



Calibrating Difficulty Across Lower-Level ELT Coursebooks: A Coh-Metrix Analysis of Texts by Major International Publishers*

Andrew Garth (Dankook University)



This is an open-access article distributed under the terms of the Creative Commons License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received: July 14, 2026
Revised: August 1, 2026
Accepted: August 18, 2026

Garth, Andrew
Assistant Professor, Department
of English
Dankook University
119 Dandae-ro, Anseo-dong,
Cheonan-si, Chungnam, Korea
Tel: +82-41-550-3130
Email: andy.garth@gmail.com

* The present research was supported by the research fund of Dankook University in 2025.

ABSTRACT

Garth, Andrew. 2026. Calibrating difficulty across lower-level ELT coursebooks: A Coh-Metrix analysis of texts by major international publishers. *Korean Journal of English Language and Linguistics* 26, 1098-1116.

The current study investigated how textual difficulty progresses across internationally published English language teaching coursebooks designed for lower-level learners. In total, 1,154 reading passages from 60 coursebooks were examined at the beginner, elementary, and pre-intermediate levels, corresponding respectively to levels A1, A2, and B1 of the Common European Framework of Reference for Languages. Nine Coh-Metrix indices representing text length, syntactic complexity, lexical diversity, lexical sophistication, referential cohesion, connective cohesion, and semantic coherence were measured. Linear mixed-effects models were employed to examine differences between adjacent proficiency levels (i.e., A1-A2 and A2-B1) while accounting for variation across publishers, series, and textbooks. The results indicated that textual difficulty progressed selectively rather than uniformly across the lower levels of the Common European Framework of Reference for Languages. Text length, sentence length, and lexical diversity increased consistently across both transitions and accumulated to large differences from A1 to B1, whereas modifiers per noun phrase, referential cohesion, and semantic coherence did not change at any point. Furthermore, direct comparison of the two transitions found no significant difference in magnitude for any index, indicating that where difficulty increased, it increased by a comparable amount at each step.

KEYWORDS

Coh-Metrix, ELT reading passages, international ELT publishers, lower-level learners

1. Introduction

Reading plays a central role in second language (L2) development by providing learners with opportunities to acquire vocabulary, grammatical structures, and discourse knowledge through meaningful language input (Grabe 2009). For reading materials to effectively support language development, the linguistic demands of texts should increase in a manner that is both gradual and developmentally appropriate (Tomlinson and Masuhara 2018). This principle is reflected in influential theories of second language acquisition, such as Vygotsky's (1978) Zone of Proximal Development and Krashen's (1985) Input Hypothesis, both of which emphasize the importance of providing learners with input that extends beyond their current abilities while remaining sufficiently comprehensible to facilitate learning. If textual difficulty increases too rapidly, learners may struggle to construct meaning from the text. Conversely, if increases in textual difficulty are insufficient, learners may receive inadequate exposure to the increasingly complex linguistic features needed for continued development (Krashen 1982). Consequently, understanding how textual difficulty progresses across proficiency levels is essential for understanding how linguistic and discourse features are introduced to learners as their reading proficiency develops. To be clear, textual difficulty refers to the linguistic and discourse features of a text that are associated with greater processing demands, rather than to reading difficulty as experienced by a particular learner, which also depends on proficiency, background knowledge, and purpose for reading.

The progression of textual difficulty is particularly important during the early stages of language learning (Grabe and Stoller 2020). As beginning learners develop their reading proficiency, they gradually acquire the lexical knowledge, syntactic processing skills, and discourse comprehension abilities that provide the foundation for more advanced reading (Grabe 2009). Accordingly, the linguistic and discourse demands of reading materials are expected to increase as learners' reading proficiency develops while remaining sufficiently comprehensible to facilitate continued learning. Examining how textual features change across these early stages of language development can provide valuable insights into how reading materials introduce progressively greater linguistic and discourse demands as learners' reading proficiency develops.

Fortunately, advances in computational text analysis now make it possible to examine a wide range of linguistic and discourse features simultaneously. In particular, the Coh-Metrix program enables researchers to analyze texts across dimensions such as lexical sophistication, syntactic complexity, cohesion, readability, and referential overlap (McNamara et al. 2014). Although previous research has examined differences among textbooks designed for learners at different proficiency levels, much of this work has focused on reading passages from materials intended for intermediate-level learners and above (Cheng and Chang 2022, Hanifah et al. 2022, Ryu and Jeon 2020a, 2020b), leaving comparatively little attention devoted to textbooks designed for lower-level learners. Consequently, relatively little is known about how textual difficulty progresses during the early stages of L2 reading development. Furthermore, most previous studies have analyzed regionally marketed textbooks designed for specific learner populations, such as Korean learners (Ryu and Jeon 2020b) or Taiwanese learners (Cheng and Chang 2022). Although these studies provide valuable insight, analyses of internationally marketed textbooks offer a broader perspective on how textual features associated with reading difficulty progress across proficiency levels in widely used English language teaching (ELT) materials that are potentially less influenced by the needs of specific regional learner populations.

Accordingly, the present study examines the progression of textual features associated with reading difficulty across internationally published ELT coursebooks designed for lower-level learners. By analyzing a range of linguistic and discourse features, the study seeks to identify which linguistic and discourse features distinguish reading passages across lower proficiency levels in international ELT coursebooks.

2. Background Literature

2.1 Reading Passages in ELT Coursebooks and Lower-Proficiency Learners

ELT coursebooks occupy a central and well-documented role in language learning, serving as the primary vehicle through which curriculum goals, language input, and pedagogical tasks are delivered to learners (Tomlinson and Masuhara 2018). One of the most consequential dimensions of textbook design is the calibration of difficulty. This concern is grounded in two foundational theoretical frameworks in second language acquisition. First, Vygotsky's (1978) notion of the Zone of Proximal Development (ZPD) posits that learning is optimized when learners are challenged just beyond their current independent capacity, with appropriate support. In addition, Krashen's (1985) Input Hypothesis similarly argues that comprehensible input that is slightly beyond a learner's current level, which has been famously formalized as $i+1$, is the essential condition for acquisition to occur. Together, these frameworks underscore a widely held assumption in ELT materials development that the difficulty of learning materials should be calibrated to learners' proficiency levels and should increase progressively as proficiency develops. Although textbooks often grade vocabulary and grammatical structures using word-frequency data and grammatical complexity or order of acquisition (Richards 2001), it is increasingly recognized that reading input difficulty is an equally important dimension of textbook design (McNamara et al. 2014).

While providing progressively difficult textual input is clearly necessary for more advanced L2 learners, it is especially important for lower-level learners and their developing reading competency (Byrnes 1998). Before addressing the textual input provided to lower-level learners, a brief outline of lower-level learners and their L2 reading capabilities is necessary. The Council of Europe's (2001) Common European Framework of Reference for Languages (CEFR) outlines six 'bands' or levels of L2 proficiencies from A1 and A2 (labeled as 'Basic Users'), to B1 and B2 ('Independent Users'), to C1 and C2 ('Proficient Users'). While the CEFR's *Companion Volume* (2020) describes learners' capabilities across many aspects of language development (e.g., listening, writing, and speaking) of all six proficiency bands, its descriptions of reading capabilities for the lower three proficiency levels are most pertinent to the current study. To be clear, although A1 and A2 are formally categorized as basic users, the inclusion of B1 is also necessary because the transition from A2 to B1 represents the final stage of lower-level reading development before learners move beyond the basic proficiency band. Examining both the A1-A2 and A2-B1 transitions, therefore, provides a more complete picture of how reading texts differ across the full lower-level range, from materials for novice readers to those at the threshold of independent reading. Despite the CEFR's lengthy list of descriptors, the A1, A2, and B1 reading capabilities can be summarized as follows.

