2

Probing the Reading Paper of the Advanced Level EFL School-Leaving Examination

Gyula Tankó
tanko.gyula@btk.elte.hu

Zsuzsanna Andréka
andreka.zsuzsanna@btk.elte.hu

Abstract

The two-level school-leaving examination system introduced in Hungary in 2005 and modified in 2017 is the most important gatekeeping examination for English major tertiary education programs. Experience shows that the language proficiency of the students admitted into the English major education programs is insufficient for effective learning and teaching. Therefore, this exploratory investigation analyzed the reading paper of the advanced EFL school-leaving examination to find out the extent to which it is suitable for the level-appropriate assessment of the aspects of language ability it is intended to assess. For this purpose, 32 tasks from eight reading papers administered in the 2017–2020 period were analyzed. With the use of statistical data available on the scores obtained by test takers alongside a close critical analysis and blind double-coding of test items, several shortcomings of the reading paper were revealed. The flaws uncovered undermine the construct validity of the examination and call into question the generalizability of its scores and its predictive validity. The findings may partly explain the ineffective reading ability of the students admitted into English major programs and should aid in the development of better reading papers.

Keywords: Hungarian EFL school-leaving examination, university admission, construct validity, reading assessment task analysis

DOI: 10.21862/ELTE.DEAL.2022.2.23

Probing the Reading Paper of the Advanced Level EFL School-Leaving Examination

The current two-level school-leaving examination system was introduced in Hungary in 2005. It originally measured foreign language proficiency at the intermediate (A2/B1) and advanced (B2) levels (Council of Europe, 2020), but in 2017, amongst other smaller modifications, the level of the intermediate examination was changed to B1. The foreign language school-leaving examination is compulsory for all secondary school leavers in Hungary (Hungarian Government, 2021), who must start learning a foreign language from grade 4 and a second one from grade 9 (Hungarian Government, 2020)—with English being the most popular language in all secondary level school types (Hungarian Central Statistical Office, 2022)—and who must take either an intermediate or advanced level foreign language school-leaving examination in grades 11 or 12 (Hungarian Government, 2020). In the spring of 2022 alone, 64,958 secondary school leavers took the English school-leaving examination, of which 18,791 chose the advanced examination (Educational Department, 2022b).

The aim of foreign language education in Hungary, as stated in the National Core Curriculum (Hungarian Government, 2020) is to develop the communicative competence of language learners so as to enable the appropriate realization of their communicative goals in authentic language use situations. Further foreign language education goals set by the curriculum are to capacitate language learners to access information efficiently, to assist their mobility for study purposes, to facilitate their entry into the labor market, and to aid them in the pursuit of their tertiary level studies.

An advanced level EFL secondary school leaving certificate is necessary for admission to English major tertiary education BA programs in Hungary, and either an intermediate or advanced level certificate is required for MA teacher training programs depending on the combination of subjects selected by the applicants (Educational Department, 2022c).

For example, secondary school graduates applying for an English-German teacher training MA program can meet the English language proficiency admission requirements even with an intermediate level certificate if they have an advanced level school-leaving certificate in German. However, even if the admission decisions are made on the basis of an advanced level school-leaving certificate in English, the question remains as to what the reason for the low language ability of first-year English major students admitted into English medium instruction programs may be. One of the likely explanations suggested by research findings (e.g., Dávid, 2008; Szabó & Kiszely, 2010; Tankó & Andréka, 2021) is that the advanced level EFL school-leaving examination itself does not assess language ability as intended. Given that further empirical investigation of the advanced level EFL school-leaving examination is needed, especially because high-stakes admission decisions are made based on it, the current study was carried out to add to the rather limited body of research (see Illés, 2011; Szabó, 2019) available on the reading paper of the advanced level EFL school-leaving examination.

Theoretical Background

Reading in Language Assessment

As Alderson (2000) discussed, assessing reading ability is a complex process determined primarily by the model of reading based on which assessment developers define the construct of reading for a specific purpose or setting. In order to differentiate poor and good readers or to make predictions based on their test performance about how they are likely to perform in other settings on other reading tasks, assessment developers must also make decisions on whether they wish to assess the reading process (i.e., how comprehension is reached) or product (i.e., the poor, fair, or good comprehension that is reached). This requires them to consider what levels of understanding they want to measure (i.e., whether they wish to assess the comprehension of literal or inferred text meaning, or the ability to critically evaluate a text) and how to distinguish acceptable and unacceptable interpretations of a text. Furthermore, so as not to test them explicitly, assessment developers have to separate reading ability from other cognitive abilities or functions (e.g., reasoning or working memory); from the readers' background, topical, and cultural knowledge; as well as from other components of language ability—such as grammar and vocabulary—that have been found to correlate with reading ability measures. This list of factors, variables, and decisions to be made is far from complete, but it is sufficient to illustrate the complex nature of reading proficiency assessment and why assessment scholars tend to shy away from formulating a definition of reading ability.

A rare exception is the definition proposed by Urquhart and Weir (1998), according to whom "reading is the process of receiving and interpreting information encoded in language form via the medium of print" (p. 22). However, a mere cursory look at this definition is enough to appraise the difficulties it poses for an assessment developer who attempts to operationalize it. Instead, reading assessment developers typically measure reading strategies, skills, or sub-processes that are part of the model of reading they adapt (see Alderson, 2000).

An influential and modern taxonomy of reading behaviors used to assess general and academic English reading ability was proposed by Khalifa and Weir (2009). It is part of the cognitive processing model for reading comprehension and differentiates reading behaviors across two dimensions: careful versus expeditious and local versus global. The faster, goal-driven, selective reading behaviors—namely, skimming, search reading, scanning, and browsing—are expeditious reading strategies. The slower ones—which include the understanding of lexis and grammar which is usually not explicitly assessed—are careful reading skills.

Context in Reading Tasks

Alderson and Cseresznyés (2003) described readers in a modern language examination as being actively engaged with a variety of texts similar to the ones they may potentially encounter in real life situations. Their reading process can be selective due to either the reading goal based on which they determine whether an idea or stretch of text is irrelevant or not, or because they have to skip unknown lexis. Moreover, the reading process is also understood to be flexible; that is, it varies according to the reading goals set by the task instructions. These reading purposes also aim to be authentic, which means that they simulate real life texts and the purposes for which they are normally read.

In modern examinations, readers are expected to use their language ability meaningfully in context. Therefore, much the same as in the case of the assessment of Use of English (UoE), the role of context is crucial for assessing reading ability. However, in the case of a UoE task, test takers have to read and understand a context (i.e., the text surrounding the item) that is markedly easier than in the case of a reading task. As Alderson and Cseresznyés (2003) stated, in a UoE task the context ″will normally not contain unknown words, apart from those being tested, and will not have complex structures, other than those that might be being tested″ (p. 27). In such a task, it is the meaningful use of a grammatical structure or lexical item that is in focus. In contrast, in a reading task the focus is on the use of various reading behaviors while processing a text, or on specific abilities such as reading critically and being able to differentiate fact from opinion or main idea from supporting detail. The context of the items in a reading test task can be below, at, or even above and, consequently, considerably more difficult than the assessed level. However, the part of the text targeted by the test item must be at the assessed level. Finally, it is also based on context that readers may be expected to make plausible inferences about meanings not explicitly expressed in the text.

