

Corpus-Based Analysis of Reading Text Selection Criteria in the National Matriculation English Test

Dandan Qin

*Hubei Business College, Wuhan, China
danna.qin@foxmail.com*

Keywords: Corpus; language testing; text selection; mixed methods

Abstract: As China's most influential large-scale high-stakes assessment, the National Matriculation English Test (NMET) has garnered extensive attention. This study investigates the principles of text selection for the reading component of NMET. Based on a corpus analysis of 30 reading texts in the NMET Volume I (2016-2025) and interviews with 10 experienced teachers, the findings reveal a consistent sourcing of texts from authentic English-language media. These texts are further characterized by close alignment with the high school curriculum in both thematic scope and lexical demand, ensuring pedagogical coherence. The study provides practical insights for instructional material selection and test design refinement.

1. Introduction

In China, standardized high-stakes language tests are widely adopted to evaluate students' English proficiency and act as a core criterion for their academic advancement [1]. Known to the public as Gaokao, China's National College Entrance Examination (NCEE) attracts the participation of millions of high school graduates each June. The National Matriculation English Test (NMET), as a mandatory component of the Gaokao, assumes a critical role in verifying whether test-takers' English competence satisfies the academic demands of university curricula [1][2]. Notably, reading comprehension has long been a key section of the NMET, accounting for a substantial proportion of the total score [2]. As a result, the English reading test in the NMET is of vital significance to Chinese examinees.

It is very common in language teaching and assessment to select and adapt appropriate reading texts for students of various language levels [3]. Existing studies mainly assess the difficulty of reading texts [4][5] and explore the vocabulary learning through reading [6], and the cognitive processing of examinees when they complete reading tests [7][8]. However, few studies have focused on the process of text selection for language testing purposes [9]. Existing relevant studies are also focused on international exams like IELTS, TOEFL [10][11] or CET-4 and CET-6 in the context of China [12], or analysis of the text adaptation methods of the NMET reading test [13] and the complexity of texts after adaptation [2]. Nevertheless, few researchers pay attention to the selection of reading texts in the NMET of the Gaokao examination which is one of the large-scale high-stakes language examinations in China.

To address this research gap, this study investigates the reading text selection process of the NMET

Volume I. It is a pivotal examination taken by the largest cohort of test-takers in China over the decade spanning its transition from the traditional to the new curriculum (2016–2025). Adopting a novel mixed-methods approach integrating computerized text analysis and teacher interviews, it aims to provide pedagogical insights for senior high school English reading instruction and practical guidance for teachers and students in selecting Gaokao-aligned reading materials.

2. Text Selection for L2 Reading Tests

Text selection is a core link in designing the reading test in the NMET. From a testing perspective, studies confirm that text features directly shape a test’s reliability and validity [8][14]. Green and Hawkey [9] further note that text selection and adaptation significantly influence test validity, and selecting appropriate source texts is the primary step in test design.

Over 30 years, educators have advocated for authentic texts in L2 learning across proficiency levels. Defined as discourse serving social functions in the target language community [15], authentic texts cover diverse genres (e.g., news, manuals) targeting native audiences [16]. They are more comprehensible to L2 learners and carry higher communicative value [17][18]. Newspapers, for instance, are widely recognized as valuable learning resources [19][20]. In testing, scholars propose reading materials should align with real-world scenarios [21][22]. Bachman and Palmer [21] define “authenticity” as correspondence between test tasks and real-language use tasks, while Larsen-Freeman [23] emphasizes that authentic texts immerse learners in natural linguistic contexts. These studies collectively highlight the necessity of appropriate text selection.

However, text selection for the NMET reading test is particularly complex. Reading tests aim to assess learners’ real-world comprehension [22], and reading involves multi-level processes (e.g., word recognition, working memory) [24]. Among relevant factors, background knowledge, vocabulary proficiency, and individual interest are closely tied to text topic selection [25].