A1 readers are able to understand very short, simple texts containing familiar names, words, and basic phrases, provided the material relates to everyday situations and is presented in a clear and predictable manner. Reading at this level relies heavily on high-frequency language, concrete meanings, and explicit textual support, with learners primarily extracting straightforward information rather than integrating ideas across extended passages. A2 readers are able to understand short, simple texts on familiar topics, including personal correspondence, notices, and straightforward factual materials. Although they can process more connected discourse than A1 learners, comprehension remains largely dependent on familiar vocabulary, transparent organization, and explicitly signaled relationships between ideas. B1 readers are expected to understand texts consisting mainly of high-frequency language that describe events, feelings, wishes, and opinions, as well as straightforward factual texts related to their personal, academic, or professional interests. At this level, learners are increasingly capable

of following extended discourse, connecting information across sentences, and constructing meaning even when ideas are presented with less explicit linguistic support than at the basic user levels (Council of Europe 2020).

In summary, the progression from A1 to A2 and from A2 to B1 reflects a gradual expansion in learners' reading capabilities rather than simply an increase in vocabulary knowledge. As learners develop, they become capable of reading longer and more conceptually demanding texts, understanding increasingly complex sentence structures, integrating information across sentences, and constructing meaning with progressively less explicit linguistic support. They also become better able to comprehend a broader range of vocabulary, including words that are less frequent, acquired later in language development, and more abstract in meaning. Overall, the CEFR reading descriptors portray reading development as a multidimensional process in which learners gradually acquire the ability to comprehend texts that place increasingly greater linguistic, syntactic, lexical, and discourse-level demands on the reader. These descriptors characterize the reading capabilities associated with each proficiency band rather than the process by which individual learners move between bands. In the present study they serve as a framework for interpreting the levels to which publishers have assigned their coursebooks, and the analysis concerns the reading materials designated for each level rather than the development of any learner over time.

Having outlined the reading capabilities described at each of the lower proficiency levels, attention now turns to the learning materials designed to support that development. Importantly, the CEFR is not merely a descriptive framework of learner proficiency. It has been widely adopted in language education systems and assessment frameworks (North 2014), and major international publishers have developed numerous ELT coursebook series explicitly aligned with its proficiency levels (Mishan 2022). Indeed, many coursebooks identify their intended proficiency level (e.g., A1, A2, or B1) in the title itself. Consequently, these materials represent reading input that has been intentionally designed to correspond to learners' developing reading abilities. Therefore, examining texts across successive CEFR levels provides an opportunity to investigate how publishers adjust linguistic, lexical, syntactic, and discourse characteristics to reflect the progression of learner proficiency.

2.2 Coh-Metrix

Considering the aforementioned importance of reading input and in recognition of the need to analyze and calibrate textual difficulty, a number of analytical tools have been developed. While readability indices such as the Flesch-Kincaid Grade Level and alternative software tools such as WordSmith Tools, Range, and VocabProfile have contributed useful insights (Crossley et al. 2011, Nation 2013), these tools predominantly operate at the word or sentence level. In other words, they measure surface-level features such as word length, word frequency, and mean sentence length without capturing the deeper cognitive and discourse-level properties that can impact how difficult a text is to process (McNamara et al. 2014). Fortunately, a more sophisticated computational tool, known as Coh-Metrix, has been developed that affords considerably richer analyses of textual difficulty.

Developed at the University of Memphis, Coh-Metrix is a computational text analysis tool that calculates a wide range of linguistic and discourse indices from written texts (McNamara et al. 2014). These indices capture linguistic and discourse characteristics that are widely recognized as important contributors to reading comprehension and textual difficulty, including lexical sophistication, syntactic complexity, and cohesion (Graesser et al. 2011). Because Coh-Metrix applies the same computational procedures to every text, it provides an objective and reproducible means of characterizing and comparing linguistic and discourse characteristics

across texts, provided the same software version and input texts are used (Graesser et al. 2004). To be clear, previous research (Tomlinson 2012) has found that graded ELT materials intentionally employ controlled vocabulary and simplified language, which may result in some indices behaving differently than they would with authentic texts. Nevertheless, Coh-Metrix has been widely adopted in readability, discourse, and educational research and provides a useful framework for characterizing and comparing the linguistic features that contribute to textual difficulty in textbook reading passages (Crossley et al. 2008). In short, Coh-Metrix is a tool for providing measurable indices that represent a number of factors. Although the list of factors being found to impact textual difficulty is extensive, several have received special attention such as text length, syntactic complexity, lexical diversity, lexical sophistication, referential cohesion, connective cohesion, and semantic coherence (Graesser et al. 2014).

2.3 Previous Studies on Textual Difficulty in ELT Materials

Although the investigation of textual difficulty in language teaching materials was historically underexplored relative to vocabulary and grammar analyses, a body of research has now emerged examining this question. Of particular relevance to the present study, several studies have investigated reading passages across different coursebook levels to determine which linguistic and discourse features change as proficiency level increases. Rather than treating readability or textual difficulty as a single global construct (e.g., the Flesch-Kincaid Grade Level), these studies examine how individual linguistic and discourse features contribute to the progression of textual difficulty across proficiency levels. In doing so, they provide insight into how textbook reading passages are graded to develop learners' proficiency, consistent with the principles underlying Vygotsky's (1978) Zone of Proximal Development and Krashen's (1985) Input Hypothesis.

As previously mentioned, a number of studies have focused on textbooks developed by regional publishers or used within specific national educational contexts. Lee (2013), for example, used Coh-Metrix to compare the reading passages in Korean high school English I and English II textbooks. Although the analysis found significant differences in only a small number of indices, including text length, semantic cohesion, and readability, the findings suggested that progression between textbook levels was less systematic than was expected. Building on this work, Ryu and Jeon (2020b) used a substantially broader range of Coh-Metrix indices to analyze the difficulty of reading and listening materials across Korean middle school English textbooks. They found that higher-grade textbooks generally contained longer texts, more sophisticated vocabulary, greater syntactic complexity, and changes in several cohesion-related features, although many Coh-Metrix indices remained stable across grades. Similar findings have been reported in analyses of national textbook series in Indonesia (Hanifah et al. 2022) and Taiwan (Cheng and Chang 2022), where increases in textbook difficulty were likewise achieved through changes in selected lexical, syntactic, and discourse features rather than through uniform increases across all dimensions of textual complexity.

Despite their contributions, these studies have almost exclusively examined textbooks developed for specific national curricula or regional educational markets. Consequently, the findings may reflect the priorities and constraints of particular educational systems and local publishing contexts rather than broader principles of ELT materials development. By contrast, major international publishers, such as Oxford University Press, Cambridge University Press, and Pearson, produce coursebook series that are marketed globally and collectively used by millions of learners across diverse educational settings. Therefore, examining the textual difficulty of these internationally distributed materials addresses an important gap in the literature and offers insights with broad relevance for ELT coursebook design and development.

To date, only one study has examined textual difficulty specifically in the reading materials of internationally published ELT textbooks using Coh-Metrix: Plakans and Bilki (2016). Their study investigated reading passages from 27 ELT textbooks across beginning, intermediate, and advanced levels, using Coh-Metrix to compare textual features via Multivariate Analysis of Variance (MANOVA). The results indicated that although the books differed greatly in terms of text length as well as cohesion-related indices (e.g., referential cohesion and connectives), there were no meaningful differences across a number of other constructs, such as syntactic similarity, syntactic complexity, and lexical complexity. However, Plakans and Bilki's (2016) study, while pioneering, is subject to an important limitation that constrains the conclusions that can be drawn from it. Simply put, the study does not appear to control for consistency across series/publishers. In terms of series, while there are four cases (and 12 books) of a full series being used such as *Quest 1*, *Quest 2*, and *Quest 3*, the remaining books do not appear to use full series. For example, while *Tapestry 2* and *Tapestry 3* are analyzed for the intermediate and advanced levels, the book *For Your Information* appears to have taken the place of *Tapestry 1* within the beginner book corpus. In some cases, the book that does not appear to conform to other books in a series is part of the same publishing house, while in other cases, the book is produced by a different publisher. Thus, without controlling for publisher and series, it is difficult to determine the extent to which the observed differences reflect intended progression across proficiency levels rather than inherent differences among textbook series or publishing practices. Consequently, some of the reported effects may be attributable to characteristics of individual textbook series or publishers rather than to proficiency level itself.