Inferencing and Reading Comprehension

A simple definition of *inference* is ″any piece of information that is not explicitly stated in a text″ (McKoon & Ratcliff, 1992, p. 440). During the reading process, information present in a text can be condensed and information not explicitly present in the text can be added by readers. Information reduction as well as accretion processes (i.e., generation and construction) are controlled by macrorules and result in macropropositions (van Dijk, 1980), namely, inferences that reduce information in and add information to a text (see Kintsch, 1993). Both types of inference aid reading comprehension and the storage of information in short- and long-term memory (Kintsch & van Dijk, 1978).

Research evidence unequivocally revealed that comprehension was obstructed without the use of relatively simple bridging inferences (Singer, 1994). These backward inferences help create discourse coherence by relating an idea to the discourse that precedes it (e.g., the explicitly signaled anaphoric reference). Contrarily, elaborative or forward inferences are not a requirement for comprehension, but empirical research findings also demonstrated that they do improve it (Singer, 1994). Moreover, whereas deductive inferences are controlled by formal rules operating on information explicitly present in the text and result in verifiable presumptions, elaborative inferences, also referred to as pragmatic inferences, depend on the reader's world knowledge (e.g., schemata, scripts, frames, memory organization packages, or stereotypes; Graesser & Kreuz, 1993) and add probable information to a text, such as a prediction about the contents of the text based on its title or a conjecture concerning the writer's attitude (Schmalhofer et al., 2002; Singer, 1994; Singer & Lea, 2012).

Reading Assessment Task Features Affecting Comprehension

Two components of the reading assessment task, titles and images, function as advance organizers. When they are clear, age- and content-appropriate, and functional both in terms of thematic relevance and reproducibility through test booklet printing (Alderson & Cseresznyés, 2003; Tankó, 2005), both foretell the topic of the reading passage and by priming it help readers to relate the topic to their background knowledge. Meyer (1982) noted that a title has a signaling function as it ”prematurely reveals information abstracted from the content occurring later in the text" (p. 77). A title, therefore, aids comprehension as it facilitates the construction of a hypothetical coherent mental representation of the macrostructure (i.e., predicted content) of a text (Kintsch, 1988; Soederberg Miller & Stine-Morrow, 1998). Specifically, macrostructures achieve this effect by activating (i) relevant background information (i.e., schema) stored in long-term memory (Kintsch, 1998) and relevant vocabulary—both of which ease the comprehension of even less well composed texts—as well as (ii) cognitive frameworks that can be modified with new information (van Dijk & Kintsch, 1983). A number of research studies confirmed that on-line comprehension and recall (e.g., Bransford & Johnson, 1972; Miller et al., 2006; Smith & Swinney, 1992; Wiley & Rayner, 2000), reading time (e.g., Soederberg Miller & Stine-Morrow, 1998; Wiley & Rayner, 2000), processing of ambiguous words (Wiley & Rayner, 2000), and working memory demands (Miller et al., 2006) improve if a title is provided prior to reading and if it activates background knowledge. Naturally, what is implied here is that the title must be rhetorically functional, as it was intended by the writer of the text. Research evidence showed that altering or changing original text titles resulted in the construction of different mental representations of the same text (Bock, 1980); this should be avoided in assessment as it affects the justifiability of the results.

In a meta-analytic study of experimental research covering 33 years (i.e., 1985–2018), Guo et al. (2020) reported that graphics had a positive effect on reading comprehension. Notably, when graphic types were compared, pictures were found to have the most pronounced effect. However, pictures that were not organically related to the content of the text were also found to impact comprehension, albeit negatively (Wiley, 2019). A study on the effect of decorative pictures (i.e., those with a mainly aesthetic function) and instructional ones (i.e., those with an informative function) revealed that the participants paid little attention to decorative pictures and that such pictures had no effect on comprehension and learning. In contrast, informative pictures had an effect that, interestingly, was enhanced by the presence of decorative ones (Lenzner et al., 2013). In a similar study investigating readers' metacomprehension accuracy of expository science texts, Jaeger and Wiley (2014) found that decorative images negatively affected metacomprehension accuracy.

In addition to titles and images, the task instructions and the linguistic accuracy of the input also affect comprehension. When provided in the assessed language, the level of the instructions must not be higher than the assessed proficiency level (Bachman & Palmer, 2010). Furthermore, operational test tasks must be reviewed to ensure that the efficient processing of the input is not hindered by spelling errors, incorrect grammar, or formatting problems (Fulcher, 2010). Additionally, to help processing, the input must also be well-formed and correctly punctuated because, as Tankó (2022) noted, punctuation aids text processing and affects comprehension through its disambiguating function. Another reason why accurate input must be provided is that test takers are believed to learn even while taking a test (Bachman & Palmer, 2010).

Given that they affect comprehension, the task features discussed above jeopardize the justifiability of the interpretations made regarding reading ability. Justifiability being a validity issue, the last part of the review discusses construct validity.

Construct and Criterion-Related Validity

Messick (1995) differentiated between two major types of threats to construct validity that can occur simultaneously: (i) construct underrepresentation incurred by the narrow and therefore ungeneralizable assessment of a construct and (ii) construct-irrelevant variance induced by the broad assessment of a construct. The latter threat has two subtypes: construct-irrelevant difficulty, which is caused by chance factors unrelated to the measured construct that make the completion of a task difficult; and construct-irrelevant easiness, which is caused by task formats that allow a test taker to answer an item correctly without engaging the assessed construct or processes. Construct-irrelevant variance is especially important in the case of assessments where context is important as it ″matters whether the contextual clues that people respond to are construct-relevant or represent construct-irrelevant difficulty or easiness″ (p. 743).

Messick (1980) also described two types of criterion-related validity: ″concurrent validity and predictive validity, which differ respectively in terms of whether the test and criterion data were collected at the same time or at different times″ (p. 1016). Given that the advanced level examination certificate is used for making admission decisions to universities, the implied claim is that it indicates the test takers′ future level on the criterion, namely how well they will function in an English medium education context. The relationship between construct and criterion-related validity is that if the former is undermined, the latter collapses.

Purpose of the Present Study

The English major programs at Eötvös Loránd University are popular. In the autumn semesters of 2021 and 2022, 223 and 209 (*N* = 432) students were admitted to the MA in English teacher training programs in addition to the 270 and 265 (*N* = 535) students admitted to the BA in English program (Educational Department, 2022a). This means that close to half of the admitted students were not required to have a B2 level certificate in English, which is a problem in itself because the minimum proficiency level needed for academic purposes is B2 (Kirkland & Saunders, 1991). This can partly explain the low language ability of first year English major students that has been causing problems for both students and teachers. However, an additional concern is the increasing body of evidence indicating that the B2 level EFL school-leaving examination does not assess language ability as intended (e.g., Dávid, 2008; Szabó & Kiszely, 2010; Tankó & Andréka, 2021).

