading Assessment, Validity, Cambridge FCE, CEFR, Qualitative Item Analysis

Abstract

This article presents a qualitative, document-based evaluation of the reading component of a historical Cambridge First Certificate in English (FCE) sample examination. It examines how task formats and selected items represent reading constructs associated with B2 proficiency and considers content representation, construct relevance, face validity, potential ambiguity, and apparent task demands. The analysis revisits an existing item-level critique of the three-part reading paper and situates it within contemporary argument-based understandings of assessment validity and the updated Common European Framework of Reference for Languages (CEFR). Findings suggest that the sample covers a meaningful range of reading processes, including comprehension of detail, inference, textual cohesion, and locating specific information. At the same time, certain item interpretations raise questions about ambiguity, the inference demands of individual questions, and the limits of judging difficulty without response data. The study does not establish psychometric validity or the performance of the current B2 First examination. Instead, it illustrates the value and limitations of transparent qualitative item review and identifies priorities for empirical follow-up and classroom assessment literacy.

1. Introduction

Language tests do more than produce marks: their results inform educational progression, certification, and decisions made by institutions and employers. Because decisions may carry significant consequences, the interpretation of scores requires a defensible relationship between the language ability of interest, the tasks used to elicit evidence, and the conclusions drawn from performance. Reading assessment offers a particularly instructive setting for this problem. A response may depend on lexical knowledge, comprehension of discourse structure, inferencing, topic familiarity, and the ability to manage time and distractors. A task can therefore appear to assess reading while also imposing demands that are not equally relevant to every intended interpretation.

The present article examines a historical First Certificate in English (FCE) reading sample, an examination associated with level B2 of the Common European Framework of Reference for Languages (CEFR). Although some of the previously conducted research has established frameworks for evaluating reading constructs and score interpretations, the application of these frameworks to individual items in historical examination materials are seen as a useful tool of inquiry. More precisely, a retrospective analysis can distinguish what

is observable in the task design from what can solely be established through candidate performance. The present analysis considers construct, content, criterion, and face validity alongside perceived item difficulty within an explicitly bounded document analysis. Rather than treating a single sample as evidence for or against the validity of the entire examination, it asks what can reasonably be inferred from task specifications, texts, and item features when no candidate responses are available.

The historical distinction is essential. The source sample comprises three reading parts with 30 questions, whereas the current B2 First examination integrates Reading and Use of English into a seven-part paper. The reading tasks are now identified as Parts 5–7, with different item counts (University of Cambridge ESOL Examinations, 2006, pp. 6–11). Accordingly, observations made about the historical paper cannot be transferred uncritically to the present examination. The case is nevertheless valuable because it makes possible a close examination of how apparently appropriate reading tasks may raise questions about inference, ambiguity, and the interpretation of proficiency descriptors. Its contribution lies in showing how these questions can be formulated as testable hypotheses rather than treated as established weaknesses.

The study addresses three questions: (1) Which reading processes and task demands are represented in the selected historical FCE reading paper? (2) What potential strengths and threats to validity emerge from a qualitative review of its items? (3) What claims remain unsupported in the absence of candidate responses, piloting, or independent evidence? The contribution is methodological and interpretive: it demonstrates how a document-based review can generate testable hypotheses without overstating what the available evidence establishes.

2. Literature Review and Conceptual Framework

2.1 Validity as an argument about interpretation and use

An argument-based perspective offers a useful qualification: validity concerns the defensibility of specific interpretations and uses of scores, rather than an unconditional property of an examination (Kane, 2012). The present documentary review therefore evaluates the plausibility of selected task interpretations; it does not establish empirical validity evidence.

Validity has been described through several overlapping traditions. Earlier language-testing accounts distinguished content, construct, criterion-related, and face validity (Alderson et al., 1995; Brown, 2003; Weir, 1993). These categories remain useful for organizing questions about a test, but should not be treated as independent labels that a paper can simply pass or fail. As Fulcher and Davidson (2007) emphasize in their discussion of language-test design and evaluation, judgments about assessment need to be related to the evidence on which they are based. The central issue is therefore whether the available evidence supports a particular interpretation and use of test results. A judgment about a reading task based solely on its surface characteristics is narrower than a claim that examination scores measure a stable level of reading proficiency.