Furthermore, in addition to the general lack of research involving internationally available textbooks, very little previous research has focused on the textual difficulty of lower-level learners. While Ryu and Jeon's (2020b) study focused on books designed for Korean middle school learners of a presumably pre-intermediate proficiency, Hanifah et al. (2022) and Cheng and Chang (2022) focused on coursebooks for more advanced upper-secondary English L2 learners. While Plakans and Bilki's (2016) examination of beginner, intermediate, and advanced levels helped to elucidate differences between beginner books and more advanced coursebooks, the wide gap between those proficiency levels and the exclusion of books at intervening levels (e.g., pre-intermediate) leaves the question of how book series progress in different facets of textual difficulty unresolved. As previously mentioned, while understanding the input difficulty of more proficient learners is valuable, lower-level learners' development is of special importance.

As a whole, previous research suggests that textual difficulty in ELT materials develops through selective changes in linguistic and discourse features rather than uniform increases across all dimensions. Studies using tools such as Coh-Metrix have provided valuable insight into how features such as lexical sophistication, syntactic complexity, and cohesion vary across textbook levels, but this work has largely focused on regionally produced materials and higher-proficiency learners. As a result, relatively little is known about how internationally published ELT coursebooks calibrate reading passage difficulty. Moreover, despite the foundational importance of early reading development, the progression of textual difficulty for lower-level learners remains underexamined. Addressing these gaps is essential for understanding how contemporary ELT materials support the development of reading proficiency at the earliest stages of language learning.

3. Current Study

The present study is designed to address the previously outlined limitations of the aforementioned related studies by answering the following central research question:

Which linguistic and discourse features distinguish reading passages across lower proficiency levels in international ELT coursebooks?

Considering that the progression through A1-A2 and A2-B1 (of the CEFR guidelines) represents the progression of ELT reading input within lower-level learners' learning process, the current study is framed specifically to examine the increases in textual factors influencing difficulty during those two progressions. Furthermore, because these features may not operate independently of one another, a supplemental correlation analysis was also conducted to examine how they relate to each other.

Fortunately, many major international ELT publishers (e.g., Oxford or Macmillan) produce a full series of books that correspond to the proficiency levels designated by the CEFR (Mishan 2022). Thus, these series provide examples of textual input that can be analyzed to determine which textual features progress in difficulty throughout a lower-level learner's L2 educational process in reading (i.e., A1-A2 and A2-B1). Furthermore, as previously summarized, the descriptors of the A1, A2, and B1 reading competencies provided by the CEFR imply numerous changes in difficulty in terms of various constructs. As previously outlined, A1 learners work with very short, highly supported texts and mostly familiar, concrete vocabulary, which suggests the importance of analyzing text length and lexical sophistication. At A2, readers are expected to handle slightly longer texts, identify main ideas, and infer unfamiliar words from context, making lexical diversity, syntactic complexity, and cohesion increasingly relevant. By B1, learners must follow lines of reasoning, interpret logical and temporal connectors, and integrate information across paragraphs, which makes referential cohesion, connective cohesion, and semantic coherence especially important. Taken together, these competencies justify examining a set of constructs that capture how reading passages gradually become longer, more lexically demanding, more syntactically complex, and more discourse-based across proficiency levels.

Fortunately, the Coh-Metrix tool provides measurable indices that are related to the previously cited constructs depicted in the CEFR descriptors of A1, A2, and B1 levels. By selecting and measuring indices that represent such constructs, the current research can help to examine how such factors progress in difficulty along the development of lower-level learners. In doing so, it seeks to identify which dimensions of textual difficulty change across proficiency levels and which remain relatively stable, thereby providing a more comprehensive understanding of how lower-level reading materials are graded in internationally published ELT coursebooks.

3.1 Coursebook Selection and Text Collection

The first step was the selection of publishers and series. Five prominent international ELT publishers that often differentiate their coursebook series along CEFR levels (Oxford University Press, Cambridge University Press, Macmillan Education, Pearson, and National Geographic Learning) were identified, and their current catalogs were examined to determine which series contained books corresponding discretely to the A1, A2, and B1 bands. In many cases, level alignment was explicit, with individual books bearing the CEFR designation in their title or series documentation. In other cases, symbols, badges, or other indications of CEFR bands were often included. Still, in other cases, publisher catalogs clarified the CEFR levels corresponding to the books in their series. From the pool of eligible series identified for each publisher, four series were selected at random, resulting in four series per publisher and three books per series (one at each of the three target proficiency levels). With five publishers, four series each, and three books per series, the total corpus comprised 60 coursebooks. A full list of the included books is provided in the Appendix.

Having established the corpus of 60 coursebooks, the relevant reading passages were identified and collected

throughout each volume. Passages were included if their corresponding activity was explicitly titled ‘reading,’ ‘reading skills,’ or a functionally equivalent label. Additionally, activities in which reading was predicated as a primary instructional goal, as indicated by titles such as ‘reading to write’ or ‘reading and listening,’ were also included, provided that the activity featured a sustained expository, narrative, or dialogic passage. In total, this process yielded 327 texts from the 20 A1 coursebooks, 396 texts from the 20 A2 coursebooks, and 431 texts from the 20 B1 coursebooks. Following collection, all passages were transcribed into plain text (.txt) format in preparation for analysis using Coh-Metrix. The Appendix also references how many passages pertain to each book, series, and publisher. Although passage count was relatively similar per publisher, ranging from 208 to 259 total passages, the total passage counts were not evenly distributed across series and books, ranging from 8 to 38 passages per book and 33 to 93 passages per series. However, this imbalance was not associated with proficiency level. Neither publisher nor series was significantly associated with level (publisher: $\chi^2(8) = 5.96$, $p = .651$; series: $\chi^2(38) = 38.24$, $p = .458$), indicating that heavily represented series contributed comparably at all three levels. Furthermore, the linear mixed-effects models described below accommodate unequal group sizes and partition variance attributable to publisher, series, and book from variance attributable to proficiency level.

3.2 Data Analysis

Table 1. Coh-Metrix Indices Included in Analysis

Construct	Variable Measured	Interpretation	Coh-Metrix Index
Text Length	Text word count	Higher values indicate longer passages, which correspond to greater general reading difficulty.	DESWC
Syntactic Complexity	Mean sentence length	Higher values indicate longer sentences, which generally increase syntactic processing demands.	DESSL
	Mean number of modifiers per noun phrase	Higher values indicate more elaborated noun phrases, which generally increase syntactic complexity.	SYNNP
Lexical Diversity	Measure of lexical diversity	Higher values indicate greater vocabulary diversity, which generally increases lexical demands.	LDMTLDa
Lexical Sophistication	Mean age of acquisition for content words	Higher values indicate later-acquired vocabulary, which generally increases lexical difficulty.	WRDAoAc
	Mean concreteness for content words	Higher values indicate more concrete vocabulary; lower values reflect more abstract vocabulary and greater lexical difficulty.	WRDCNCc
Referential Cohesion	Mean argument overlap (adjacent sentences)	Higher values indicate greater overlap of key concepts between sentences; lower values increase the difficulty of maintaining coherence across the text.	CRFAO1
Connective Cohesion	All connectives incidence	Higher values indicate more explicit connections between ideas; lower values increase the difficulty of identifying relationships between ideas.	CNCAII
Semantic Coherence	Mean LSA overlap (all sentences in paragraph)	Higher values indicate stronger semantic links between sentences; lower values increase inferencing demands and make the text more difficult to integrate.	LSASSp

Rather than incorporating the full range of available measures, the present study selected a subset guided by the constructs identified in the CEFR as central to reading proficiency development. Because the previously

outlined descriptors for reading competency at the A1, A2, and B1 bands imply progression in text length, syntactic complexity, lexical diversity, lexical sophistication, and several distinct facets of cohesion and coherence, one or two representative indices measurable within Coh-Metrix (version 3.0) were selected for each construct. Text length was represented by text word count (TWC). Syntactic complexity was assessed using mean sentence length (MSL) and the mean number of modifiers per noun phrase (MOD-NP), capturing sentence-level and phrase-level processing demands, respectively. Lexical diversity was represented by the Measure of Textual Lexical Diversity (MTLD), a robust measure that is relatively insensitive to text length. Lexical sophistication was assessed using word age of acquisition (WRD-AoA) and word concreteness (WRD-CONC), reflecting the developmental level and the conceptual abstractness of the vocabulary, respectively. Cohesion was operationalized through two separate constructs: referential cohesion (REF-COH), represented by argument overlap between adjacent sentences, and connective cohesion (CON-COH), represented by the incidence of all connectives which capture the repetition of referents across sentences and the explicit signaling of logical relationships between ideas, respectively. Finally, semantic coherence (SEM-COH) was represented by latent semantic analysis (LSA) overlap across all sentences within a paragraph, reflecting the degree of global semantic connectedness within the text. Table 1 above displays the constructs outlined above along with their corresponding Coh-Metrix indices that are included in the study as well as a brief explanation about how results correspond to textual difficulty.