Given that high-stakes admission decisions are made on the basis of the advanced level EFL secondary school leaving certificate, the justifiability of the assessment needs to be investigated. As a consequence, this research study was carried out in order to analyze the reading paper of the B2 level EFL school-leaving examination. The research question it proposed to answer was the following: To what extent is the reading paper in the advanced level EFL school-leaving examination suitable for the level-appropriate assessment of the aspects of language ability that it intends to assess?

Methods

To answer the research question, a qualitative content analysis study was carried out to analyze reading papers from past EFL school-leaving examinations. The first section in this part gives a brief introduction to the EFL school-leaving examination. The second section describes the reading papers analyzed, and the last section presents a summary of the data analysis.

The EFL School-Leaving Examination

The EFL school-leaving examination is administered for secondary school students in Hungary twice a year, in May and in October, at two levels: The intermediate level is intended to be at level B1 and the advanced level at level B2 (Council of Europe, 2020). The examination consists of a written part (which includes Reading, Use of English, Listening, and Writing papers) administered in one sitting with a break and an oral part administered on a separate day after the written part.

A B2 level state-accredited language examination certificate is issued if a test taker achieves a minimum of 60% on both the written and oral parts of the advanced level EFL school-leaving examination. A test taker whose score is between 40%–59% receives a B1 level state-accredited language examination certificate (Hungarian Government, 2022). Admissions officers award additional points for those applying to tertiary university programs who hold a B2 level EFL school-leaving examination certificate stating that they have achieved a minimum score of 45% (Educational Department, 2022c).

The Reading Papers Analyzed

The advanced level reading papers investigated (representing the 2017–2020 period) were administered together with the Use of English paper in the first half of the written examination. Test takers had 70 minutes to complete it and—depending on the number of items—could get maximum 28 or 30 raw points, which were converted to 30 final points. The reading paper accounts for 25% of the total score for the written part, in which the four papers are equally weighted. For the successful completion of the written part of the EFL school-leaving examination, test takers must achieve a minimum of 12% (Hungarian Government, 2022). Therefore, assuming that test takers can pass the written part of the examination by scoring 12% on each paper, depending on the number of items in the paper, a minimum of three or four reading items must be answered correctly, the equivalent of merely four converted points, which is a disconcertingly low cut score.

Data Analysis

The first version of the codebook used in this study and a set of analytical decision rules were created based on the specifications available for the advanced level reading paper and the relevant literature on the reading construct and assessment. An example of an analytical decision based on the *"minimal effort necessary to solve the item correctly"* principle (Tankó & Andréka, 2021) is the one stating that in cases when the correct response to an item can be given based on a semantic, syntactic, or form matching decision, the decision type to be recorded is the one that requires the least effort. For example, in 2019-i-T1-I25 (i.e., year 2019, first take, task one, item 25), the co-text before the gap "These sites are built to be engaging, (25) _____ is addictive for others." clearly cues the option "... and what's engaging for some ...” because of the word present in both, so making the semantic lexical repetition link between “some” and “others” or the grammatical link with the repetition of the preposition “for” may only be needed as reassuring check possibilities.

Following this, a pilot sample of the tasks (25%) was solved by the two authors, the items were coded, and following a consensus-building discussion the codebook and analytical rules were revised. Then each author independently coded the items in the remaining papers. During the coding process, the codebook and the coding rules were updated when new coding issues emerged. The coding was conducted with the assistance of various tools, including the MS Excel software to count words and edit the coding form; the CEFR-based Vocabulary Level Analyzer (ver. 2.0; Uchida, 2022), which estimates the CEFR level of an input text; Multimodal Analysis Image, a software for image annotation and analysis (trial version; Tan et al., 2012); Textinspector, a web-based linguistic analysis tool that produces metrics benchmarked to the CEFR (Weblingua, 2022); and English Vocabulary Profile Online, which is a reference database based on the Cambridge Learner Corpus that assigns a CEFR level to the lexis in texts (Cambridge University Press & Assessment, 2015). The results of the coding were compared, codings which did not match were discussed, and a final set of jointly approved codings were created for analysis.

Reading Paper Test Specification

According to the detailed school-leaving examination specification (Ministry of Education, 2002) from which several of the categories were selected for the code book, the reading paper aims to measure test takers’ ability to read independently and comprehend various kinds of real-life authentic texts with the use of appropriate strategies and at the level of specificity appropriate to the set reading purposes. Although in the poorly organized specification this information is added in a seemingly random manner to a thematically unrelated section, the types of comprehension to be measured are global, selective, and detailed comprehension. It needs to be noted here that none of the constructs are defined, which raises questions about their operationalization. Furthermore, the specification not only fails to link the measured reading abilities with the types of comprehension named, but the listed abilities all denote reading activities that can only be achieved through global careful comprehension (see the taxonomy of Khalifa & Weir, 2009, described in the Theoretical Background section), which is a non sequitur. The reading paper is supposed to measure the test taker's ability to follow a train of thought, opinions, and arguments; understand information in sufficient detail (NB whether at the global or local level remains unspecified); and infer the writer's point of view as well as the feelings and emotions of the writer or characters (i.e., formulate elaborative inferences). Using Gray's (1960) phrasing, the paper aims to assess the test taker's ability to read *the lines* as well as *between the lines*.

The input is to be authentic (but may be edited), so it may contain "words, phrases and structures whose level exceeds that of the examination, *but which are not necessary for the successful completion of a task* [emphasis added]" (Ministry of Education, 2002, Advanced level examination section); in addition, the input should be straightforward in content; well-organized; concrete or abstract; and thematically suited for the experience and general interest of the age group. In terms of prior knowledge demands, it must be at the level of the general knowledge of a secondary school leaver, on a topic specified in the detailed requirements, and finally, level-appropriate as regards linguistic and content complexity. A variety of genres are also specified, ranging from user's manuals and newspaper articles to academic and fictional literature.

The specification names an impressive array of task types from which reading item writers can choose freely and which can be used in the reading paper in any combination. The list comprises matching (at least 14 subtypes), ordering (three subtypes), multiple choice, true/false/not stated statements, short answer questions, open or banked cloze, gapped summary, and grouping according to given categories task types.

The reading paper may consist of 3—4 tasks, each with an English language instruction and with one longer or several shorter input texts per task. The total input length must be between 1,300—1,500 words, and the paper must consist of 25—30 items.