Bachman and Palmer (1996) argue that language-test usefulness must be considered in relation to test purpose, reliability, construct validity, authenticity, interactiveness, impact, and practicality. These dimensions help explain why a plausible reading task alone cannot guarantee defensible score-based decisions.

Construct considerations concern the relationship between the abilities that a test intends to assess and the performances its tasks actually require. Reading is not a single observable behaviour. Comprehending a text may involve identifying explicit information, integrating ideas across sentences, recognizing an author's stance, and inferring meanings not directly stated. The task designer must determine which processes are relevant to the target proficiency and how items can elicit evidence of them. The evaluator must then consider

alternative explanations for correct and incorrect responses, including distractor strategies, vocabulary knowledge, background familiarity, and unintended ambiguity.

Alderson (2000) distinguishes reading processes and assessment tasks, highlighting the importance of aligning observable responses with the reading abilities a test claims to assess. In this study, item demands are therefore described as hypotheses about processes rather than direct measurements of those processes.

Content-related considerations address whether the range of texts, task types, and processes reasonably represents the target domain. A paper containing several task formats may provide broader coverage than a paper relying on one format, but variety alone does not demonstrate representative sampling. Claims about content depend on a sufficiently clear specification of the intended domain and on systematic comparison with the assessment material. Criterion-related evidence, in contrast, requires an external measure or outcome with which test performance can be compared. It cannot be established by inspecting a test booklet.

Hughes (2003) also distinguishes practical test-development considerations from claims about what scores mean. This distinction matters for the historical sample: a well-formed multiple-choice task may still invite competing interpretations of what successful performance demonstrates.

Face validity concerns how plausible and appropriate a task appears to users or stakeholders. It may influence acceptance and engagement, yet apparent authenticity does not by itself establish that the intended construct is measured. Multiple-choice questions can reduce the need for extended written production, while introducing their own demands associated with option interpretation and elimination. Likewise, a gapped-text task can target discourse organization while being influenced by candidates' familiarity with the format. These are hypotheses about task demands, not automatic conclusions about validity.

2.2 Reading proficiency and CEFR alignment

The CEFR provides a shared language for describing communicative proficiency and the activities learners can perform at different levels (Council of Europe, 2001). The Companion Volume updates and extends illustrative descriptors, including additional detail relevant to reading (Council of Europe, 2020). These descriptors provide reference points for the present examination of task demands, but do not constitute a mechanical key for assigning a single CEFR level to each item without further evidence.

At B2, reading tasks may reasonably involve understanding extended texts, locating information, recognizing perspectives, and following argumentation. Some inferencing or interpretation of attitude is therefore not automatically evidence that an item exceeds B2. The difficulty of a particular question depends on the text, language, response options, contextual support, and the characteristics of the test-taking population. A question that appears demanding to an analyst may function differently when administered to a representative group of candidates.

McNamara and Roever (2006) place language testing within its social context, reminding researchers that the consequences and stakeholders of assessment are part of a broader evaluation. Because the present study lacks test-taker and stakeholder data, it does not evaluate these social consequences directly.

Alignment analysis should consequently distinguish intended skill from observed difficulty. Intended skill is inferred from specifications and the operations apparently required to answer. Observed difficulty is an empirical property of candidate responses, often summarized through item facility and examined alongside discrimination and other evidence. Without response data, a reviewer can identify plausible sources of difficulty but cannot determine their actual magnitude or locate an item reliably on a proficiency scale.

2.3 Reading formats and the historical FCE paper

The source material presents a three-part historical reading paper: multiple-choice comprehension, gapped text, and multiple matching. These formats can also be considered in relation to the examination specifications documented in the Cambridge English Language Assessment (2015) handbook, while recognizing that the handbook describes a later version of the examination. The multiple-choice section invites judgments about explicit detail, implied meaning, attitude, and the main idea. The gapped-text section requires the reader to restore sentence-level connections and track discourse progression. The matching section requires information to be located and compared across shorter texts. This range gives the analyst an opportunity to examine both local comprehension and global textual processing.