To assess how these features varied across proficiency levels, linear mixed-effects models were used. Reading passages were nested within textbooks, which were themselves nested within series, which were in turn nested within publishers, meaning that passages drawn from the same book, series, or publisher could not be treated as fully independent observations. A separate linear mixed-effects model was fit for each of the nine indices, with proficiency level (i.e., A1, A2, B1) entered as a fixed effect and random intercepts specified for publisher, series, and book to account for this nested structure. For two indices, MTLD and TWC, the full random-effects structure produced convergence warnings. Therefore, the publisher random intercept was removed for these indices, which resolved the convergence issues and produced substantively unchanged results.

All three pairwise comparisons among proficiency levels (i.e., A1 vs. A2, A2 vs. B1, and A1 vs. B1) were computed for each model, and corrections for multiple comparisons were applied using the Holm method across the three comparisons within each model. Because the Holm-correction approaches have no directly corresponding interval estimate, confidence intervals are reported unadjusted throughout and are intended as indications of precision rather than as tests, with statistical significance determined by the Holm-corrected *p*-values. Because the study's aims concerned developmental progression between consecutive proficiency levels, interpretation focuses on the two adjacent-level contrasts, with the A1-B1 comparison reported as a supplementary indication of cumulative change across the full lower range. Effect sizes are reported for the pairwise comparisons. The effect sizes of pairwise differences are quantified using Cohen's *d*, calculated as the mean difference divided by the pooled standard deviation. For the mixed-effects models, the pooled standard deviation is defined as the square root of the summed book, series, publisher, and residual variance components, following the approach recommended by Westfall et al. (2014).

Because the study's main aims concerned characterizing the two adjacent transitions (i.e., A1-A2 and A2-B1), whether there was any difference in magnitude was also examined. This was necessary because a difference in statistical significance between two contrasts does not itself constitute evidence that those contrasts differ (Gelman and Stern 2006). Therefore, an additional contrast was computed for each model. This quadratic contrast, specified with the weight vector (1, -2, 1) across the three proficiency levels, tests the null hypothesis that the increment from A1 to A2 is equal to the increment from A2 to B1. Holm corrections were applied across

the nine indices. Effect sizes for this contrast are reported as Δd , which were computed using the same pooled standard deviation as the pairwise contrasts.

Lastly, to further examine the relationships among the nine indices, Pearson correlations were also calculated at the book level ($N = 60$), with each index averaged across all passages within a book to avoid the non-independence present in the passage-level data. Because books from all three proficiency levels are pooled in such an analysis, any two indices that both increase across levels will correlate for that reason alone. Therefore, correlations were computed twice. One correlational analysis was computed on the book means, and another analysis after centering each index within proficiency level, which removes between-level differences and leaves only variation among books at the same level. Because each correlation matrix involved 36 pairwise tests, Holm corrections were applied across the full family of comparisons within each matrix.

4. Results

As previously stated, linear mixed-effects models were applied separately for each of the nine Coh-Metrix indices across seven constructs of textual complexity, with proficiency level (A1, A2, B1) as a fixed effect and random intercepts for publisher, series, and book to account for the nested structure of the corpus (passages within books, books within series, series within publishers). Full model results, including omnibus tests, Holm-corrected pairwise comparisons for adjacent levels (A1-A2, A2-B1), and effect sizes, are reported in Table 2. To be clear, A1-B1 results are also included but are reported for completeness rather than as a focus of the present analysis, since all three comparisons formed part of the same Holm-corrected family within each model. They are drawn on later only as a supplementary indication of cumulative change across the full lower range.

Text length, one measurement of syntactic complexity (MSL), and lexical diversity emerged as the most consistent markers of progression across proficiency levels. The number of words increased significantly and substantially at both transitions, with similar increases from A1 to A2 ($d = 0.60$) and A2 to B1 ($d = 0.63$). Lexical diversity also increased significantly at both transitions, although the gain was noticeably larger from A1 to A2 ($d = 0.59$) than from A2 to B1 ($d = 0.36$). Interestingly, regarding the two measures of syntactic complexity, mean sentence length followed a similar trajectory ($d = 0.63$ for A1-A2, $d = 0.55$ for A2-B1) of comparable significant increases in difficulty across both the A1-A2 and A2-B1 transitions, while the number of modifiers per noun phrase showed no significant pairwise differences at either A1-A2 or A2-B1 transition. This divergence suggests that the progression of syntactic difficulty from A1 to A2 is driven primarily by increasing sentence length rather than by greater phrasal density within noun phrases.

The remaining constructs showed more selective patterns. Word age of acquisition increased significantly from A1 to A2 ($d = 0.33$, $p = .027$) but not from A2 to B1 ($p = .080$), whereas word concreteness showed the reverse timing, remaining flat between A1 and A2 ($p = .509$) before decreasing sharply and significantly from A2 to B1 ($d = -0.43$, $p < .001$). The two indices thus reached significance at different transitions. Whether this reflected a genuine difference in magnitude was tested directly using the quadratic contrasts reported in Table 3, which compare the magnitude of the A1-A2 and A2-B1 increments for each index. Analogous to word age of acquisition, connective incidence increased significantly from A1 to A2 ($d = 0.28$, $p = .018$) but did not increase significantly from A2 to B1 ($p = .245$).

Table 2. Linear Mixed-Effects Model Results of Coh-Metrix Indices by CEFR Level

Construct	Variable	<i>M (SD)</i> by level	<i>F (2,df)</i>	<i>p</i>	Contrast: <i>b (d), p</i>	<i>R</i> ² _m / <i>R</i> ² _c
Text Length	TWC	A1: 149.6 (63.9)	54.0 (36.9)	< .001*	A1-A2: 57.1 (0.60), < .001*	.198 / .477
		A2: 198.1 (86.5)			A2-B1: 59.4 (0.63), < .001*	
		B1: 264.9 (109.8)			A1-B1: 116.5 (1.23), < .001*	
Syntactic Complexity	MSL	A1: 10.8 (3.1)	44.7 (38.9)	< .001*	A1-A2: 1.97 (0.63), < .001*	.182 / .402
		A2: 12.8 (3.0)			A2-B1: 1.72 (0.55), < .001*	
		B1: 14.5 (3.1)			A1-B1: 3.69 (1.17), < .001*	
	MOD-NP	A1: 0.7 (0.2)	2.61 (40.4)	.086	A1-A2: 0.048 (0.21), .094	.007 / .219
		A2: 0.8 (0.2)			A2-B1: -0.014 (-0.06), .519	
		B1: 0.8 (0.2)			A1-B1: 0.035 (0.15), .229	
Lexical Diversity	MTLD	A1: 61.4 (25.0)	43.5 (39.3)	< .001*	A1-A2: 14.51 (0.59), < .001*	.126 / .375
		A2: 75.6 (25.5)			A2-B1: 8.99 (0.36), < .001*	
		B1: 83.3 (22.5)			A1-B1: 23.50 (0.96), < .001*	
Lexical Sophistication	WRD-AoA	A1: 290.4 (43.8)	9.65 (40.0)	< .001*	A1-A2: 15.4 (0.33), .027*	.048 / .495
		A2: 307.7 (47.9)			A2-B1: 10.6 (0.23), .080	
		B1: 316.5 (41.3)			A1-B1: 26.0 (0.56), < .001*	
	WRD-CONC	A1: 402.6 (31.9)	20.2 (34.3)	< .001*	A1-A2: -1.81 (-0.06), .509	.048 / .113
		A2: 400.8 (31.1)			A2-B1: -13.42 (-0.43), < .001*	
		B1: 386.8 (30.0)			A1-B1: -15.23 (-0.49), < .001*	
Referential Cohesion	REF-COH	A1: 0.5 (0.2)	2.90 (32.9)	.069	A1-A2: 0.032 (0.15), .178	.006 / .047
		A2: 0.5 (0.2)			A2-B1: 0.010 (0.05), .558	
		B1: 0.5 (0.2)			A1-B1: 0.042 (0.19), .076	
Connective Cohesion	CON-COH	A1: 82.5 (26.7)	7.98 (36.3)	.001*	A1-A2: 6.55 (0.28), .018*	.026 / .100
		A2: 89.6 (23.2)			A2-B1: 2.73 (0.12), .245	
		B1: 92.4 (20.2)			A1-B1: 9.28 (0.40), .001*	
Semantic Coherence	SEM-COH	A1: 0.2 (0.1)	1.82 (38.0)	.175	A1-A2: 0.018 (0.21), .203	.007 / .113
		A2: 0.2 (0.1)			A2-B1: -0.012 (-0.13), .457	
		B1: 0.2 (0.1)			A1-B1: 0.007 (0.08), .493	