Codebook

Due to the lack of attention to technical detail and incoherence problems, the specification summarized above had to be elaborated on the basis of the available literature on reading comprehension before a coding scheme could be designed. The codebook that was written for this study consists of two main parts (i.e., task characteristics and item characteristics) featuring eight and six variables, respectively. Of these, 10 were nominal and four were interval variables. The coding scheme, indicating measurement levels and offering brief descriptions of the variables together with information on whether the coding was human only or computer assisted, can be found in the Appendix.

Coder Agreement

To check the reliability of the coding, Cohen's κ reliability coefficient was used; if this was not possible, percentage agreement was calculated instead. There was perfect agreement (Landis & Koch, 1977) in two cases (Scope of relationship, $\kappa = .845$, 95% CI [.79 to .92], $p < .001$; CEFR level of title, $\kappa = .848$, 95% CI [.76 to .93], $p < .001$]), almost perfect agreement in three cases (Reading behavior type engaged–Category B, $\kappa = .858$, 95% CI [.80 to .92], $p < .001$; Linguistic decision required by response, $\kappa = .903$, 95% CI [.84 to .96], $p < .001$), and substantial agreement in two cases (Reading behavior type engaged–Category A, $\kappa = .747$, 95% CI [.67 to .82], $p < .001$; CEFR level of item, $\kappa = .633$, 95% CI [.56 to .70], $p < .001$). A high percentage agreement was found in three cases (Comprehension level, 99%; Task type, 97%; and Image-text intersemiotic sense relations, 91%). In the case of those variables where only computer-generated indices were used (i.e., Length per task/paper and CEFR level of input) or where there was 100% agreement between the coders (e.g., Number of items per task/paper), no intercoder reliability index was calculated.

Results and Discussion

The outcomes of the analysis are presented in this section according to the task and item characteristics variables investigated.

School-Leaving Examination Results

Within the 2017–2020 period investigated, altogether 60,691 secondary school students registered for the advanced level EFL school-leaving examination. From these, 59,976 took the examination and 57,686 (96%) passed (Educational Department, 2022b). Based on the analysis of the results of the test takers who had a reading score recorded, the majority of the test takers managed to receive fairly high scores on the reading paper (*N* = 59,976; *M* = 22, *Mdn* = 23, *SD* = 5.548; *Q1* = 19, *Q2* = 23, *Q3* = 26). Furthermore, altogether 63 students had a converted reading score of 4 points, and of these 35 (56%) passed the EFL school-leaving examination and became eligible—some (*n* = 4) with additional points awarded for a minimum of 45% achievement—for admission to English major programs offered in Hungary. This serves as evidence that a student with the minimum acceptable EFL reading score can become eligible for a tertiary English major program, but the low number of such cases found is moderately reassuring.

Task Characteristics

Task Types

The advanced level reading papers were selected for this study from the eight examinations administered in the 2017–2020 period. Each reading paper analyzed consisted of four tasks, and the number of items per paper ranged from 28 to 30 ($f_{30} = 5$, $f_{29} = 2$, $f_{28} = 1$). The task types included were matching sentence segments (i.e., clauses or phrases) to gaps and true/false/not stated (each $n = 7$, 22%); matching lexical items to gaps (i.e., open cloze), multiple choice, and matching sentence beginnings to ends (each $n = 4$, 13%), filling in a list of gapped sentences ($n = 2$, 6%) or a gapped summary ($n = 1$, 3%) based on the input; and matching complete sentences to gaps, paragraphs to gaps, or questions to answers (each $n = 1$, 3%).

In spite of the fact that on the basis of the specification nine main task types were included in the codebook, the reading papers only contained five of these. As could be expected given that the matching main task type had 12 subtypes, matching tasks were used most frequently in the reading papers. Of these, three (i.e., matching sentences/paragraphs to gaps and questions to answers) tested global text organization, namely coherence, which—although not irrelevant in terms of reading comprehension—is also tested in the writing and speaking parts, which should be sufficient for decision making. Instead, other task types like short-answer would contribute more relevant information for reading comprehension assessment and improve the generalizability of the results. The true/false/not stated task type was also frequent.

What is difficult to explain is the switch from gapped summaries ($n = 2$, 2018-i-T2, 2019-ii-T1) to gapped sentences (2020-i-T2) over the years. In the case of a gapped summary, the test taker reads a continuous text and contrasts its macrostructure with that of the input text. This task is cognitively more demanding and arguably much more authentic than comparing the content of sentences from a list to an input text. The cognitive load derives from the complexity and number of operations to be performed. Its authenticity becomes obvious if we consider the relationship between the headline-and-lead advance organizer dyad and the body of a news article, or between an abstract and the full text research paper. The headline and the lead together add up to a selective (also known as guided, Tankó, 2019) summary (Bell, 1998), whereas a research article abstract is a global summary of the paper. Both these tasks illustrate common, real life reading activities from the general and academic target language use domains. Furthermore, the variation in terms of the main task types used from 2017 to 2018 is considerable ($n_{2017} = 2$, $n_{2018} = 5$), which raises justifiability issues concerning consistency across different assessment administrations.

Length of the Input

According to the specification, the overall length of the input text must be between 1,300–1,500 words per reading paper. The average length of the input per paper was 1,470 words, with a narrow range of 1,434 to 1,495 words, which is consistent with the specification. The average length of the input per task was 367 words, with a large range of 292 to 461. The multiple-choice tasks, however, add a considerable reading load with their verbose options, leading to inconsistency in the amount of input to be processed across years. Furthermore, the amount of input to be processed in the test items also varied markedly within the multiple-choice tasks; in fact, it more than doubled in the 2020 spring task compared to 2019 (2019/i/T4, *n* = 158; 2018-ii-T3, *n* = 221; 2019-ii-T3, *n* = 270; 2020-i-T3, *n* = 330 words). This raises concerns in terms of the consistency of the assessment across different takes.

CEFR Level of the Input

The overall CEFR level of the input texts was assessed with the CEFR-based vocabulary level analyzer (Uchida, 2022). The levels were found to range from B1.1—the lowest level within the B1 band (Uchida & Negishi, 2018)—to C2, the highest defined in the CEFR. As Table 1 shows, the overall CEFR level of 20 (62%) reading input texts was above the B2 band level, while five (15%) were at levels below it.

Table 1

Reading Input Text CEFR Levels in the B2 Level Examination

Level	*f*	%	*cum %*
C2	11	34	34
C1	9	28	62
B2.1	5	16	78
B1.1	3	9	87
B1.2	2	6	94
B2.2	2	6	100
Total	32	100	

Trained item writers can construct B2 level reading comprehension items for an input text whose overall difficulty level exceeds the level of the examination. However, their job becomes challenging and maybe even impossible when they have to write items for input that is barely at the B1 level as they are not supposed to counterbalance the low difficulty level of the input with a high difficulty level item. In fact, the exact opposite is recommended (Alderson, 2000).