Task format also shapes what is observable. A selected answer shows which option the candidate chose, not the reasoning used to choose it. A correct response may reflect comprehension, partial comprehension, or effective elimination; an incorrect response may reflect misreading, uncertainty, or confusion caused by a distractor. Consequently, the present analysis focuses on potential construct representation and item-design risks rather than claiming to identify the mental processes of actual candidates.

3. Materials and Methods

3.1 Design and materials

This study is a qualitative, document-based case analysis. The primary material is the FCE Reading sample paper (Paper 1) published by University of Cambridge ESOL Examinations (2006, pp. 6–11). The primary material is a historical FCE reading sample included in the appendix to this article. The examination tasks and selected questions are examined through qualitative item analysis. The sample contains three parts and 30 numbered questions. The analysis is restricted to this reading sample; no listening, speaking, writing, or language-use scores are examined. The article also draws on published descriptions of reading constructs and CEFR levels to interpret the task demands. It does not report newly collected participant data.

The unit of analysis is the individual question considered in the context of its task format and source text. The analysis considers selected questions on items 1–9, 12–14, 20–21, 23, and 29, together with comments about the three task formats. These questions provide the basis for examining reading demands and potential item-design concerns. The appendix contains photographs of the historical examination pages; the present discussion paraphrases the item-level observations rather than reproducing extended passages from the source material. The present discussion paraphrases the item-level observations rather than reproducing extended passages from the source material.

3.2 Analytic procedure

The analysis proceeds in four stages. First, the stated or apparent purpose of each format is described: multiple-choice comprehension, textual reconstruction, or information matching. Second, the reading operations identified in selected items are grouped into descriptive categories: explicit detail, global comprehension, inference, authorial attitude, cohesion/coherence, and locating information. Third, each interpretation is assessed for potential relevance to the intended reading construct and for possible sources of ambiguity or construct-irrelevant difficulty. Finally, claims are classified by evidential status: observations supported by the documents, interpretive hypotheses requiring verification, and psychometric conclusions that cannot be made from the available material.

This is a retrospective analytical reconstruction, not a newly administered or independently coded validation study. There were no raters, participant sample, response distributions, inter-rater reliability estimates, or statistical comparisons. The categories are used as an organizing framework rather than as a validated coding instrument. This

transparency matters: the study can identify issues worthy of closer investigation, but cannot estimate the frequency of problems across the broader test or generalize to all versions of Cambridge First.

3.3 Boundaries and ethical considerations

No human participants were recruited and no personal test-taker data were processed. The analysis concerns published or reproduced assessment material. Since the examination materials belong to their respective rights holders, any republication of full passages or questions should be subject to permission or a clear applicable license. The manuscript therefore paraphrases the item-level observations rather than reproducing the photographed test pages.

An additional limitation is the age of the historical paper. Cambridge has revised the examination since the version analysed, and the CEFR Companion Volume has expanded the reference descriptors. Comparisons with current materials are contextual, not claims that the historical and present-day tests are identical.

4. Results and Discussion

4.1 Coverage of reading processes across the three parts

The three task formats provide complementary opportunities to elicit evidence of reading comprehension. Multiple-choice items in the selected sample require recognition of detail, inferred meaning, and attitudes expressed in extended prose. The gapped-text task emphasizes the relationships between sentences and the development of an argument or narrative. Multiple matching directs attention to specific information across a collection of shorter texts. In combination, these tasks appear to sample several important dimensions of reading rather than relying on one narrowly defined skill.

This variety is a potential strength for content representation, but the extent of coverage cannot be determined by counting formats alone. The selection and difficulty of texts, balance of genres, distribution of item targets, and correspondence with a formal test specification would all need to be documented. In particular, a conclusion that the examination is content-valid solely because it includes a broad spectrum of reading performances would be too strong. A more defensible conclusion is that the selected sample shows prima facie diversity of reading demands consistent with the stated purpose of a B2 reading assessment.

