ISSN:2790-167X Volume-9-(2024)

From Coverage to Distribution: Exploring Lexical Features of the National Matriculation English Test

Tao Yang[1, a], Zhenhui Liang [2, b]

[1]Graduate School, Xi'an International Studies University, Xi'an, China;

[2] School of Translation Studies, Xi'an International Studies University, Xi'an, China.

[a] yangtao@xisu.edu.cn, [b] liangzhenhui@xisu.edu.cn

Abstract. Lexical feature has long been a pivotal element of almost all high-stakes language tests. Since the implementation of the "Experimental Curriculum Criteria" in China, few studies have reported investigating lexical features of the ensuing National Matriculation English Test with corpus methodology, and notably none was conducted on word dispersion. To address this problem, Python programming was employed in the present study to perform a corpus-based two-way coverage and visualized distribution analysis between the National Matriculation English Test and Experimental Curriculum Criteria lexicon. It was found that: 1) text coverage of the National Matriculation English Test reached the minimal (95%) threshold yet not the optimal (98%) one for adequate comprehension; 2) word-list coverage of the Experimental Curriculum Criteria was disproportionate and insufficient, suggesting that a large volume (42.905%) of the prescribed lexicon has never been used during the 13 years of implementation; 3) a relatively few (N = 90) high-frequency words, most (74.444%) of which were significantly overused compared with their corresponding BNC frequency, constituted over half (51.403%) of the text coverage; and 4) a vast majority (93.333%) of high-frequency words was homogeneous in dispersion, confirming the overuse with fresh distribution evidence. The results are discussed in terms of implications for test development.

Keywords: text coverage; word frequency; word dispersion; language corpora.

1. Introduction

Designing and developing a language test is never easy and requires a great deal of expertise [1]. Nevertheless, test developers tend to depend largely on personal experiences, knowledge, and intuition when writing test items [2, 3, 4]. As an authoritative large-scale selective high-stakes test in China [5], the National Matriculation English Test (NMET) makes inferences about candidates' English language ability [6] and serves as a major criterion for the admission to tertiary education [7].

Since the restoration of Gaokao (the national college entrance examinations), NMET has evolved from a uniform national version to a coexistence of several national and regional versions [8]. The present study concentrates on the most widely used national version entitled "the 2nd version of NMET based on the new curriculum criteria" (NMET-v2), which initiated in 2007 [9] in consequence of the reform of China's "Experimental Curriculum Criteria (ECC)" in 2004 [10]. Moreover, ECC lexicon was prescribed at each stage of primary and secondary education by Ministry of Education [10, 11] and the National Education Examinations Authority [9] for teaching reference, which imposed the authoritative requirements for candidates' NMET lexicon.

Owing to the fact that candidates' performance in Gaokao is the sole criterion for college admission [7], NMET exerts a strong washback effect on primary and secondary English teaching and learning in China [6]. Therefore, it requires a comprehensive assessment of the lexicon used in NMET for the sake of test development. Nevertheless, a large body of literature has simply discussed classroom teaching of NMET lexicon or candidates' preparation for NMET lexicon with little or no empirical or statistical evidence [12, 13, 14]. Inspired by the development of corpus linguistics, recent years have witnessed a surge in the adoption of quantitative approaches in lexical

studies within the domain of language testing [15, 16, 17]. Accordingly, research on lexical features of NMET has arisen with the assistance of corpus-based tools or computer programs [18, 19].

Considering the high-stakes nature of NMET [7] and the common practice of examination-oriented teaching in China's primary and secondary education [20], lexical coverage has come into the research focus of NMET when fairness was taken into consideration [21]. According to Nation [22], text coverage refers to "the percentage of running words in the text known by the readers". Various studies based on empirical evidence have established that 95% coverage is the minimal threshold for adequate comprehension, while 98% coverage is suggested as the optimal threshold [22, 23, 24]. The other side of the coin, word-list coverage, demonstrates particular importance in China under the context of examination-oriented teaching and authoritativeness of language testing [20]. Word-list coverage was defined as "the percentage of items in the word lists covered by the lexicon in NMET" in the present study. To date, researchers have investigated text coverage and word-list coverage of some particular parts [19] or a specific period [25] of NMET, but none of them has addressed the issue from a panoramic and rigorous NMET-ECC relationship.