2.1. Background Knowledge

Bartlett’s [26] Schema Theory and Kintsch’s [27] Construction-Integration Model indicate that interactions between readers’ prior knowledge and text information facilitate comprehension. Specifically, content consistent with prior knowledge is easier to grasp [27]. Burgoyne et al. [28] find that readers’ background knowledge levels affect L2 comprehension. Those with richer knowledge perform better in deep-comprehension tasks [29]. Thus, test designers should consider test-takers’ academic backgrounds [22], as topic familiarity significantly impacts high-stakes test performance.

2.2. Vocabulary Proficiency

Vocabulary proficiency is a critical constraint. L2 learners may have advanced reading skills and topic knowledge, but insufficient vocabulary hinders comprehension [29]. L2 learning requires linking linguistic elements to world knowledge [31], and L1 topic knowledge does not guarantee English comprehension. L2 learners also rely more on vocabulary (rather than memory) to extract meaning [32]. Thus, topic-specific vocabulary mastery impacts high-stakes performance, and developers must ensure text vocabulary aligns with the syllabus to maintain validity [22].

2.3. Individual Interest

Individual interest (triggered by text features or habitual preferences) positively affects text memory [33]. There is a strong correlation between background knowledge and topic interest. People know more about topics they care about, and grow interested in familiar topics [34]. Even with limited

background knowledge, interest helps organize text information at basic comprehension levels [34]. Alderson [22] suggests test texts should be engaging, as topic relevance to test-takers' interests strongly influences high-stakes performance.

In summary, existing research clarifies that background knowledge, vocabulary proficiency, and interest are key to L2 reading test design. Therefore, these factors must be central to text selection, particularly when using authentic texts. However, systematic corpus-based analysis of the text selection criteria in the NMET reading test remains scarce. This study addresses this gap by exploring the actual criteria and features of NMET reading texts by corpus methods.

3. Methodology

This study adopts a mixed-methods approach, which integrates quantitative and qualitative strategies to address research gaps comprehensively [35]. For quantitative analysis, 30 English reading texts from the NMET Volume I over the past decade (2016-2025) were selected. Their genres (e.g., newspapers, journals) were identified, and text subjects were classified using the Common Aggregation Hierarchy (CAH), which is a standardized tool for consistent subject grouping to explore selection preferences.

Qualitatively, semi-structured interviews were conducted with 10 teachers experienced in NMET English teaching to elicit their perspectives on text selection [36][37]. Cross-validating quantitative text analysis results with qualitative interview insights ensured the reliability of findings by aligning automated data with practical teaching experience.

4. Results

4.1. Sources of Selected Texts

The analysis of 30 NMET Volume I reading texts (2016–2025) shows dispersed sources, predominantly from English-speaking countries. As Table 1 illustrates, the texts derive from 24 sources, including social media (varying scales), international websites (diverse types), and other sources (e.g., print journals, academic systems). Notably, The New York Times and NPR News are the most frequently used, each appearing at least three times.

Table 1. Frequency Distribution of Text Sources

Text Source Category	Specific Source	Frequency
High-frequency Sources	The New York Times	4
	NPR News	3
	The Economist	2
Low-frequency Sources	Other Text Sources	21

Note: “Other Text Sources” include 21 sources with a frequency of 1 each: UK Express, The Science, Newsweek, CNN News, ABC News, New Scientist, The Guardian, The Atlantic, MIT Technology Review, The Conversation, Bloomberg, WIRED, U.S. Fish & Wildlife Service, Academic System, National Geographic, Queensland traffic route map, US travel website, government website, public welfare website, hotel website, print journal.

These findings align with interviewees' views. They uniformly noted that test texts draw on diverse western media, predominantly from target-language countries, to ensure authenticity. Several highlighted a shift from occasional concentration on single outlets (five years prior) to greater dispersion in recent years.