Note. Pairwise *p*-values are Holm-corrected across the three level comparisons within each model. **p* < .05; *R*²_m is the variance explained by proficiency level; *R*²_c is the variance explained by the full model including random effects.

By contrast, neither referential cohesion nor semantic coherence exhibited meaningful progression across the three proficiency levels. The omnibus tests for both constructs were non-significant, and neither adjacent-level comparison reached statistical significance. These findings indicate that, unlike lexical sophistication, text length, or sentence length, discourse-level connectedness remained remarkably stable across the A1-B1 sequence. Taken together, the results suggest that progression in these internationally published ELT coursebooks is driven primarily by changes to lexical and sentence-level characteristics rather than by adjustments to cohesion or semantic coherence between sentences.

Figure 1 below displays Cohen's *d* with 95% confidence intervals for the A1-A2 and A2-B1 progressions across all nine indices, allowing the magnitude of each change to be compared directly. At the A1-A2 transition, the largest effects are in text length (*d* = 0.60, 95% CI [0.37, 0.84]), mean sentence length (*d* = 0.63, CI [0.38, 0.87]), and lexical diversity (*d* = 0.59, CI [0.39, 0.80]), all falling in the moderate range and tightly enough estimated that values below 0.35 are implausible for any of the three. In practical terms, an effect of this size means that a randomly chosen A2 passage is longer, built from longer sentences, and drawn from a wider vocabulary than a randomly chosen A1 passage roughly two-thirds of the time. A second tier of appreciably

smaller effects appears in word age of acquisition ($d = 0.33$, CI [0.08, 0.58]) and connective cohesion ($d = 0.28$, CI [0.08, 0.48]), each roughly half the magnitude of the first tier. The remaining indices show effects near zero, with word concreteness ($d = -0.06$), referential cohesion ($d = 0.15$), and semantic coherence ($d = 0.21$) all small enough that their intervals accommodate change in either direction.

At the A2-B1 transition, text length ($d = 0.63$, CI [0.40, 0.86]) and sentence length ($d = 0.55$, CI [0.31, 0.79]) sustain effects of comparable size to the earlier transition, while lexical diversity weakens to roughly two-thirds of its earlier magnitude ($d = 0.36$, CI [0.17, 0.56]). Word concreteness produces the only substantial negative effect in the figure ($d = -0.43$, CI [-0.60, -0.27]). However, it should be noted that since lower concreteness indicates more abstract vocabulary, this represents an increase in difficulty of magnitude comparable to the lexical diversity increase. Word age of acquisition ($d = 0.23$, CI [-0.02, 0.48]) and connective cohesion ($d = 0.12$, CI [-0.08, 0.31]) show effects smaller than at the earlier transition but still positive in direction, with intervals that extend well above zero and only marginally below it. As a whole, Figure 1 shows a narrow band of effect magnitudes. Text length, sentence length, and lexical diversity produce the largest effects at both transitions, clustering between $d = 0.36$ and 0.63 . Modifiers per noun phrase, referential cohesion, and semantic coherence remain below $d = 0.25$ at both transitions and in both directions. No index reaches a large effect by conventional standards ($d \geq 0.8$), and the largest single contrast is $d = 0.63$. The differences these coursebooks build in at each step are therefore consistent and detectable, but moderate in size.

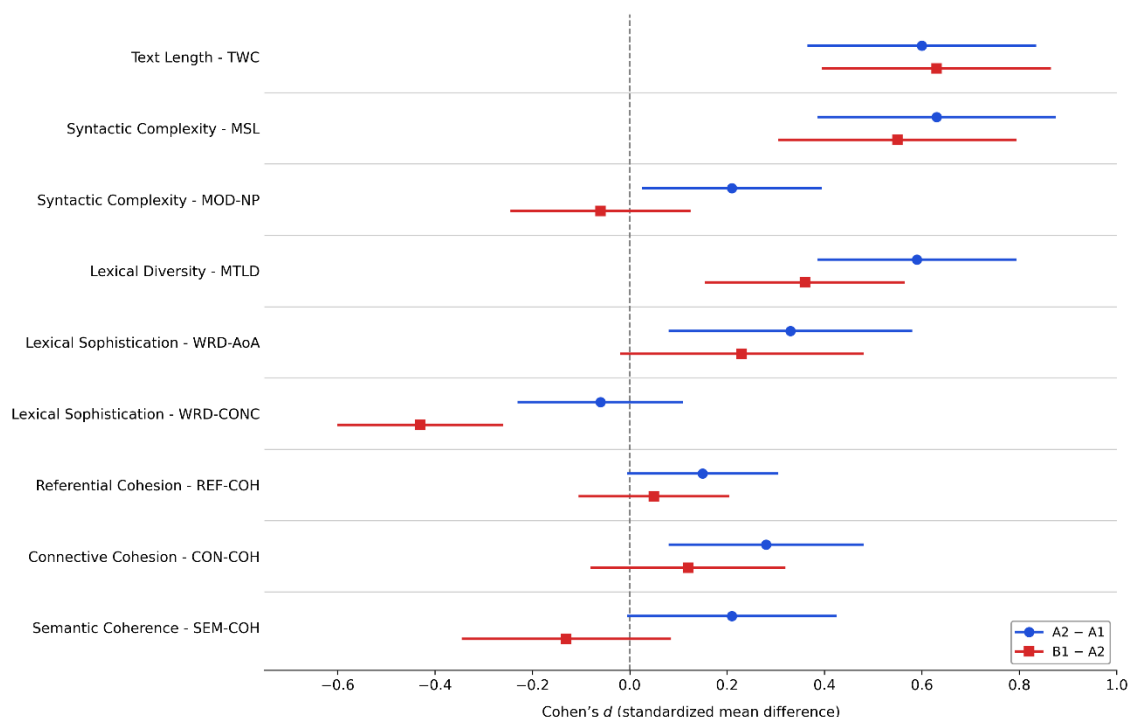


Figure 1. Effect Sizes for CEFR-Level Comparisons Across Measured Coh-Metrix Indices

Note. Points represent Cohen's d ; bars represent 95% confidence intervals (unadjusted) for each individual contrast.

The preceding results identify which indices distinguish adjacent proficiency levels, but they do not establish that the two transitions differ from one another. In short, knowing that an index increased significantly from A1 to A2 but not from A2 to B1 does not establish that the first increase was larger than the second. This is because

a contrast (e.g., A1-A2) may fall on either side of a significance threshold without its magnitude differing meaningfully from another contrast (e.g., A2-B1) (Gelman and Stern 2006). Therefore, testing whether the transitions differ requires comparing the A1-A2 and A2-B1 contrasts directly rather than comparing their significance. To do this, a quadratic contrast was computed for each model, and the results are displayed below in Table 3.