The format-specific distinction is important for interpretation. For example, locating a detail is not identical to understanding a writer's attitude, and restoring a missing sentence may involve recognizing discourse markers as well as integrating meaning across paragraphs. If the final score combines these performances, an explicit account is needed of why they constitute appropriate evidence of the target reading construct. The historical sample makes this issue visible but does not resolve it.

4.2 Multiple-choice comprehension: detail, inference, and attitude

Several first-part questions provide examples of reading operations broadly associated with B2. Item 1 is interpreted as requiring recognition of a narrator's regret about a choice made during a bus journey. Item 3 is treated as an inference from a figurative phrase; items 4 and 5 require attention to relevant details; and item 6 involves comparing the description of a place in a guidebook with what the narrator observes. These examples illustrate the distinction between retrieving a stated fact and constructing an interpretation from several textual cues.

Item 2 is described as requiring the reader to infer that a writer has been selected for an interview. This demand is broadly compatible with B2, although a general descriptor does not specify every sub-operation involved in arriving at an answer. This illustrates the limits of descriptor-based alignment. A CEFR level is a statement about patterns of communicative performance, not a complete inventory of every possible inferential act. The analyst should

therefore ask whether the inference is accessible from contextual cues and whether the alternatives distinguish successful comprehension from plausible but unintended readings.

Item 7 receives particular attention because it is said to require inference of the meaning of "ivy" from context. Assigning a precise CEFR level to this demand is problematic. A more precise interpretation is that lexical inferencing may be relevant to reading across levels, while the difficulty of the specific item remains unmeasured. The relevant evidence would include the surrounding sentence, the extent to which the target word is necessary to answer, and candidates' actual patterns of selection. An isolated unfamiliar word does not establish a mismatch with the intended proficiency level.

Item 8 is interpreted in the present analysis as demanding global comprehension of an extended text and recognition of attitude. This demand might appear to exceed B2. Such a claim should remain provisional: global understanding and attitude recognition can be legitimate B2 targets, and their difficulty varies with discourse complexity and the explicitness of cues. To evaluate this question more securely, reviewers would need to examine the exact item stem, all options, the passage, and evidence from representative B2 candidates.

An important consequence of these observations is that content and construct issues are intertwined. If a question is answered mainly through superficial option elimination, its score may not reflect the intended reading process. If it requires a complex but clearly signposted inference, it may still be a suitable B2 item. The presence of inference, by itself, is not a defect. The appropriate validation question concerns the quality and relevance of the evidence elicited by the particular item.

4.3 Gapped text: coherence, contrast, and discourse development

The second part of the historical sample consists of a text with removed sentences and an additional distractor. The task requires candidates to reconstruct a coherent sequence by recognizing continuity, contrast, and relationships among ideas. Item 9 is described as requiring the reader to follow a sequence of setbacks and developments in an account of a downhill racer. Item 14 is said to involve a contrastive relationship. Items 12 and 13 are discussed as requiring recognition of coherence and cohesion across adjacent portions of the text.

These observations suggest that the gapped-text format has a plausible relationship to discourse-level reading. Successful completion may require integration of propositional content, reference tracking, temporal order, and rhetorical relationships. Such operations are more demanding than matching a single repeated word, and they offer a rationale for including this format alongside local comprehension questions. Nevertheless, the format may also reward familiarity with test-taking conventions, such as comparing pronouns and connectors across answer choices. Validation should consider both possibilities.

The relative position and apparent difficulty of this section might be regarded as a potential weakness if easier tasks are expected to come first. This proposition requires more careful treatment. Task order can affect motivation and performance, but the claim that a section is harder cannot be established from its position or scoring weight alone. Difficulty and weight are distinct: weight determines how responses contribute to a score, whereas difficulty concerns how often candidates answer correctly under specified conditions. A controlled comparison or operational item statistics would be required before describing the ordering as a demonstrated validity problem.