Another underexplored domain of NMET lexical research is word distribution, which, based on the premise that "a word that is both frequent and widely distributed across the entire corpus is more 'important' than a high-frequency word that is restricted to just one or two texts" [26], is proposed in the present study to incorporate both word frequency [27, 28] and word dispersion [26]. Nevertheless, all of the current literature on NMET distribution pays exclusive attention to word frequency rather than word dispersion. To this end, more research into dispersion evidence is required to fill the research gap in NMET lexical study. Considering all that has been mentioned so far, much uncertainty still exists about the relationship between NMET and ECC lexicon in terms of coverage and distribution. Therefore, the following two research questions were addressed in this study:

1. What are the text coverage of NMET-v2 and the word-list coverage of ECC?
2. To what degree do word frequency and dispersion in NMET-v2 affect coverage?

2. Method

2.1 Materials

Since its inception in 2007, NMET-v2 has been a major and the most widely used national version developed by the National Education Examinations Authority [29] for China's Gaokao system. Due to a new revision of the curriculum criteria in 2018 [5], the NMET syllabus for ECC ceased to update after 2019 [30]. Accordingly, NMET-v2 papers from 2007 to 2019, all of which were typeset and double-checked on the basis of official publications of the National Education Examinations Authority, were selected as the corpus for the present study. Listening comprehension was excluded due to the fact that it is a discretionary part for candidates in different provinces of China [31].

On the other hand, lexicon prescribed by ECC was stratified into four categories according to the requirements imposed for different stages of education. Word List I consisted of lexicon for primary school [11]. Word List II was comprised of lexicon for junior high school [11]. Word List III included lexicon for graduation requirements imposed for senior high school [10]. Word List IV incorporated lexicon for test-taking purpose [9, 10]. The number of lexicon in Word List I-IV was 430, 1085, 1148, and 931 respectively.

2.2 Instruments

The open-source programming language Python (v3.7.3) was employed for text mining and visualization. Python code was written by the researchers to conduct noise removal (i.e. eliminating irrelevant symbols, punctuations, Chinese characters, numeric digits, and whitespaces). Several

integrated or third-party open-source libraries were incorporated in the code for text processing. To be specific, tokenization was performed by jieba (v0.40) library. POS tagging and lemmatization were facilitated by UnigramTagger and WordNetLemmatizer in NLTK (v3.4.5) library, based on the machine learning of the annotated Brown Corpus. Mathematical and statistical calculations were implemented by NumPy (v1.16.5) library and SciPy (v1.3.1) library. All these steps were written by the researchers into a single Python algorithm for effectiveness and accuracy.

2.3 Data analysis

To investigate the first research question, NMET-v2 corpus were processed in the sequence of noise removal, tokenization, lowercasing, POS tagging, and lemmatization to get the lemmas that were comparable to the items in Word List I-IV. Two-way comparison was subsequently made between NMET-v2 corpus and Word List I-IV by the Python algorithm in order to compute text coverage and word-list coverage.

To investigate the second research question, occurrences of the lemmas were calculated first and then listed in reverse frequency order. British National Corpus (BNC) was employed as the reference corpus to compare inter-corpus word frequencies, since BNC, considering its balanced and representative nature, was in concordance with NMET's objective of presenting "real-life" and "authentic" language use [5]. Previous research has shown that log-likelihood ratio (G2) test performs better than Pearson's Chi-square test in analyzing non-normal distribution samples, reducing overestimation of the importance of rare events, and showing insensitivity to differences of size between two samples [32], G2 was therefore computed by NumPy and SciPy libraries in the Python algorithm to check the significance ($p < 0.05$) of the overuse of high-frequency words. With regard to dispersion, Juilland's D was selected on the grounds that "it has been shown to be the most reliable of the various dispersion coefficients that are available" [32]. The claimed theoretical range for Juilland's D is 0.000 to 1.000, with values close to 1.000 indicating a completely homogeneous dispersion, and values close to 0.000 reflecting a maximally skewed dispersion [26]. Juilland's D was computed by NumPy and SciPy libraries in the Python algorithm.