In short, no construct reached significance after Holm correction in this analysis. The largest difference was observed for word concreteness, where the decline from A2 to B1 exceeded the change from A1 to A2 ($\Delta d = -0.37$, 95% CI [-0.67, -0.07], uncorrected $p = .016$, $p = .142$ after Holm correction) with this being the only contrast whose confidence interval excluded zero. For the remaining eight indices, the intervals for Δd comfortably included zero, including those for word age of acquisition ($\Delta d = -0.10$, CI [-0.55, 0.35]) and connective incidence ($\Delta d = -0.16$, CI [-0.52, 0.20]) which were the two indices that reached significance at A1-A2 but not at A2-B1. Therefore, where difficulty increased, it did so by a statistically comparable amount at each transition. To be clear, it should be noted that the quadratic contrast is estimated less precisely than the individual pairwise contrasts. Therefore, these results indicate that the two transitions cannot be distinguished at the present sample size, not that they are equivalent.

Table 3. Tests of whether A1-A2 and A2-B1 Transitions Differ in Magnitude

Construct	Variable	d (A1-A2)	d (A2-B1)	Δd	95% CI for Δd	t	df	p (Holm)
Text Length	TWC	0.60	0.63	0.03	[-0.39, 0.45]	0.15	36.9	1.000
Syntactic Complexity	MSL	0.63	0.55	-0.08	[-0.51, 0.35]	-0.37	38.9	1.000
Lexical Diversity	MOD-NP	0.21	-0.06	-0.27	[-0.60, 0.06]	-1.65	40.4	.752
Lexical Sophistication	MTLD	0.59	0.36	-0.23	[-0.59, 0.13]	-1.28	39.3	1.000
	WRD-AoA	0.33	0.23	-0.10	[-0.55, 0.35]	-0.45	40.0	1.000
	WRD-CONC	-0.06	-0.43	-0.37	[-0.67, -0.07]	-2.54	34.3	.142
Referential Cohesion	REF-COH	0.15	0.05	-0.10	[-0.39, 0.19]	-0.70	32.9	1.000
Connective Cohesion	CON-COH	0.28	0.12	-0.16	[-0.52, 0.20]	-0.90	36.3	1.000
Semantic Coherence	SEM-COH	0.21	-0.13	-0.34	[-0.73, 0.05]	-1.77	38.0	.682

Note. $\Delta d = d$ (A2-B1) - d (A1-A2); Confidence intervals are unadjusted; p-values are Holm-corrected across the nine contrasts. No contrast is significant at $\alpha = .05$ after correction.

To examine relationships among the nine indices, Pearson correlations were calculated at the book level ($N = 60$), with each index averaged across all passages within a book to avoid the non-independence present in the passage-level data. Because books from all three proficiency levels are pooled in such an analysis, any two indices that both increase across levels will correlate for that reason alone. Therefore, correlations were computed twice with one analysis on the book means (as seen in the left-hand chart below in Figure 2) and another analysis after centering each index within proficiency level, which removes between-level differences and leaves only variation among books at the same level (as seen in the right-hand chart below in Figure 2). Holm corrections were applied across the 36 pairwise comparisons within each matrix.

There are a few noteworthy findings via the two correlation analyses in Figure 2. First, these coursebooks adjust text length, sentence length, and lexical diversity independently of one another. The three indices appeared to form a coherent group when books from all levels were analyzed together ($r = .50$ to $.63$, all significant after correction), but the group dissolved once level was removed ($r = .21$ to $.38$, none significant). In short, longer passages are not linked with longer sentences or a wider vocabulary. Therefore, these three indices

represent separable means of adjusting difficulty rather than three symptoms of one. A second set of indices, namely modifiers per noun phrase, word age of acquisition, and semantic coherence, is related for reasons unconnected to proficiency level. These three remained strongly intercorrelated after centering (MOD-NP/WRD-AoA: $r = .82$ pooled and $.83$ within-level; MOD-NP/SEM-COH: $.69$ and $.68$; WRD-AoA/SEM-COH: $.55$ and $.56$), and none progressed consistently across levels in the mixed-effects models. Because their association holds among books at the same level, it suggests a source other than CEFR grading. One possible explanation is that publishers and series differ in stylistic conventions, with some publishers employing denser, more expository prose at every level. However, it should be clearly stated that this is a plausible explanation rather than a confirmed finding. A third pattern was visible only within levels with cohesion declining as passages become longer and lexically denser, but only within levels. Furthermore, while referential cohesion was negatively associated with word age of acquisition ($r = -.50$) and lexical diversity ($r = -.42$), and semantic coherence with text length ($r = -.44$), none of these three relationships was detectable when levels were pooled.

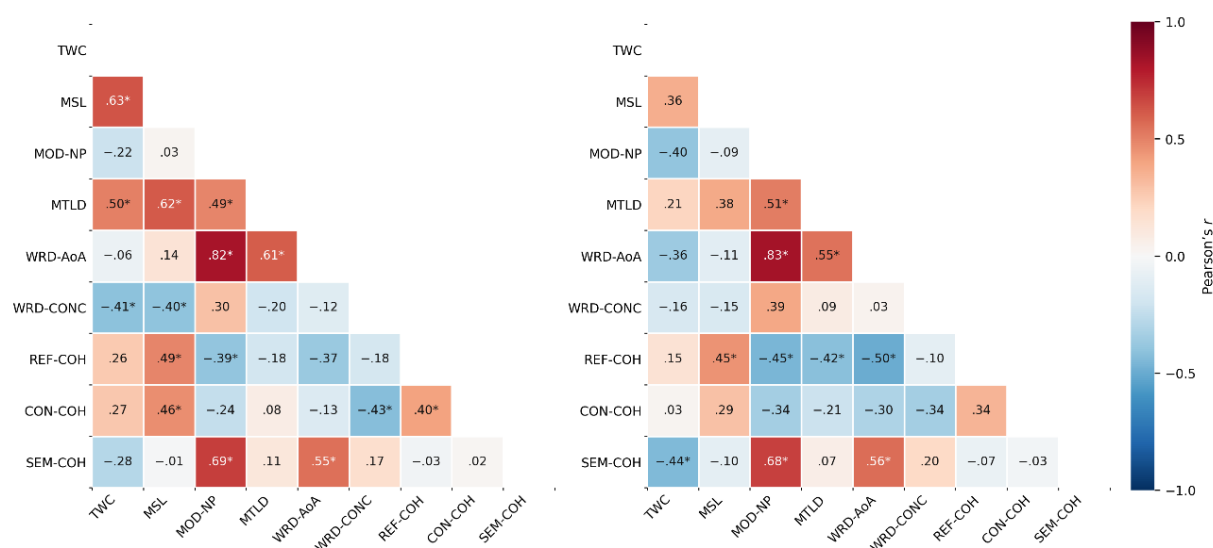


Figure 2. Correlations Among Measured Coh-Metrix Indices (Book Level, N = 60)

Note. The left matrix displays correlations among book means pooled across the three proficiency levels ($df = 58$); the right matrix displays the same correlations after each index was centered within proficiency level, removing between-level differences ($df = 56$). * $p < .05$ after Holm correction for 36 pairwise comparisons within each matrix.

5. Discussion and Conclusion

The present study sought to examine how the textual difficulty of reading passages progresses across internationally published ELT coursebooks designed for lower-level learners. Overall, the findings indicate that publishers increase textual difficulty in a selective rather than uniform manner, with only certain linguistic and discourse features consistently distinguishing adjacent proficiency levels. Text length, sentence length, and lexical diversity emerged as the most reliable indicators of progression, increasing significantly across both the A1-A2 and A2-B1 transitions, whereas other features either changed at only one transition or remained stable throughout. These findings support previous research suggesting that textual difficulty is a multidimensional

construct that develops through targeted adjustments to particular linguistic features rather than simultaneous increases across all dimensions of complexity. However, direct comparison of the two transitions found no significant differences in magnitude for any index, indicating that where difficulty increased, it increased by a comparable amount at each step. Therefore, differences in which indices reached significance at each transition should be read as differences in statistical detectability rather than as evidence that publishers calibrate the two stages differently. In total, five of the nine indices distinguished the reading passages in A1 and A2 coursebooks, indicating these textual passages increase the demands placed on learners across multiple dimensions simultaneously during this early stage of reading development. This pattern aligns closely with the CEFR's characterization of A2 readers, who move beyond understanding very short, highly predictable texts toward comprehending short connected texts containing a wider range of vocabulary and more complex linguistic structures.