Regarding the rationale for scattered sources, interviewees cited three key aims: enhancing thematic diversity to rationalize material selection, reducing predictability of text sources for teachers

and students, and encouraging exposure to authentic, diverse English across contexts while discouraging overreliance on specific sources. As one interviewee emphasized: “Text source dispersion is an innovative strategy reminding teachers and students that textbooks alone are insufficient. To excel in the Gaokao English test, regular extensive reading is essential.” (T4)

However, the analysis of top text source categories shows media accounts for 73.3% (Table 2), making it the primary source over the decade. International websites follow at 16.7%, with print journals and other sources comprising 10%. Within media, mainstream outlets (63.6%) dominate, including *The Economist*, *Newsweek*, *The New York Times*, and CNN, followed by professional media (31.8%; e.g., *The Science*, *National Geographic*, *New Scientist*) and mass media (4.6%; e.g., *The Express*, NPR News, ABC News) (Table 3).

Table 2. Proportional Distribution of Text Source Categories

Category	Frequency	Proportion
Media	22	73.3%
Website	5	16.7%
Others	3	10%

Table 3. Proportional Distribution of Media Type

Media Subcategory	Percentage of Media Category (%)	Percentage of Total Sample (%)
Mainstream Media	63.6	46.7
Professional Media	31.8	23.3
Mass Media	4.6	3.3
Total	100.0	73.3

Most interviewees, while not focusing on media categories in daily teaching, noted mainstream media’s authority, formality, and low error rate as key to its preference by test designers. In brief, reading texts in the NMET are dispersed (predominantly English-speaking nations), with international mainstream media as test designers’ top choice.

4.2. Subject Areas of Selected Texts

With the assistance of the CAH version 1.3.4, the subjects of 30 reading texts are classified into 11 groups as shown in the table below (Table 4).

Education and teaching, language and area studies are the top two themes preferred by test designers (each with over four passages). These align with senior high school students’ academic backgrounds and core learning tasks, as they are familiar and relevant to students’ experiences. Psychology, geography (Earth/Env), business and management, and biological and sport science are also popular, given their proximity to daily life and public concerns. In contrast, themes like media, engineering, agriculture, computing, and design are less frequently selected.

Interviewees noted that the Gaokao English papers adhere to high school English curriculum standards. They reinforce basic language knowledge and cultural understanding, assess overall language proficiency grounded in linguistic fundamentals, and guide teachers to prioritize students’ communicative competence. Therefore, theme selection firstly adheres to curriculum standards, an observation echoed by respondents, who noted partial overlap between the Gaokao English reading themes and those in senior high school English textbooks. Second, all emphasized themes are life-relevant and academically appropriate, with language, education, and business topics being more accessible.

Table 4. Discipline Categories in Texts (Frequency & Proportion)

Frequency Group	Discipline(s)	Frequency	% of Total
≥ 3	Education and Teaching (5); Language and Area Studies (4); Biological and Sport Science; Psychology, Geography (Earth/Env), Business & Management (3 each)	5/4/3	18.5/14.8/11.1
2	Agriculture (Food); Media (Journalism); Engineering; Computing (2 each)	2	7.4
1	Design (Creative/Performing Arts)	1	3.7
Total	-	30	100.0

In summary, text analysis shows that the reading texts in the NMET are authentic, dispersed, and predominantly from international media in English-speaking countries. Themes are diverse but concentrated in education and teaching as well as language and area studies, closely tied to students' backgrounds. These findings largely align with interviews, where teachers offered interpretations based on practice and noted textbook-theme overlap.

5. Discussion

5.1. Sources of Selected Texts

The analysis reveals three principal characteristics of reading text selection in the NMET. First, texts are predominantly sourced from English-speaking countries, ensuring authenticity and enabling the assessment of learners' ability to comprehend and apply language in real-world contexts, aligning with established language testing principles [21][22]. Second, text origins are diverse, spanning various media types, commercial and governmental websites, and academic journals. This diversity serves three key purposes: it enhances test validity by covering a range of topics; protects test fairness by reducing predictability; and exposes students to multiple genres within authentic settings, supporting prior research on input variety [17][22]. Finally, a clear preference exists for internationally recognized mainstream global media. This preference is driven by such media's use of accurate, formal, and textbook-aligned language, which minimizes comprehension barriers. This finding corroborates studies advocating newspapers and reputable media as valuable language learning resources [19]. Collectively, the text selection strategy underscores a commitment to authenticity, diversity, and the use of reputable global media from target-language environments, thereby supporting test validity, fairness, and pedagogical alignment with real-world language use.