The main finding concerning the gapped-text section is therefore not a verdict about task order but an account of the textual relations candidates are asked to reconstruct. Those relations can be articulated in a test specification and subsequently investigated through item review, cognitive interviews, or think-aloud protocols. This would help determine whether the intended discourse skills, rather than unintended shortcuts, explain candidate success.

4.4 Multiple matching: locating information and possible ambiguity

The third part uses shorter texts and matching prompts. Item 20 is examined as a detail-location question involving a speaker who tests whether sewing machines function but otherwise does not handle them. Item 23 is interpreted as requiring attention to a distinguishing word that separates the correct passage from superficially similar alternatives. These examples indicate how the format can assess careful reading across multiple sources rather than simple recognition of repeated vocabulary.

Item 21 is the clearest potential item-design concern in the present analysis. The question appears to target an attitude of regret but that none of the response options expresses that attitude directly. The implication is that stronger candidates may select the intended answer through elimination, while others may find the options misleading. This is a plausible hypothesis about ambiguity, not a confirmed flaw. A full judgment requires close inspection of the complete question and passages, independent expert review, and ideally response-process or item-statistical evidence.

Item 29 is also characterized as involving implicit attitudes and opinions, which may suggest a proficiency demand above B2. Again, the precise level cannot be established from the label "implicit" alone. The key issues are the linguistic cues provided, the extent of integration required, and whether successful B2 readers can reasonably perform the operation. Such distinctions show why qualitative reviews should record the reasoning behind each judgment and avoid equating an analyst's impression of sophistication with empirical item difficulty.

The matching section further highlights a possible construct-representation tension. Efficient scanning is useful for locating information, but a candidate who matches isolated keywords without understanding the passage may produce the same answer as a candidate who has read carefully. Well-designed distractors can reduce superficial matching, yet they can also introduce ambiguity if distinctions are insufficiently clear. A robust evaluation therefore needs both textual analysis and evidence about how test takers actually respond.

4.5 Difficulty, scoring, and the interpretation of B2

The Cambridge English Scale allows stronger performance to be reported above the B2 band. The general point remains relevant: a B2-targeted examination may distinguish performance at different levels rather than consist exclusively of items that every B2 candidate can answer. However, the relationship between a particular question and a reported CEFR level cannot be read directly from score bands. Operational scaling and standard setting rely on broader evidence than qualitative inspection of a few questions.

For current B2 First reporting, Cambridge describes Grade A performance as corresponding to C1 and Grades B and C as corresponding to B2; the present article does not use those current reporting conventions to reclassify individual historical items (Cambridge University Press & Assessment, n.d.). The earlier claim that an item must be invalid because it resembles a C1 descriptor is therefore not retained as a finding. Instead, potentially demanding questions are identified as candidates for empirical study.

Without administration data, this study cannot calculate facility values, item discrimination, internal consistency, differential item functioning, or relationships with external measures. It also cannot establish whether a given item disadvantages any group of candidates. These absences do not make document analysis unhelpful, but they define its proper role: generating explicit, examinable concerns and guiding later validation rather than replacing it.

4.6 Implications for assessment design and teaching

The case offers several implications for practitioners who design or evaluate reading tests. First, assessment specifications should state the intended reading processes clearly enough that an item can be reviewed against them. Broad labels such as "understanding a text" are insufficient for deciding whether an item targets explicit detail, inference, attitude, or

discourse organization. Second, item reviewers should inspect the entire response context, including the passage and distractors, rather than infer validity from an item stem in isolation.

Third, test writers should distinguish accessibility from simplicity. A B2 reading task may legitimately require sustained interpretation, but the relevant cues should be available in the text and the options should avoid unintended competing interpretations. Fourth, teachers should be cautious about interpreting success on multiple-choice or matching tasks as direct evidence of a single underlying strategy. Classroom follow-up activities that ask learners to justify answers with textual evidence can make reasoning more visible, although these activities serve instructional rather than operational test-scoring purposes.