Input Text Titles

The titles of eight input texts (26%) were above B2 level ($n_{C1} = 4$; $n_{C2} = 4$) according to the coding rule which specified that the level of the highest CEFR level lexical item should be recorded as the indicator of overall title difficulty level. The rule was formulated with awareness of the fact that readers skip lexis they do not understand (Alderson & Cseresznyés, 2003); however, this is not exactly the case with titles, which are macropropositions with important discourse functions. As discussed in the review of the literature, understanding a title is important because it loads knowledge frames and activates vocabulary that enhances comprehension. The problem above could be mitigated with the inclusion of the topic of the input text in the task instructions. However, not all the instructions were found to do this (e.g., the discourse topic announcement is missing from the instruction of task 2017-i-T1), and it is common knowledge that—most likely due to a sense of security deriving from testwiseness acquired through classroom test preparation—most students do not read the instructions.

Since the level of some of the lexical items appearing in the titles could not be estimated with the English Vocabulary Profile, the analysis was most likely unable to reveal all the level-related problems with titles (e.g., an extreme instance of this is "THE **A1** PERILS **??** OF 'TABOO' **??** GIFTS **A2**;" 2017-ii-T2 eventually coded as A2, where *taboo* is guessable, so easy, but *perils* is more likely a C1 level item like *hazard* or *threat*). Nevertheless, the analysis did reveal several other issues of which the most important are presented here: One input text had no title at all (2020-ii-T1). Modifications of the original titles and functionally related components resulted in distorted discourse topic signaling. For example, the title "The owl thieves of Sweden" (2020-ii-T2) should introduce a text about cash not being used anymore. However, it fails to do so because a fully functional lead present in the original "As the country ditches cash, criminals turn to stealing owls" was deleted, which disconnected the title from the text and raised the difficulty level of the input in an inauthentic way. Another modification type compromised the macroproposition function of titles. For example, only the first three of the six paragraphs in the text entitled "What's in the queen's handbag?" (2017-i-T1) discuss what is in the handbag; the remaining ones provide explanations about the functions of the bag (e.g., signaling device). The input bears close similarity to an online article entitled "What's inside the Queen's handbag and why is it so significant?" (Hello! magazine), which—unlike the test task version—does anticipate the discussion of reasons. Furthermore, several titles contained mistakes introduced by item writers. A title which was originally "It's a WET wedding! Hero groom jumps into a river during photoshoot with the bride to save drowning boy" (Daily Mail Online) was changed to "Canada groom rescues boy from lake" (2018-i-T1)—note the incorrect use of a noun instead of the adjective. A punctuation mark such as a colon might have been intended to be added after the first word.

Because it is beyond the scope of this paper to discuss punctuation errors in detail, it can only be noted here that the instructions in 23 tasks contained one punctuation mistake, 68 punctuation mistakes were found in the body of the input texts (e.g., $n = 7$ in 2017-i-T4 and 2019-ii-T2 each), and 26 were found in the test items (e.g., $n = 5$ in 2019-ii-T3). Not only do the tasks become unduly difficult when punctuation cannot perform its text disambiguating function (Tankó, 2022), but it also potentially teaches test takers incorrect English use—if not during the test, then when teachers use the tasks in their classes.

Image-Text Intersemiotic Sense Relations

Each reading task featured an image. Some of these ($n = 11$, 34%) were acceptable as they set the context and potentially helped the activation of the schema necessary for comprehension. Such images illustrated the input (e.g., a picture of Christopher Marlow with a text about the playwright, 2018-ii-T2; or a picture of a hornet that illustrated the insect discussed in the text, 2019-ii-T3). Attempts made to illustrate more complex text content failed, and the remaining images were not functional because they were indiscernible due to their size (e.g., 2019-i-T3), quality (e.g., 2018-i-T4), or because by turning color images into black and white ones, important information was lost (e.g., 2020-ii-T4, where the image is supposed to be a heat map illustrating climate change with colors). Other images required age-inappropriate prior knowledge and failed to cue the discourse topic (e.g., 2017-i-T3). Instead of being informative, some provided irrelevant and misleading details (e.g., 2020-i-T1, where an image depicting a meeting held in the Whitehouse accompanied a text about bureaucracy in the UK, Austria, and companies in general).

Item Characteristics

Altogether 236 regular test items and 32 example items provided in the reading tasks were double-coded; the coding was finalized and the dataset analyzed. The number of items per task ranged from five to nine

(*M* = 7.38, *Mo* = 7, *SD* = 1.212), and there were 28 to 30 items in a reading paper (only three papers had less than 30 items: $n_{2017\text{-i}} = 28$, $n_{2018\text{-I} \,\&\, 2020\text{-ii}} = 29$), so the number of items per paper matches the test specification.

Comprehension Level

The results of the analysis showed that except for four items (1%), the reading papers tested literal comprehension. This does not match the specification to a desirable extent because, as summarized in the methodology section, the specification emphasizes that the reading paper assesses an extensive range of inference types. One of the inference items found in a True/False/Not stated task targeted the last paragraph of the text entitled "Three ways to train your brain to cope with heavy travel." The paragraph and the item are the following:

> **Input text paragraph #6**: **[(1)** If you feel sleepy during daylight hours when you first arrive somewhere new, try and do some aerobics.**]** Even if you do not feel tired in the evening, try to sleep anyway. **[(2)** And avoid drinking a coffee when you hit that wall in the afternoon.**]** Caffeine will only make the process much harder when it's time for bed. Smartphone use before bed is the ultimate no-no. The blue light emitted from it can trick your brain into thinking it's daytime and therefore block the production of the hormone melatonin, which would normally help you sleep.
>
> **Item 28**: Exercising or having coffee will have similar effects if you feel sleepy during the day. (2018-i-T4)

The idea that exercise helps reduce daytime sleepiness is implied only in Text Segment 1 because it does not explicitly state that exercise will wake up the jetlagged traveler. Nor does Text Segment 2 explicitly state that a coffee in the afternoon has the same effect—it also only cues this information, so it needs to be retrieved from prior knowledge. However, additional inferencing ability is required from the reader to understand the macro-level relationship between the item and the paragraph. The reader must infer the analogy implicitly present in the first part of the paragraph that the item targets: exercise and coffee will have the same positive effect during the day (but not in the evening). In order to answer such items, readers must combine information across sentences within a paragraph (i.e., engage in global reading).

Scope of the Relationship

Most of the items in the reading papers analyzed measured local comprehension ($n = 160$, 60%). This means that most items could be answered by reading individual sentences with little need to take into consideration the context. Given that global reading (which requires the construction of a coherent meaning representation across sentences) is more cognitively demanding, the relatively low frequency of items engaging global reading behaviors may make the reading paper easier than it is intended to be.