3. Results

3.1 Text coverage and word-list coverage

Text coverage calculated the ratio of tokens in NMET-v2 covered by items in the word lists to the total number of tokens in NMET-v2.

Table 1. Text Coverage of NMET-v2

Year	Token	Text coverage by word list				Total coverage	PNAW coverage	Adjusted coverage
		I	II	III	IV			
2007	2459	62.464	28.670	3.253	0.854	95.241	2.725	97.966
2008	2494	60.946	26.945	6.014	1.123	95.028	2.646	97.674
2009	2572	65.941	23.095	4.743	0.933	94.712	3.149	97.861
2010	2485	64.064	24.547	4.467	1.247	94.325	2.938	97.263
2011	2518	60.127	28.912	4.369	1.271	94.679	2.025	96.704
2012	2476	60.905	26.939	6.058	2.100	96.002	1.212	97.214
2013	2420	62.066	26.364	4.256	1.322	94.008	4.050	98.058
2014	2348	62.095	25.809	4.514	1.448	93.866	2.896	96.762
2015	2242	57.761	27.431	5.932	2.676	93.800	3.568	97.368
2016	2310	59.481	23.896	5.758	2.468	91.603	3.896	95.499
2017	2409	58.863	25.820	5.853	2.698	93.234	3.694	96.928
2018	2355	58.132	24.161	7.261	3.057	92.611	3.270	95.881
2019	2457	58.771	23.240	5.454	3.826	91.291	4.925	96.216

Results indicated that text coverage ranged between a low of 91.291% in 2019 and a high of 96.002% in 2012 (M = 93.877%). Considering that words "Not in the lists" pose different comprehension difficulties to test-takers, further categorization was made to the "Not in the lists" words. The first category was proper nouns signifying names of people, places, and organizations etc.; the second category was words with annotations in Chinese characters in the original NMET papers. In terms of comprehension difficulty, both categories are "easily understood" [22] by the reader and therefore was labeled collectively as "proper nouns and annotated words (PNAW)". By contrast, the third category was completely new words to test-takers and was thus labeled as "out-of-curriculum words". Taking PNAW into account, the adjusted text coverage (i.e. original word-list coverage plus PNAW coverage) ranged between a low of 95.499% in 2016 and a high of 98.058% in 2013 (M = 97.030%). Text coverage of NMET-v2 and more details are presented in Table 1.

Table 2. Word-list coverage of ECC

Year	Type	Word List I		Word List II		Word List III		Word List IV	
		No.	Perc.	No.	Perc.	No.	Perc.	No.	Perc.
2007	674	226	52.558	307	28.295	55	4.791	17	1.826
2008	692	219	50.930	307	28.295	78	6.794	20	2.148
2009	683	229	53.256	299	27.558	72	6.272	18	1.933
2010	745	245	56.977	308	28.387	78	6.794	25	2.685
2011	736	218	50.698	329	30.323	77	6.707	23	2.470
2012	681	202	46.977	302	27.834	83	7.230	27	2.900
2013	673	196	45.581	288	26.544	74	6.446	22	2.363
2014	680	208	48.372	295	27.189	70	6.098	25	2.685
2015	676	195	45.349	275	25.346	80	6.969	43	4.619
2016	750	210	48.837	277	25.530	109	9.495	40	4.296
2017	756	194	45.116	312	28.756	93	8.101	53	5.693
2018	800	209	48.605	297	27.373	117	10.192	57	6.122
2019	776	187	43.488	299	27.558	94	8.188	61	6.552
Average			48.980		27.614		7.237		3.561
Total		383	89.070	889	81.935	518	45.122	262	28.142

Word-list coverage calculated the ratio of items in the word lists covered by the types used in NMET-v2 to the total number of items in the word lists. Results indicated that word-list coverage varied considerably, with the maximum in Word List I (56.977% in 2010) and the minimum in Word List IV (1.826% in 2007). Word-list coverage of ECC and more details are shown in Table 2. Results also revealed that the overall full-range word-list coverage was 57.095%, suggesting that more than 40% of ECC lexicon has never been used in NMET-v2 over the 13 years of implementation.