Finally, qualitative review should be connected to empirical evidence whenever stakes warrant it. Candidate interviews can reveal whether a question elicits the intended processes; piloting can reveal unexpected response patterns; and expert panels can evaluate alignment with the target construct. Each source of evidence addresses a different aspect of the validity argument. The historical FCE case shows why a single favorable overall judgment is less informative than a transparent account of what has been examined and what remains unknown**.**

4.7 Interpreting the available evidence

The observations can be organized by distinguishing the claims that are supported by the documents from those in need of further investigation. The three formats and the selected items provide evidence of various reading demands, including attention to detail, inference, attitude, and discourse relationships. Yet, this variety does not establish that the sample adequately represents the intended reading domain. Nor does the apparent relevance of an item prove that the candidates use the reading processes attributed to it.

Items 7 and 21 are considered as the most striking questions for further review. The former questions the issue of whether necessary contextual information supports lexical inference; the latter raises a possible mismatch between the attitude being assessed and the available response options. Both observations are plausible interpretive concerns rather than demonstrated defects. Similarly, items involving implied attitudes cannot be classified as above B2 only on the basis of their apparent inferential demands.

These distinctions are important because a document-based review can identify possible sources of construct-irrelevant difficulty without measuring their effect. Independent review, candidate responses, and response-process evidence would be needed to establish whether the concerns identified here affect item performance or score interpretation.

5. Limitations and Future Research

This article has four major limitations. First, it analyses a historical sample rather than the current B2 First test. Changes to examination structure and descriptors mean that its observations should not be used to characterize contemporary operational forms. Second, the item interpretations have not been independently verified by multiple blinded reviewers. Third, it lacks candidate responses, interviews, and external criteria; it cannot estimate empirical difficulty or validate score interpretations. Fourth, only selected items are discussed in depth, limiting claims about the whole sample.

Future research could begin by obtaining a fully documented and properly licensed test form, specifying the intended reading constructs, and applying a transparent item-review rubric with more than one trained analyst. Reviewers could record the textual evidence supporting each answer and examine whether competing interpretations are possible, particularly in item 21. Agreement and disagreement among reviewers could then be reported. A second stage could involve representative B2 candidates completing the paper while a subset participates in response-process interviews. Item-level statistics could be compared with the qualitative predictions, including whether potentially ambiguous items produce unusual distractor patterns.

A further study could compare historical and contemporary Cambridge reading specifications with the 2020 CEFR descriptors. Such work would need to avoid assuming that changes in task format necessarily imply changes in the underlying construct. It would also benefit from examining fairness, familiarity with topics, and accessibility across learner populations. These directions would move the present exploratory critique toward a fuller validation argument.

6. Conclusion

The historical FCE reading sample examined here offers varied opportunities to assess reading for detail, inference, attitude, discourse cohesion, and specific information. Its three task formats suggest a meaningful range of reading demands associated with B2, while the analysis of selected items identifies questions about contextual support and response-option interpretation that merit closer investigation. These findings do not amount to a judgment about the validity of the examination as a whole.

The practical importance of the study proves how close examination of texts and questions can support more transparent item review. By distinguishing observations from hypotheses about difficulty and ambiguity, teachers and test developers can classify which aspects of a task oblige closer study and which claims require candidate evidence. The historical case therefore proves the contribution of qualitative analysis to a broader, evidence-based approach to reading assessment.

7. References

Alderson, J. C. (2000). *Assessing reading*. Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Brown, H. D. (2003). *Language assessment: Principles and classroom practices*. Pearson Education.

Cambridge English Language Assessment. (2015). *Cambridge English: First handbook for teachers for exams from 2015*.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge University Press.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17. https://doi.org/10.1177/0265532211417210

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Blackwell Publishing.

University of Cambridge ESOL Examinations. (2006). *First Certificate in English: FCE specifications and sample papers for examinations from December 2008*. https://timetolearnesl.wordpress.com/wp-content/uploads/2014/10/fce_sample_papers_with_answers.pdf

Weir, C. J. (1993). *Understanding and developing language tests*. Prentice Hall.