Reading Skills and Strategies

The ratio of items that engaged the test-taker's reading skills versus their strategies also reflects the narrow scope of most of the items discussed in relation to the previous variable. The majority of the reading items required the use of skills ($n_{\text{Skill}} = 170$, 63%), some necessitated the joint deployment of strategies and skills ($n_{\text{Strat. \& Skill}} = 97$; 36%), and there was very little emphasis on strategies alone ($n_{\text{Strat.}} = 1$; 4%). Therefore, it can be concluded that most items measured careful reading skills, namely, the parsing of lexis and syntactic structures. Given that these behavior types are processes, the coders categorized the items based on the *"minimal effort necessary to solve the item correctly"* principle (Tankó & Andréka, 2021) that they applied as they solved the tasks themselves. For more pertinent insights, actual performance data and information about actual test taker's comprehension processes would be needed as in the case of the next variable.

Expeditious and Careful Reading Behaviors

The findings about expeditious and careful reading behaviors confirm the general overemphasis on careful reading in the reading papers discussed in relation to the previous variable. All the items could be answered by engaging careful local and global reading behaviors ($n_{\text{Car. Local/Within sentence}} = 160, 60\%$; $n_{\text{Car. Global/Across sentences}} = 93, 35\%$; $n_{\text{Car. Global/Across paragraphs}} = 14, 5\%$; $n_{\text{Car. Global/All text}} = 1, 4\%$). The fact that all the items could be solved with careful reading is disquieting in light of the narrow scope of most items and with respect to the specification according to which the reading paper measures selective reading ability. Admittedly, it could be argued that any item measuring any type of expeditious reading behavior can be answered with careful reading, providing there is sufficient time given for the test taker to substitute skimming, scanning, and search reading with careful local and global reading, but this is unlikely to apply to most test takers under the time constraints of the examination.

CEFR Levels of the Items

In spite of the computer assisted coding, the most difficult variable to code was the one involving the assessment of the CEFR levels of the reading items. Most of the problems were caused by the lexical items for which no CEFR levels were available. Once again, the rule followed during coding was that the level of a test item was to be recorded according to the level of the highest CEFR level lexical unit it contained or according to the lowest if a list of solutions was provided in the marking key, the latter based once again on the *"minimal effort necessary to solve the item correctly"* principle (Tankó & Andréka, 2021). The use of the rule is justified given that the complete comprehension of a well-written reading test item is necessary for a correct response; otherwise, most likely the assessor will be faced with construct irrelevant variance issues. Partly for this reason, the rule was not applied in the case of those lexical items that could be guessed easily based on L1 knowledge (e.g., "pyramid," C1, 2017-i-T2, "piramis" in Hungarian; "clichés," C2, 2018-ii-T1, "klisé" in Hungarian).

In such cases, the second highest CEFR level was recorded for the item. Any instance of the use of a proper name in the item and cases when an item could be answered with one word for which no estimated CEFR level was available (e.g., "archery," 2019-1-T2) was coded as "NA." The results are summarized in Table 2.

Table 2

Reading Item CEFR Levels in the B2 Level Examination

Level	*f*	%	*cum %*
B2	112	42	42
B1	53	20	62
C1	42	16	77
C2	23	9	86
A2	21	8	94
A1	13	5	98
NA*	4	1	100
Total	32	100	

*Reading test item containing a lexical unit that was a proper name or a one-word lexical unit whose estimated CEFL level was not known.

Given that 25% of the items contained C1 and C2 level lexis, these items were above the level intended to be measured by the examination. Since text difficulty is best predicted by vocabulary difficulty (Alderson, 2000), the responses given to these items provided more information about the difficulty of the task induced by the item than about the state of the test takers' reading ability. This increased level of construct complexity most likely resulted in construct-irrelevant difficulty and does not match the specification as these lexical units are necessary for the successful completion of a task. The level of the items below B2 level could actually be optimal provided the input content they targeted was of the level of difficulty intended to be measured by the examination.

Linguistic Decisions Required by the Response

The analysis of the types of decisions based on which a correct response to an item could be given showed that the majority of the items required a semantic decision ($n_{\text{Sem.}} = 207$, 77%). The second most frequent decision type required the combination of form and meaning cues ($n_{\text{Sem.}} = 28$, 10%). This means that one expeditious reading strategy behavior, scanning, could be used—even if infrequently—to respond to an item as the lexis in the item and the input were identical in form, which allows for string search. Semantic decisions, enhanced by syntactic cues, represented the third and almost equally frequent decision type ($n_{\text{Sem. \& [Syntax]}} = 23$, 9%). The low frequency of such items is actually reassuring because the reading construct, as defined in the specification, does not include syntactic ability. It is also encouraging that only two instances were found when the reverse applied, and five instances when the decision was equally informed by semantic and syntactic cues. There were three instances when the test taker had to rely on inference, using the generalization ($n = 2$) and construction macrorules ($n = 1$) for decision making. One of these is reproduced here:

> Input text sentence: The winning team is the one that completes a catch over the furthest distance, with no breakage.
>
> Item 28: Winning the championship depends on only ____________ basic criteria. (2017-i-T4)

To answer Item 28 correctly, the test taker had to infer (1) that the largest distance an egg travels in the air and (2) that it does not break are the two criteria based on which the winner can be found, and then construct a macroproposition using the generalization rule. Additional items triggering the use of these more complex decision-making mechanisms necessary for the construction of an integrated representation of the content of an input text are required for an enhanced construct representation in reading ability assessment.

Conclusion

The analysis of the high-stakes advanced level EFL school-leaving examination reading papers selected for the current study revealed a number of construct validity issues that call into question the generalizability and predictive validity of the results of the reading paper and by extension that of the entire examination—especially if the findings published on the Use of English paper are also taken into consideration (Tankó & Andréka, 2021).

Based on the main problems found in terms of construct underrepresentation, it can be concluded that the reading paper samples the construct that it is intended to measure in a markedly narrow way. In addition, the range of operational task types is poor. The items basically only test literal comprehension and disproportionately target local comprehension, making ineffective use of context. The majority of the items primarily require the use of reading skills rather than strategies or the use of inference or expeditious local reading behaviors. It is to be noted that the official test specification addresses all of these constructs in more or less detail.

The reading paper was also found to be lacking in terms of construct-irrelevant variance. Construct-irrelevant difficulty was induced partly by a lack of consistency in terms of overall length of the reading text (i.e., input text and items) due to specific task types not used in each paper. This resulted in a lack of equivalence between test forms and compromised consistency across different groups of test-takers.

The linguistic and non-linguistic weaknesses of the task context generated by the instructions, images, and titles of the reading tasks were additional sources of construct-irrelevant difficulty. The fact that a quarter of the test task items (i.e., multiple-choice items or sentences to be matched to the text) contained C1 and C2 level lexis which was likely necessary for the successful completion of the tasks were further causes of construct-irrelevant difficulty. Finally, the substantially lower CEFR level of some of the input texts than the level intended to be measured by the examination resulted in both construct-irrelevant easiness—due to inappropriately easy input text selection during the development of these assessment tasks—and to construct-irrelevant difficulty induced by the exceedingly difficult test task items written to counterbalance the low difficulty level of the input. The disconcertingly low cut score established for the paper further aggravates the problems, and in spite of its apparent beneficial consequences to the test takers, it affects secondary school leavers—and thus potential English major students—negatively.