3.2 Word frequency and word dispersion

Table 3. Word frequency of NMET-v2

Freq.	No. of Types	Cumulative occurrences	Perc. of text coverage
1~10	2541	7198	22.818
11~20	240	3521	11.162
21~30	81	1996	6.327
31~40	44	1531	4.853
41~50	24	1084	3.436
51~60	14	776	2.460
61~70	10	637	2.019
71~80	9	690	2.187
81~90	3	251	0.796

91~100	12	1133	3.592
>100	42	12728	40.349
Total	3020	31545	100

Word frequency computed the occurrences of different forms of the word in NMET-v2. As is shown in Table 3, the total number of tokens and types in NMET-v2 was 31,545 and 3,020 respectively. Word frequency of the 3,020 types varied from a minimum of 1 to a maximum of 1629 (M = 10.445). Below the mean value, 2,541 types had a cumulative frequency of 7,198, accounting for 22.818% of occurrences in NMET-v2. By contrast, 479 types with frequencies above the mean value had a cumulative frequency of 24,347, accounting for 77.182% of occurrences in NMET-v2. It is noteworthy that 90 types with frequencies above 50 had a cumulative frequency of 16,215, accounting for 51.403% of occurrences. Ensuing G2 test between frequencies of the 90 types and their corresponding frequencies in BNC corpus [32] was subsequently computed to check whether they were overused (see Table 4). Results indicated that 67 (74.444%) of the 90 types had significantly higher frequencies.

Table 4. Log-likelihood ratio and Juilland's D of high-frequency words

Type	Freq.	LL.	P	J's D	Type	Freq.	LL.	P	J's D
the	1629	54.696	0.000	0.956	work	96	51.182	0.000	0.809
be	1046	65.146	0.000	0.960	go	95	10.635	0.001	0.894
to	980	30.621	0.000	0.974	an	94	1.894	0.169	0.877
a	785	14.296	0.000	0.942	this	94	20.537	0.000	0.884
and	697	27.111	0.000	0.959	will	93	2.075	0.150	0.861
of	648	91.456	0.000	0.968	so	93	8.059	0.005	0.885
in	542	4.323	0.038	0.946	child	92	114.346	0.000	0.730
it	388	5.444	0.020	0.906	or	92	5.579	0.018	0.896
have	370	8.738	0.003	0.910	say	92	1.750	0.186	0.871
you	351	63.220	0.000	0.914	them	85	13.858	0.000	0.864
I	351	16.119	0.000	0.889	there	83	4.198	0.040	0.822
for	347	19.600	0.000	0.921	who	83	4.521	0.033	0.833
that	273	18.594	0.000	0.952	day	80	55.532	0.000	0.801
do	257	31.053	0.000	0.922	if	80	0.352	0.553	0.887
on	237	0.332	0.564	0.919	know	78	5.139	0.023	0.831
not	228	2.330	0.127	0.899	all	77	0.422	0.516	0.864
he	208	0.212	0.645	0.844	she	77	17.131	0.000	0.715
with	203	0.092	0.762	0.905	other	77	8.575	0.003	0.842
at	175	3.480	0.062	0.897	would	75	3.119	0.077	0.892
they	173	8.614	0.003	0.875	look	74	21.816	0.000	0.814
as	165	0.078	0.780	0.846	like	72	7.387	0.007	0.832
from	163	7.293	0.007	0.888	help	67	81.669	0.000	0.819
what	154	54.088	0.000	0.920	learn	65	174.092	0.000	0.820
your	154	158.214	0.000	0.860	see	65	0.307	0.580	0.878
we	149	10.146	0.001	0.910	find	65	26.647	0.000	0.819
when	145	63.825	0.000	0.900	some	64	1.689	0.194	0.889
make	141	56.677	0.000	0.906	than	63	21.378	0.000	0.801
can	140	28.934	0.000	0.925	after	63	15.062	0.000	0.920
time	132	65.691	0.000	0.893	come	63	4.306	0.038	0.862
but	128	2.032	0.154	0.920	then	61	1.903	0.168	0.867
about	126	48.362	0.000	0.902	new	61	13.280	0.000	0.822
my	126	80.770	0.000	0.822	student	60	140.352	0.000	0.748
his	124	1.196	0.274	0.839	how	57	15.147	0.000	0.836
their	123	16.863	0.000	0.779	want	57	18.764	0.000	0.834
up	119	44.664	0.000	0.892	school	56	54.207	0.000	0.823
get	115	23.656	0.000	0.861	me	56	3.449	0.063	0.876
more	109	27.527	0.000	0.851	first	56	7.510	0.006	0.823