5.2. Themes of Selected Texts

The reading texts in the NMET exhibit three core thematic characteristics, with findings aligning with curriculum and second language assessment principles.

First, thematic diversity reflects two key design rationales. Test designers aim to assess learners' language proficiency in authentic, varied contexts, supporting Alderson's [22] argument that high-quality reading assessments require multi-thematic texts. That adheres to senior high school curriculum guidelines, confirming a close alignment between test content and syllabus [1].

Second, themes are strongly linked to students' background knowledge, with education and teaching, language and area studies as dominant categories, alongside life-relevant fields (e.g., geography, business, biology). This echoes prior research emphasizing the role of background knowledge in facilitating text comprehension, as test designers prioritize themes familiar to students'

daily learning and lives [22][29]. However, this study does not corroborate the claim that student-interest-aligned themes enhance exam performance [34].

Third, teacher interviews confirm that test themes align with curricula to assess basic language and cultural knowledge authentically and to suit students' vocabulary proficiency. It further validates the literature's emphasis on integrating linguistic elements with prior knowledge for language acquisition [22][34].

In short, themes of the reading test in the NMET are diverse, closely tied to students' background knowledge, and aligned with high school curricula, reflecting the integration of assessment goals with curriculum standards.

6. Conclusion

This mixed-methods study examines the text selection process for the reading test in China's NMET over the past decade. Findings reveal that the NMET reading text selection strategy embodies a focus on authenticity, diversity, and reliable global target-language media sources, which in turn enhances test validity, fairness, and alignment with real-world language use in pedagogy. Meanwhile, the diverse themes of NMET reading tests, closely linked to students' background knowledge and aligned with high school curricula, reflect the integration of assessment objectives with curriculum standards.

The research offers both theoretical and practical contributions. On a theoretical level, it addresses a gap in the literature on high-stakes language testing. Practically, it provides evidence-based insights to aid teachers in material selection and assists test designers in refining the authenticity and relevance of assessment texts.

Future studies should address two key limitations. First, it should be acknowledged that although the research spans ten years, the sample size is relatively small. Second, test designers' perspectives are lacking. To address these gaps, researchers should broaden the data scope and incorporate direct insights from test designers to refine text selection frameworks.

References

- [1] Cheng, L. (2008). *The key to success: English language testing in China*. *Language Testing*, 25(1), 15-37.
- [2] Yu, X. (2021). *Text complexity of reading comprehension passages in the national matriculation English test in China: The development from 1996 to 2020*. *International Journal of Language Testing*, 11(2), 142-167.
- [3] Bunch, G. C., Walqui, A., and Pearson, P. D. (2014). *Complex text and new common standards in the United States: Pedagogical implications for English learners*. *TESOL Quarterly*, 48(3), 533-559.
- [4] Hu, Y. (2006). *Two thoughts on secondary school English reading instruction*. *Policy and Scientific Consult*, 6(2), 39-40.
- [5] Luo, C. K. (2007). *Vocabulary instruction in secondary schools*. *Heilongjiang Science and Technology Information*, 13(1), 203-207.
- [6] Miao, X. H. (2010). *Problems and strategies on vocabulary instruction in secondary schools*. *Journal of Lanzhou Institute of Education*, 26(2), 150-153.
- [7] Bax, S. (2013). *The cognitive processing of candidates during reading tests: Evidence from eye-tracking*. *Language Testing*, 30(4), 441-465.
- [8] Weir, C., Hawkey, R., Green, A., and Devi, S. (2012). *The cognitive processes underlying the academic reading construct as measured by IELTS*. *IELTS collected papers*, 33(2), 212-269.
- [9] Green, A., and Hawkey, R. (2012). *Re-fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading*. *Language Testing*, 29(1), 109-129.
- [10] Salisbury, K. (2005). *The edge of expertise: towards an understanding of listening test item writing as professional practice* (Doctoral dissertation, University of London).
- [11] Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., and Roullion, V. (2010). *A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual item writers' characteristics*. *Language Assessment Quarterly*, 7(2), 160-174.
- [12] Gu, X. D. (2020). *A comparative study on the source and adaptation of CET-6 and IELTS reading texts*. *Foreign*

languages and translation, 67(3), 52-64.