Several individual findings further reinforce this interpretation. The substantial increases in text length and sentence length suggest that A2 learners are expected to sustain comprehension over longer stretches of discourse while processing more syntactically demanding sentences. At the same time, the increase in lexical diversity and the use of later-acquired vocabulary indicate that learners encounter a broader and developmentally more advanced vocabulary repertoire than at A1, consistent with the CEFR expectation that they can infer the meaning of some unfamiliar words from context. Notably, the increased incidence of connectives provides readers with more explicit signals of logical relationships between ideas, potentially offsetting some of the increased lexical and syntactic demands introduced at this level. In contrast, neither referential cohesion nor semantic coherence changed significantly, suggesting that although texts become longer and linguistically more demanding, the overall connectedness of discourse does not decrease. This balance between increasing linguistic complexity and preserving explicit discourse support is consistent with theoretical perspectives such as Vygotsky's (1978) Zone of Proximal Development and Krashen's (1985) Input Hypothesis, both of which emphasize providing learners with input that extends beyond their current abilities while remaining sufficiently comprehensible to promote continued development.

At the A2-B1 transition, only four indices reached significance, namely text length, sentence length, lexical diversity, and word concreteness. Notably, the increases in later-acquired vocabulary and connective incidence observed at the earlier transition did not reach significance at B1. Learners at this stage continue to encounter longer texts, longer sentences, and a broader vocabulary, while also being introduced to increasingly abstract lexical content. This pattern aligns well with the CEFR's description of B1 readers, who are expected to comprehend texts describing events, feelings, wishes, and opinions, as well as straightforward factual information, all of which frequently require vocabulary referring to concepts that are less immediately tangible than those typically encountered at the basic user levels.

Notably, word age of acquisition and connective incidence reached significance only at the earlier transition (i.e., A1-A2). However, direct comparison of the two increments found no significant difference for either index as shown in Table 3. Therefore, these results should not be read as evidence that publishers stop increasing lexical difficulty or explicit signaling after A2. The more cautious interpretation is that any continued increase at A2-B1 was too small to detect at the present sample size. Finally, the continued lack of change in referential cohesion and semantic coherence suggests that the overall connectedness of discourse remains relatively stable across lower-level materials, despite increases in other aspects of textual difficulty.

A further consideration concerns the magnitude of the observed effects. Although the adjacent-level contrasts reported above were statistically reliable, none reached the conventional threshold for a large effect. Indeed, each contrast compares consecutive proficiency levels representing a single increment within a graded series, so

moderate effects at each step are consistent with, rather than contrary to, the principle of gradual progression that underlies materials development. Considered across the full lower range, the accumulated differences are substantially larger. The A1-B1 comparison, which was not the focus of the present analysis, indicates large effects for the three indices that progressed at both transitions, namely, text length ($d = 1.23$, 95% CI [1.00, 1.47]), sentence length ($d = 1.17$, CI [0.93, 1.42]), and lexical diversity ($d = 0.96$, CI [0.75, 1.16]). In descriptive terms, mean passage length rose from approximately 150 words at A1 to 265 at B1, mean sentence length from 10.8 to 14.5 words, and lexical diversity from 61.4 to 83.3. Therefore, a learner progressing through these three levels encounters passages roughly three-quarters longer, composed of noticeably longer sentences and drawn from a wider vocabulary range, which represents a pedagogically meaningful increase in processing load. Notably, the three indices that showed no adjacent-level progression, namely modifiers per noun phrase, referential cohesion, and semantic coherence, remained non-significant across the full A1-B1 span as well, indicating that their stability reflects a consistent feature of these materials rather than effects too small to detect between adjacent levels.

As previously mentioned, three indices showed no detectable change at any point in the A1-A2 or A2-B1 transitions (or even when comparing A1 and B1 textual input). Modifiers per noun phrase, referential cohesion, and semantic coherence were non-significant in all contrastive analyses, indicating that their stability is a consistent property of these materials rather than a series of increments too small to detect between neighboring levels. This warrants attention because all three were selected precisely because the CEFR descriptors imply they should change. B1 readers are expected to follow lines of reasoning, interpret logical and temporal connectors, and integrate information across paragraphs, and A2 readers to process more connected discourse than A1 readers, all of which suggests texts that place progressively greater demands on inferencing and on integrating meaning across sentences. The materials examined here do not appear to introduce those demands through reduced cohesion or greater phrasal elaboration. There are several interpretations readily available to explain this finding. Despite CEFR levels implying differences in these textual aspects, the observed pattern, in which discourse support is maintained while lexical and sentence-level demands increase, is consistent with the principle of keeping input comprehensible as it becomes more challenging. Indeed, the stability of these indices may reflect a consistent standard of discourse support maintained across the lower levels rather than a failure to progress. Materials that preserve referential and semantic connectedness while increasing lexical and syntactic demands provide learners with stable discourse support as those demands increase. This may be a desirable property of graded materials rather than an omission. A further consideration applies specifically to modifiers per noun phrase. Phrasal elaboration varies systematically with text type, since descriptive and expository passages typically contain more heavily modified noun phrases than dialogic or narrative ones. The stability of this index may therefore indicate that broadly similar text types recur across the three proficiency levels rather than that phrasal complexity is held constant within a given type. Because text type was not coded in the present study, this possibility cannot be evaluated directly and would require a genre-annotated corpus to test.

The correlation analyses add a further qualification to the picture of selective progression. Text length, sentence length, and lexical diversity were the three indices that increased at every transition, and when books from all levels were analyzed together they appeared to form a coherent group. That grouping disappeared once proficiency level was removed. In short, among books at the same level, a longer passage was not a passage with longer sentences or a wider vocabulary. These indices covary across the corpus because each independently tracks the CEFR sequence, not because they reflect a single underlying dimension of difficulty. Therefore, progression in these materials is selective in a second sense. Not only do some features increase while others remain stable, but the features that do increase vary independently of one another within levels, such that a series

could in principle extend text length without lengthening sentences, or broaden vocabulary without doing either. A second group of indices behaved in a way that raises a caution about interpreting textual measures as evidence of grading at all. Modifiers per noun phrase, word age of acquisition, and semantic coherence were strongly intercorrelated regardless of whether proficiency level was controlled, yet none progressed consistently across levels. Therefore, their association reflects something other than CEFR calibration. One plausible explanation is that differences exist in style across publishers and series, with some materials employing denser, more expository prose at every level.

The present findings have several implications for ELT materials development and classroom practice. Because the analysis examined textual features rather than learner comprehension, it does not indicate which pattern of progression is most effective. However, it does describe how a large sample of internationally published coursebooks currently calibrates reading input across the lower CEFR levels, which may serve as a reference point for publishers and materials writers reviewing their own series. Difficulty in this corpus progressed selectively rather than uniformly, with some features increasing at both transitions, others shifting at only one, and several remaining stable throughout. A series evaluated against a single readability measure would not reveal this pattern, so materials developers may find it useful to consider several dimensions of difficulty separately when examining how reading input develops across a series. The findings also have implications for teachers selecting or supplementing classroom reading materials. In particular, teachers introducing supplementary reading materials may benefit from considering the particular dimensions of textual difficulty emphasized in a text, such as text length, sentence length, lexical diversity, or lexical abstraction, in relation to learners' current reading abilities and instructional goals.