one	107	1.249	0.264	0.860	him	56	0.288	0.591	0.725
by	107	20.510	0.000	0.879	use	56	2.842	0.092	0.878
good	106	68.150	0.000	0.865	show	56	34.548	0.000	0.817
people	104	68.823	0.000	0.929	which	56	38.820	0.000	0.855
her	103	85.268	0.000	0.721	our	54	14.948	0.000	0.817
take	98	23.756	0.000	0.871	no	52	3.512	0.061	0.812
year	97	30.268	0.000	0.852	may	52	6.197	0.013	0.830
out	97	35.731	0.000	0.857	give	52	2.892	0.089	0.848

Word dispersion measured the even distribution of the 90 high-frequency types. As is presented in Table 4, Juilland's D of the types ranged between a low of 0.715 and a high of 0.974 (M = 0.866). Specifically, there were 23 types with Juilland's D above 0.900 (M = 0.930); 61 types with Juilland's D between 0.800 and 0.900 (M = 0.855); 6 types with Juilland's D between 0.700 and 0.800 (M = 0.736); suggesting that the vast majority (93.333%) of the high-frequency types had homogeneous distributions in NMET-v2.

4. Discussion and Conclusion

The first research question sought to investigate the text coverage of NMET-v2 and the word-list coverage of ECC. The findings revealed that only a full knowledge of ECC lexicon (Word List I-IV) could hardly reach the minimal threshold (95%) suggested by previous studies [22, 23, 24] for adequate comprehension in most of the years (N = 10), while PNAW, which pose "a minimal learning burden" [22] for comprehension, plus ECC lexicon would virtually ensure adequate comprehension with the minimal threshold yet not the optimal one (98%) suggested by previous studies [22, 23, 24] in most of the years (N = 12). A note of caution is due here since researchers [33] argue that further evidence is needed to assess the difficulty in understanding the proper nouns to secure reliable results. The findings are consistent with that of Huang et al. [19] but lower than that of Li [25]. One possible explanation for this might be that whether NEMT corpus selected in Li's study [25] was rigidly confined to test papers based on ECC was not clarified. On the other hand, the findings revealed that word-list coverage of ECC lexicon was disproportionate, and the overall full-range word-list coverage was insufficient on the grounds that 42.905% of ECC lexicon has never been used during the 13 years of implementation.

To further verify the insufficient coverage of ECC lexicon, the second research question investigated word frequency of NMET-v2 and the dispersion of high-frequency words. Findings of word frequency indicated that a relatively few (N = 90) high-frequency words, most of which (74.444%) were significantly overused according to G2 test, constituted over half (51.403%) of the text coverage. It should be noted that all of the 90 high-frequency words are Word List I or Word List II lexicon. Comparison of the findings with those of other studies confirms a similar high proportion of functional words among the high-frequency words [34]. One previous study [25] investigated the overuse of NMET high-frequency content words using log-likelihood ratio test, yet the research findings are not comparable since most of the words in the study did not exist in the list of the present research. It is understandable considering that high-frequency content words have much lower frequency, compared with the functional words [34]. However, simple word-frequency measurements can be misleading in judging the significance of difference, thus dispersion statistics are needed to avoid distorting effects of overuse tested by frequency alone [32]. To be specific, high-frequency words with high dispersion values indicate high currency in the language, while high-frequency words with low dispersion values should be interpreted with caution [32]. Further evidence of dispersion in the present study confirmed the overuse of high-frequent words on the grounds that a vast majority (93.333%) of high-frequency words was homogeneous in dispersion.