[13] Zhang, H. Q. and Gu, X. D. (2019). An Analysis of the Text Adaptation Methods of the National Matriculation English Test Reading Materials: A Case Study of NMET 2018. *China Examinations*, 331(11), 36-42.

[14] Green, A. (2013). Adapting or developing source material for listening and reading tests. *The companion to language assessment*, 10(2), 830-846.

[15] Lee, W. Y. C. (1995). Authenticity revisited: Text authenticity and learner authenticity. *ELT Journal*, 49(4), 323-328.

[16] Crossley, S. A., Louwerse, M. M., McCarthy, P. M., and McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1), 15-30.

[17] Krashen, S. (1981). *Second language acquisition and second language learning*. Oxford Pergamon Press.

[18] Gilmore, A. (2007). Authentic materials and authenticity in a foreign language. *learning. Language teaching*, 40(2), 97-118.

[19] Sanderson, P. (1999). *Using Newspapers in the Classroom*. Cambridge University Press.

[20] Kathiresan, K. (2022). Use of Newspapers in Developing Language Skills. *Journal of Positive School Psychology*, 6 (4), 7215-7219.

[21] Bachman, L.F. and Palmer, A.S. (1996). *Language testing in practice*. Oxford University Press.

[22] Alderson, J. C. (2000). *Assessing Reading*. Cambridge University Press.

[23] Larsen-Freeman, D. (2002). *Techniques and principles in language teaching*. Oxford University Press.

[24] August, D., Francis, D.J., Hsu, H.-Y.A. and Snow, C.E. (2006). Assessing reading comprehension in bilinguals. *The Elementary School Journal*, 107(2), 221-238. doi:10.1086/510656.

[25] Alexander, P. A., and Jetton, T. L. (2000). Learning from text: A multidimensional and developmental perspective. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research*, Vol. 3 (pp. 285-310). Lawrence Erlbaum Associates Publishers.

[26] Bartlett, F.C. (1932). *Remembering*. Cambridge University Press.

[27] Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163-182. doi:10.1037/0033-295X.95.2.163.

[28] Abu-Rabia, S. (2003). Cognitive and social factors affecting Arab students learning English as a third language in Israel. *Educational Psychology*, 23(4), 347-360.

[29] Burgoyne, K., Whiteley, H. E., and Hutchinson, J. M. (2013). The role of background knowledge in text comprehension for children learning English as an additional language. *Journal of Research in Reading*, 36(2), 132-148.

[30] McNamara, D. S., and Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse processes*, 22(3), 247-288.

[31] Alptekin, C., Erçetin, G., and Bayyurt, Y. (2007). The effectiveness of a theme-based syllabus for young L2 learners. *Journal of Multilingual and Multicultural Development*, 28(1), 1-17.

[32] Nassaji, H. (2002). Schema theory and knowledge-based processes in second language reading comprehension: A need for alternative perspectives. *Language learning*, 52(2), 439-481.

[33] Catrysse, L., Gijbels, D., and Donche, V. (2018). It is not only about the depth of processing: What if eye am not interested in the text? *Learning and Instruction*, 58(1), 284-294.

[34] Boscolo, P., and Mason, L. (2003). Topic knowledge, text coherence, and interest: How. they interact in learning from instructional texts. *The Journal of Experimental Education*, 71(2), 126-148.

[35] Sadan, V. (2014). Mixed methods research: A new approach. *International Journal of Nursing Education*, 6(1), 254.

[36] Hochschild, J. L. (2009). Conducting intensive interviews and elite interviews. In *Workshop on interdisciplinary standards for systematic qualitative research* (pp. 124-127). Washington, DC: National Science Foundation.

[37] Cohen, L., Manion, L., and Morrison, K. (2018). *Research methods in education* (Eighth;8; ed.). Routledge.