Lastly, several limitations of the present study should be acknowledged. First, the analysis was limited to coursebooks corresponding to the CEFR proficiency levels A1, A2, and B1. Consequently, the observed patterns of progression may not extend to more advanced materials (i.e., B2-C2), where different dimensions of textual difficulty may become more prominent. Second, although the selected Coh-Metrix indices were chosen because they closely aligned with the reading competencies described in the CEFR and represented theoretically motivated dimensions of textual difficulty, they constitute only a subset of the measures available within Coh-Metrix. Additional indices may provide further insight into how reading materials are graded across proficiency levels. Third, the A1, A2, and B1 classifications used in this study were based on the CEFR designations and alignments provided by the publishers themselves, as indicated in book titles, series documentation, or publisher catalogs, rather than on an independent assessment of each text's proficiency level. The findings therefore describe the reading passages contained in coursebooks that international publishers have developed and designated for these CEFR levels, rather than passages that have been independently verified as corresponding to A1, A2, or B1. A further limitation is that text type was not coded. Coursebooks vary in the proportion of dialogic, narrative, and expository passages they contain, and these text types differ systematically in sentence length, phrasal elaboration, and patterns of cohesion. Some of the differences observed across levels may therefore reflect shifts in the distribution of text types rather than adjustments to difficulty within a given type. A genre-annotated analysis would allow these two sources of variation to be separated. Finally, the study relied exclusively on computational analyses of textual characteristics rather than direct measures of learner comprehension. While Coh-Metrix provides well-established indicators of linguistic and discourse complexity, future research should examine how these textual features influence learners' actual reading performance through measures such as comprehension assessments or reading times. Future research addressing these limitations would provide a more comprehensive understanding of how textual difficulty develops across proficiency levels and further refine our understanding of how ELT reading materials support learners' reading development.

References

- Byrnes, H. 1998. Reading in the beginning and intermediate college foreign language class. In G. S. Burkart, ed., *Modules for the Professional Preparation of Teaching Assistants in Foreign Languages*, 227-260. Washington, DC: Center for Applied Linguistics.
- Cheng, Y.-S. and S.-C. Chang. 2022. Do senior high school English textbooks sufficiently prepare students for the high-stakes college entrance examinations? A corpus-based analysis of text difficulty. *Journal of Textbook Research* 15(1), 43-80.
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Strasbourg, France: Council of Europe Publishing.
- Crossley, S., D. Allen and D. McNamara. 2011. Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language* 23(1), 84-101.
- Crossley, S., J. Greenfield and D. McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly* 42(3), 475-493.
- Gelman, A. and H. Stern. 2006. The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician* 60(4), 328-331.
- Grabe, W. 2009. *Reading in a Second Language: Moving from Theory to Practice*. New York: Cambridge University Press.
- Grabe, W. and F. L. Stoller. 2020. *Teaching and Researching Reading*. 3rd ed. New York: Routledge.
- Graesser, A., D. McNamara, Z. Cai, M. Conley, H. Li and J. Pennebaker. 2014. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal* 115(2), 210-229.
- Graesser, A., D. McNamara and J. Kulikowich. 2011. Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher* 40(5), 223-234.
- Graesser, A., D. McNamara, M. Louwerse and Z. Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2), 193-202.
- Hanifah, M., D. Wahyuni and S. Haryati. 2022. The readability of reading texts in English textbook for Indonesian senior high school using Coh-Metrix. *English Education* 10(2), 81-92.
- Krashen, S. 1982. *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon Press.
- Krashen, S. 1985. *The Input Hypothesis: Issues and Implications*. London: Longman.
- Lee, S. 2013. A corpus-based analysis of the difficulty of the reading materials in high school English 1 and 2 textbooks using Coh-Metrix. *Studies in Foreign Language Education* 27(2), 131-148.
- McNamara, D., A. Graesser, P. McCarthy and Z. Cai. 2014. *Automated Evaluation of Text and Discourse with Coh-Metrix*. Cambridge: Cambridge University Press.
- Mishan, F. 2022. The global ELT coursebook: A case of Cinderella’s slipper? *Language Teaching* 55(4), 490-505.
- Nation, I. S. P. 2013. *Learning Vocabulary in Another Language*. 2nd ed. Cambridge: Cambridge University Press.
- North, B. 2014. Putting the Common European Framework of Reference to good use. *Language Teaching* 47(2), 228-249.
- Plakans, L. and Z. Bilki. 2016. Cohesion features in ESL reading: Comparing beginning, intermediate and advanced textbooks. *Reading in a Foreign Language* 28(1), 79-100.

- Richards, J. 2001. *Curriculum Development in Language Teaching*. Cambridge: Cambridge University Press.
- Ryu, J. and M. Jeon. 2020a. A Coh-Matrix analysis of the first grade middle school English textbooks revised by the 2015 National Curriculum. *Korean Journal of English Language and Linguistics* 20, 89-121.
- Ryu, J. and M. Jeon. 2020b. An analysis of text difficulty across grades in Korean middle school English textbooks using Coh-Matrix. *Journal of Asia TEFL* 17(3), 921-936.
- Tomlinson, B. 2012. Materials development for language learning and teaching. *Language Teaching* 45(2), 143-179.
- Tomlinson, B. and H. Masuhara. 2018. *The Complete Guide to the Theory and Practice of Materials Development for Language Learning*. Hoboken, NJ: Wiley Blackwell.
- Vygotsky, L. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.
- Westfall, J., D. Kenny and C. M. Judd. 2014. Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General* 143(5), 2020-2045.

Examples in: English

Applicable Language: English

Applicable Level: Tertiary

Appendix. List of Coursebooks Included in Analysis (N = number of passages)

Publisher	Series Title	A1 (327)	A2 (396)	B1 (431)
Cambridge (208)	<i>Empower</i> (83)	<i>A1 Starter</i> (18)	<i>A2 Elementary</i> (29)	<i>B1 Pre-intermediate</i> (36)
	<i>Face2Face</i> (49)	<i>Starter</i> (11)	<i>Elementary</i> (23)	<i>Pre-intermediate</i> (15)
	<i>Four Corners</i> (33)	1 (12)	2 (11)	3 (10)
	<i>Touchstone</i> (43)	1 (15)	2 (13)	3 (15)
Macmillan (259)	<i>New Inside Out</i> (67)	<i>Beginner</i> (16)	<i>Elementary</i> (28)	<i>Pre-intermediate</i> (23)
	<i>New Inspiration</i> (93)	1 (23)	2 (32)	3 (38)
	<i>Open Mind</i> (36)	<i>Beginner</i> (8)	<i>Elementary</i> (12)	<i>Pre-intermediate</i> (16)
	<i>Straightforward</i> (63)	<i>Beginner</i> (18)	<i>Elementary</i> (20)	<i>Pre-intermediate</i> (25)
National Geographic Learning (213)	<i>Keynote</i> (34)	1 (12)	2 (11)	3 (11)
	<i>Life</i> (75)	<i>Beginner</i> (24)	<i>Elementary</i> (29)	<i>Pre-intermediate</i> (22)
	<i>Outcomes</i> (45)	<i>Beginner</i> (12)	<i>Elementary</i> (17)	<i>Pre-intermediate</i> (16)
	<i>Voices</i> (59)	<i>Beginner</i> (13)	<i>Elementary</i> (23)	<i>Pre-intermediate</i> (23)
Oxford (253)	<i>English File</i> (64)	<i>Beginner</i> (15)	<i>Elementary</i> (23)	<i>Pre-intermediate</i> (26)
	<i>Headway</i> (53)	<i>Beginner</i> (15)	<i>Elementary</i> (11)	<i>Pre-intermediate</i> (27)
	<i>Navigate</i> (70)	<i>A1 Beginner</i> (26)	<i>A2 Elementary</i> (23)	<i>B1 Pre-intermediate</i> (21)
	<i>Wide Angle</i> (66)	1 (18)	2 (24)	3 (24)
Pearson (221)	<i>Cutting Edge</i> (44)	<i>Starter</i> (9)	<i>Elementary</i> (15)	<i>Pre-intermediate</i> (20)
	<i>Roadmap</i> (68)	<i>A1</i> (21)	<i>A2</i> (18)	<i>B1</i> (29)
	<i>Speakout</i> (42)	<i>Starter</i> (11)	<i>Elementary</i> (14)	<i>Pre-intermediate</i> (17)
	<i>Wider World</i> (67)	1 (30)	2 (20)	3 (17)