Owing to the "authoritativeness of NMET" and the "inaccessibility of the test data" [35], the National Education Examinations Authority has undertaken the major responsibility for the development, implementation, interpretation and use of NMET; therefore, few studies have sought

to question the validity of word use in the test [20]. Findings of the present study, while preliminary, suggested an insufficient coverage of the ECC lexicon and the overuse of high-frequency words in NMET. When writing test items for high-stakes language tests in China, it is urgent that the official agencies for test development in China take the washback effect on classroom teaching into account, considering the common practice of examination-oriented classroom teaching in China's primary and secondary education epitomized by the popular tendency of "teaching only what they test" [20]. The National Education Examinations Authority and test developers were therefore suggested to gradually reduce the excessive reliance on lexicon of compulsory education (Word List I and Word List II) in NMET so that more uncovered ECC lexicon would be used, which is beneficial for a more comprehensive and legitimate evaluation of candidates' mastery of ECC lexicon. One possible solution may be that the scope of topics in NMET be reasonably broadened to reduce the recurrence of high-frequency words.

This study set out to investigate lexical coverage and distribution in NMET with a Python-based corpus approach. The insufficient coverage of the ECC lexicon is supported by current findings, and the overuse of high-frequency words in NMET is identified as a major cause of the insufficiency, verified by novel distribution data to avoid over-interpretation of the results. The present study provides the first dispersion statistics of NMET lexicon and is one of the first attempts to thoroughly examine the relationship between NMET and ECC from a lexical perspective. The insights gained from this study may be of assistance to lexical features research in language testing. Despite being informative, an issue that was not addressed in this brief research report was whether other dimensions of lexical features in NMET contributed or jointly contributed to the insufficient coverage of the ECC lexicon. Therefore, further research on these components would be of great help in determining the causal relationship between the overuse of high-frequency words and the insufficient word-list coverage.

Acknowledgments

This study was supported by Project of Humanities and Social Sciences of Ministry of Education, P.R. China (Grant No.: 17YJC740107).

References

[1] Schmitt N, Nation P, Kremmel B. Moving the field of vocabulary assessment forward: the need for more rigorous test development and validation. Language Teaching, 2020, 53(1): 109-120.

[2] Fulcher G, Davidson F. Language Testing and Assessment: An Advanced Resource Book. New York: Routledge, 2007.

[3] Bachman L F, Palmer A S. Language Testing in Practice: Designing and Developing Useful Language Tests. Oxford: Oxford University Press, 1996.

[4] Bachman L F. What does language testing have to offer?. TESOL Quarterly, 1991, 25(4): 671-704.

[5] Ministry of Education. English Curriculum Criteria for General High School (2017 version), Beijing: People's Education Press, 2018.

[6] Cheng Liying, Qi Luxia. Description and examination of the National Matriculation English Test. Language Assessment Quarterly, 2006, 3(1): 53-70.

[7] Wang Jianlan, Li Qiqi, Luo Ying. Physics identity of Chinese students before and after Gaokao: the effect of high-stake testing. Research in Science Education, 2022, 52: 675–689.

[8] Pan Mingwei, Qian David D. Embedding corpora into the content validation of the grammar test of the National Matriculation English Test (NMET) in China. Language Assessment Quarterly, 2017, 14(2): 120-139.

[9] National Education Examinations Authority. National Matriculation Test Syllabus (2007 Experimental Curriculum Criteria Version). Beijing: Higher Education Press, 2006.

[10] Ministry of Education. Experimental English Curriculum Criteria for General High School. Beijing: People's Education Press, 2003.

[11] Ministry of Education. English Curriculum Criteria for Compulsory Education. Beijing: Beijing Normal University Publishing Group, 2012.