The direct effect of the school leaving examination on those secondary school leavers who become English majors is that, contrary to the goals set for the two-level school-leaving examination system, it fails to aid them in the pursuit of their tertiary level studies. Specifically in terms of reading ability, it does not benefit—as it should—only those prospective university students who are able to access information efficiently for study purposes. At university, students have to read long, complex texts and combine content extracted from these texts across paragraphs and texts by using their full range of expeditious and careful reading behaviors both globally and locally (see Tankó, 2019; Weir et al., 2000). The indirect effect of the school-leaving examination on the same stakeholder group is that while it may facilitate their mobility for study purposes or entry into the labor market through certification, because of its low generalizability and predictive validity it ultimately does a disservice to those who do not have the functional competencies required by either of these domains.

The limitations of the study are that the CEFR level of each test task item could not be determined with the right level of accuracy because an estimated CEFR level was not available in the English Vocabulary Profile database for every lexical unit used in the analyzed test items. Moreover, more pertinent insights could be gained with actual performance data and information about test taker's comprehension processes. With the above limitations considered, a crucial practical outcome of the study is that it will be easier to improve the examination with the help of the now identified and explained shortcomings of the specifications and operational reading papers.

Note: This study was conducted as a pilot study for the research project (K 142536) financed by the Hungarian Scientific Research Fund (OTKA).

References

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

Alderson, J. C., & Cseresznyés, M. (2003). *Into Europe: Reading and use of English*. Teleki László Foundation.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Bell, A. (1998). The discourse structure of news stories. In A. Bell & P. Garrett (Eds.), *Approaches to media discourse* (pp. 64–104). Blackwell Publishers.

Bock, M. (1980). Some effects of titles on building and recalling text structures. *Discourse Processes*, *3*(4), 301–311. https://doi.org/10.1080/01638538009544494

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning & Verbal Behavior*, *11*, 717–726. https://doi.org/10.1016/S0022-5371(72)80006-9

Cambridge University Press & Assessment (2015). *EnglishProfile—The CEFR for English: English vocabulary profile online:* [Software]. https://www.englishprofile.org/wordlists/evp

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

Dávid, G. (2008). *Az emelt szintű idegen nyelvi érettségi és az államilag elismert nyelvvizsgák a vizsgázói teljesítmények tükrében: Összegző tanulmány, bővített változat* [The advanced level foreign language school-leaving examinations and the state accredited language examinations in the light of the candidates' performance: A synthesis study—extended version]. Nyelvvizsgáztatási Akkreditációs Központ.

Educational Department (2022a). *Elmúlt évek felvételi statisztikái* [University admission statistics of recent years]. https://www.felvi.hu/felveteli/ponthatarok_statisztikak/elmult_evek/!ElmultEvek/index.php/elmult_evek_statisztikai/

Educational Department (2022b). *Statisztikák, vizsgaeredmények* [Statistical data and school leaving examination results]. https://www.ketszintu.hu/publicstat.php

Educational Department (2022c). *Felsőoktatási felvételi tájékoztató - 2022. szeptemberben induló képzések* [Higher education admission bulletin for September 2022]. https://www.felvi.hu/felveteli/jelentkezes/korabbi_elj_archivum/felveteli_tajekoztatok/FFT_2022A

Fulcher, G. (2010). *Practical language testing* (1st ed.). Routledge.

Graesser, A. C., & Kreuz, R. J. (1993). A theory of inference generation during text comprehension. *Discourse Processes*, *16*(1–2), 145–160. https://doi.org/10.1080/01638539309544833

Gray, W. S. (1960). The major aspects of reading. In H. Robinson (Ed.), *Sequential development of reading abilities* (Vol. 90, pp. 8–24). Chicago University Press.

Guo, D., Zhang, S., Wright, K. L., & McTigue, E. M. (2020). Do you get the picture? A meta-analysis of the effect of graphics on reading comprehension. *AERA Open*, *6*(1). https://doi.org/10.1177/2332858420901696

Hungarian Central Statistical Office (2022). 23.1.1.16. *Idegen nyelvet tanulók a középfokú iskolákban* [Foreign language learners in secondary schools]. https://www.ksh.hu/stadat_files/okt/hu/okt0016.html?utm_source=kshhu&utm_medium=banner&utm_campaign=theme-oktatas

Hungarian Government (2020). *A Kormány 5/2020. (I. 31.) Korm. Rendelete a Nemzeti alaptanterv kiadásáról, bevezetéséről és alkalmazásáról szóló 110/2012. (VI. 4.) Korm. rendelet módosításáról* [Government decree 5/2020. (I.31.) on the modification of Government decree 1010/2012. (VI.4.) on the issuing, introduction and application of the new National Curriculum]. https://magyarkozlony.hu/dokumentumok/3288b6548a740b9c8daf918a399a0bed1985db0f/megtekintes

Hungarian Government (2021). *A Kormány 65/2021. (II. 15.) Korm. Rendelete egyes köznevelési és szakképzési tárgyú, valamint a jelnyelvoktatói névjegyzék vezetésével összefüggő kormányrendeletek módosításáról* [Government Decree amending the government decrees on public education, vocational training and management of the list of sign language teachers]. Magyar Közlöny, 2021. február 15.

Hungarian Government (2022). A Kormány 100/1997. (VI. 13.) *Korm. rendelete az érettségi vizsga vizsgaszabályzatának kiadásáról* [Government Decree 100/1997. (VI. 13.) on the issuance of the examination regulations for the school leaving examination]. https://njt.hu/jogszabaly/1997-100-20-22

Illés, É. (2011). A szövegértés pragmatikája: Egy érettségi feladat elemzése [The pragmatics of text comprehension: The analysis of a school-leaving examination task]. *Iskolakultúra*, *21*(4–5), 144–156. http://www.iskolakultura.hu/index.php/iskolakultura/article/view/21150

Jaeger, A. J., & Wiley, J. (2014). Do illustrations help or harm metacomprehension accuracy? *Learning & Instruction*, *34*, 58–73. https://doi.org/10.1016/j.learninstruc.2014.08.002

Khalifa, H., & Weir, C. J. (2009). *Examining reading: Research and practice in assessing second language reading*. Cambridge University Press.

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

Kintsch, W. (1993). Information accretion and reduction in text processing: Inferences. *Discourse Processes*, *16*, 193–202. https://doi.org/10.1080/01638539309544837

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363

Kirkland, M. R., & Saunders, M. A. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly*, *25*(1), 105–121. https://doi.org/10.2307/3587030

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. https://doi.org/10.2307/2529310

Lenzner, A., Schnotz, W., & Müller, A. (2013). The role of decorative pictures in learning. *Instructional Science*, *41*, 811–831. http://www.jstor.org/stable/43575400

McKoon, G., & Ratcliff, R. (1992). Inference during reading. *Psychological Review*, *99*(3), 440–466. https://doi.org/10.1037/0033-295x.99.3.440

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*(11), 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

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

Meyer, B. J. F. (1982). Signalling the structure of text. In D. H. Jonassen (Ed.), *The technology of text* (pp. 64–89). Educational Technology Publications.