[12] Qin Yinghua. An analysis of vocabulary of 2021 new NMET version I and its implications for NMET preparation. Education of Guangdong Province, 2022, 13(1): 36.

[13] Huang Fang. An analysis of NMET cloze and its implications for vocabulary teaching and learning. Test and Research, 2020, 33(18): 48-49.

[14] Chen Xiyan. The implications of 2017 NMET (Jiangsu version) for high school vocabulary teaching and learning. Window of Knowledge, 2017, 9(10): 31.

[15] Szudarski P. Corpus Linguistics for Vocabulary: A Guide for Research. New York: Routledge, 2018.

[16] Egbert J. Corpus linguistics and language testing: navigating uncharted waters. Language Testing, 2017, 34(4): 555-564.

[17] Xi Xiaoming. What does corpus linguistics have to offer to language assessment?. Language Testing, 2017, 34(4): 565-577.

[18] Yu Xiaoli. 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, 2021, 11(2): 142-167.

[19] Huang Liyan, Wang Jiaying, Hua Xiaochen. "A corpus-based analysis of text complexity in NMET reading and its implication on reading instruction," in Technology in Education: Innovations for Online Teaching and Learning, eds. Lee et al. (Singapore: Springer), 2020: 23-34.

[20] Yang Huizhong. Valid testing, effective teaching, and valid test use. Journal of Foreign Languages, 2015, 38(1): 2-26.

[21] Kunnan A. "Test fairness," in Europe Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference, eds. M. Milanovic and C. J. Weir (Cambridge: Cambridge University Press), 2004: 27–48.

[22] Nation I S P. How large a vocabulary is needed for reading and listening?. Canadian Modern Language Review, 2006, 63(1): 59-82.

[23] Nurmukhamedov U, Webb S. Lexical coverage and profiling. Language Teaching, 2019, 52(2): 188–200.

[24] Laufer B, Ravenhorst-Kalovski G C. Lexical threshold revisited: lexical text coverage, learners' vocabulary size and reading comprehension. Reading in a Foreign Language, 2010, 22(S1): 15-30.

[25] Li Qingshen. A study on characteristics of vocabulary in English tests of National College Entrance Examination (2007–2016) in China. Journal of Teaching and Management, 2018, 35(18): 118-121.

[26] Biber D, Reppen R, Schnur E, Ghanem R. On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. International Journal of Corpus Linguistics, 2016, 21(4): 439-464.

[27] Rayson P, Berridge D, Francis B. "Extending the Cochran rule for the comparison of word frequencies between corpora," in Le Poids des mots: Actes des 7es Journées internationales d'analyse statistique des données (JADT vol. 2), eds. G. Purnelle, C. Fairon and A. Dister (Louvain: Presses Universitaires de Louvain), 2004: 926-936.

[28] Laufer B, Nation P. Vocabulary size and use: lexical richness in L2 written production. Applied linguistics, 1995, 16(3): 307-322.

[29] National Education Examinations Authority. National Matriculation Test Analysis for Candidates of Science (2008 Experimental Curriculum Criteria Version). Beijing: Higher Education Press, 2008.

[30] The State Council. Guiding principles on promoting education reform in general high schools in the new era. Available online at: http://www.gov.cn/zhengce/content/2019-06/19/content_5401568.htm.

[31] National Education Examinations Authority. National Matriculation Test Analysis (2020 NMET Volume). Beijing: Higher Education Press, 2020.

[32] Leech G, Rayson P, Wilson A. Word Frequencies in Written and Spoken English: Based on the British National Corpus. London: Routledge, 2001.

[33] Brown D. An improper assumption? The treatment of proper nouns in text coverage counts. Reading in a Foreign Language, 2010, 22(2): 355-361.

[34] Wang Rong. Study on the vocabulary of reading comprehension texts in Jiangsu National Matriculation English Test from 2008 to 2017. Educational Measurement and Evaluation, 2018, 11(01): 19-25+53.

[35] Liu Jianda, He Manzu. New development of validity theories in language testing. Modern Foreign Languages, 2020, 43(4): 565-575.