Miller, L. M., Cohen, J. A., & Wingfield, A. (2006). Contextual knowledge reduces demands on working memory during reading. *Memory and Cognition*, *34*(6), 1355–1367. https://doi.org/10.3758/bf03193277

Ministry of Education (2002). *40/2002. (V. 24.) OM rendelet az érettségi vizsga részletes követelményeiről* [Ministry of Education Decree on the detailed requirements of the secondary school leaving examination]. https://njt.hu/jogszabaly/2002-40-20-45

Schmalhofer, F., McDaniel, M. A., & Keefe, D. (2002). A unified model for predictive and bridging inferences. *Discourse Processes*, *33*(2), 105–132. https://doi.org/10.1207/S15326950DP3302_01

Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 479–515). Academic Press.

Singer, M., & Lea, R. B. (2012). Inference and reasoning in discourse comprehension. In Schmid, H.-J. & Geeraerts, D. (Eds.), *Handbook of cognitive pragmatics* (pp. 85–119). Mouton de Gruyter.

Smith, E. E., & Swinney, D. (1992). The role of schemas in reading text: A real-time examination. *Discourse Processes*, *15*(3), 303–316. https://doi.org/10.1080/01638539209544814

Soederberg Miller, L. M., & Stine-Morrow, E. A. L. (1998) Aging and the effects of knowledge on on-line reading strategies. *Journal of Gerontology: Psychological Sciences*, *53B*(4), 223–233. https://doi.org/10.1093/geronb/53B.4.P223

Szabó, G., & Kiszely, Z. (2010). Államilag elismert nyelvvizsgarendszerek, illetve az emelt szintű érettségi összevetése próbavizsgázói teljesítmények tükrében német és angol nyelvből [A comparison of the state accredited language examination systems and the advanced level school-leaving examinations in German and English in the light of mock test performance.]. *Modern Nyelvoktatás*, *16*(4), 19–38.

Szabó, G. (2019). A szövegnehézség vizsgálata az emelt szintű angol érettségi olvasáskomponensében [Examining text difficulty in the reading component of the advanced level school-leaving exam in English]. *Modern Nyelvoktatás*, *25*(3–4), 102–119.

Tan, S., E., M, K. L. E., & O' Halloran, K. (2012). *Multimodal analysis image* [Software]. Multimodal Analysis Company. https://multimodal-analysis.com/products/multimodal-analysis-image/index.html

Tankó, G. (2005). *Into Europe: The writing handbook*. Teleki László Foundation.

Tankó, G. (2019). *Paraphrasing, summarising and synthesising skills for academic writers: Theory and practice* (Rev. ed.). Eötvös University Press.

Tankó, G. (2022). *Professional writing: The academic context* (Rev. 2nd ed.). Eötvös University Press.

Tankó, G., & Andréka, Z. (2021). Probing the advanced level EFL school-leaving examination: The Use of English paper. In G. Tankó & K. Csizér, (Eds.), *DEAL 2021: Current explorations in English applied linguistics* (pp. 65–105). Eötvös Loránd University. https://edit.elte.hu/xmlui/handle/10831/62555

Uchida, S. (2022). *CVLA: CEFR-based vocabulary level analyzer* (ver. 2.0) [Software]. https://cvla.langedu.jp/

Uchida, S., & Negishi, M. (2018). Assigning CEFR-J levels to English texts based on textual features. *In Y. Tono & H. Isahara (Eds.), Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC 2018)*, 463–467. APCLA.

Urquhart, A. H., & Weir, C. J. *(*1998). *Reading in a second language: Process, product and practice*. Longman.

Van Dijk, T. (1980). *Macrostructures: An interdisciplinary study of global structures in discourse interaction and cognition*. Erlbaum.

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* Academic Press.

Weblingua (2022). *Textinspector* [Software]. https://textinspector.com/

Weir, C. J., Huizhong, Y., & Yan, J. (2000). *An empirical investigation of the componentiality of L2 reading in English for academic purposes* (Studies in Language Testing, Vol. 12). UCLES/Cambridge University Press.

Wiley, J. (2019). Picture this! Effects of photographs, diagrams, animations, and sketching on learning and beliefs about learning from a geoscience text. *Applied Cognitive Psychology*, *33*(1), 9–19. https://doi.org/10.1002/acp.3495

Wiley, J., & Rayner, K. (2000). Effects of titles on the processing of text and lexically ambiguous words: Evidence from eye movements. *Memory & Cognition*, *28*(6), 1011–1021. https://doi.org/10.3758/BF03209349

Appendix

Reading Assessment Task Coding Scheme

TASK CHARACTERISTICS

- Task type [nominal]: *20 task types identified in the test specification (e.g., short answer, gapped summary, multiple choice) and subdivided into nine main types of which one, matching, had 12 subtypes (e.g., banked cloze)*
- Length per task [interval; computer assisted coding]: *total number of words in a complete input text (i.e., reconstructed text with the title also included)*
- Length per paper [interval; computer assisted coding]: *total number of words in all the complete input texts within one paper*
- Number of items per task [interval]: *total number of items in a task*
- Number of items per paper [interval]: *total number of items in a paper*
- CEFR level of input [nominal; computer assisted coding]: *A1–C2 (Uchida, 2022)*
- CEFR level of title [nominal; computer assisted coding]: *A1–C2 (CUP & Assessment 2015; Weblingua, 2022)*
- Image-text intersemiotic sense relations [nominal; computer assisted coding]: *four relationship types (e.g., illustration, contrast) (Tan et al., 2012)*

ITEM CHARACTERISTICS

- Comprehension level [nominal]: *literal / inference (Gray, 1960; Khalifa & Weir, 2009; Schmalhofer et al., 2002; Singer & Lea, 2012)*
- Scope of relationship [nominal]: *global-broad / local-narrow (Bachman & Palmer, 2010; Khalifa & Weir, 2009; Urquhart & Weir, 1998)*
- Reading behavior type engaged–Category A [nominal]: *strategy / skill (Bachman & Palmer, 2010; Khalifa & Weir, 2009; Urquhart & Weir, 1998)*

- Reading behavior type engaged–Category B [nominal]: *expeditious / careful (Khalifa & Weir, 2009)*
- CEFR level of item [nominal; computer assisted coding]: *A1–C2 (CUP & Assessment 2015; Weblingua, 2022)*
- Linguistic decision required by response [nominal]: *8 subtypes (e.g., correct answer can be given based on a semantic or syntactic decision, or a combination of the two is needed; or if a superordinate term is generated) (van Dijk, 1980; van Dijk & Kintsch, 1983)*