

A Further Look into the Use of a Dictionary APP in EFL Writing: A Replication Study

Yuzhen Chen, *College of Foreign Languages,
Putian University, Fujian, P.R.C.*
(287323222@qq.com)

and

Suping Liu, *College of Foreign Languages,
Putian University, Fujian, P.R.C.*
(Corresponding Author, 260033359@qq.com)

Abstract: The study replicated the experiment by Chen and Liu (2022), investigating the effect of dictionary use on EFL writing. It involved the same research variables as the original study except for adopting a different dictionary. Sixty-two English majors took two writing tests, one without dictionary assistance, the other with access to a mobile phone dictionary application which features a combination of an L1–L2 and an L2–L1 dictionary for bidirectional search. The application can keep a record of users' search inputs and entry clicks. A questionnaire was also conducted to survey the students' evaluation of the dictionary application. Different from the negative results found by the original study, the replication revealed a non-significant effect of dictionary use on writing performance, providing solid evidence that a better dictionary leads to fewer consultation errors, although the improvement in writing scores brought about by dictionary use was only marginal. The study confirmed the original finding about the positive impact of dictionary use on lexical sophistication. It also identified some differences in dictionary lookup patterns between the participants of the replication and the original study in terms of search frequency, preference for language search, preference for search items, and use of source dictionaries. The implications of the study for dictionary making are discussed.

Keywords: REPLICATION, DICTIONARY USE, EFL WRITING, WRITING PERFORMANCE, LEXICAL SOPHISTICATION, LOOKUP PATTERNS

Opsomming: 'n Verdere kyk na die gebruik van 'n woordeboektoepassing in EVT-skryfwerk: 'n Repliseringstudie. In hierdie studie is die eksperiment van Chen en Liu (2022) waarin die effek van woordeboekgebruik op EVT-skryfwerk bestudeer is, gerepliseer. Buiten die gebruik van 'n ander woordeboek, het dit dieselfde navorsingsveranderlikes as die oorspronklike studie behels. Twee-en-sestig studente met Engels as hoofvak het twee skryfvoete afgeleë, een sonder die hulp van 'n woordeboek, die ander een met toegang tot 'n selfoonwoordeboektoepassing wat 'n kombinasie van 'n L1–L2- en L2–L1-woordeboek vir tweerigtingsoektogte bevat. Die

toepassing kan 'n rekord hou van gebruikers se soektogte en klikke op inskrywings. 'n Vraelys is ook voltooi om die studente se evaluering van die woordeboektoepassing te bepaal. Anders as die negatiewe resultate wat deur die oorspronklike studie verkry is, is daar in die replisering nie 'n beduidende effek van woordeboekgebruik op skryfprestasie nie, wat goeie bewyse verskaf dat die gebruik van 'n beter woordeboek tot minder naslaanfoute lei, alhoewel die verbetering in skryfprestasie deur die gebruik van 'n woordeboek slegs marginaal was. Die studie het die oorspronklike bevinding rakende die positiewe impak van woordeboekgebruik op leksikale sofistikasie bevestig. Dit het ook enkele verskille in woordeboeknaslaanpatrone tussen die deelnemers van die replisering en dié van die oorspronklike studie rakende soekfrekwensie, voorkeur vir taalsoektogte, voorkeur vir soekitems, en gebruik van bronwoordeboeke geïdentifiseer. Die implikasies van die studie vir woordeboekmaak word bespreek.

Sleutelwoorde: REPLISERING, WOORDEBOEKGEBRUIK, EVT-SKRYFWERK, SKRYFPRESTASIE, LEKSIKALE SOFISTIKASIE, NASLAANPATRONE

1. Introduction

1.1 Importance of replication studies

Within the empirical sciences, replication plays a major role in assessing the internal and external validity of findings and establishing predictable exceptions (Lindsay and Ehrenberg 1993, Gast 2009, Abbuhl 2012). It also helps to expose the weaknesses of the original study and improve the way we interpret empirical research (LTRP 2008: 1). However, such research is seldom attempted because it is difficult to successfully accomplish and it carries more risk than potential reward for both the replicator and the originator of the research (Park 2004: 194). In the field of second language acquisition (SLA), due to the lack of prestige and rewards associated with replication, it is not widely practiced either. In particular, replications in second language writing are virtually non-existent (LTRP 2008).

As regards dictionary use research, there are also relatively few studies openly acknowledged to be replications of some previous investigations (Dziemianko 2012: 199). Yet, similar to other areas of SLA research, increasingly "diverse in scope and investigation of topics" and thus resulting in "divergent and at times fragmented research results" (LTRP 2008: 11), a replication of dictionary use research is even more needed, valued and encouraged today than before.

Dziemianko (2010, 2011, 2012, 2017) conducted a series of approximate replications¹ to evaluate the role of dictionary media in language learning which involved the same battery of tests, participants with the same English proficiency and linguistic background, and the same experimental setting. Regardless of the different, if not contradictory results, the series of replications exhibit steady improvement in research methodology and give fascinating in-

sights into the way dictionary form affects language comprehension, production and retention. Such self-replications, rare as they are, prove worthwhile due to their potential to motivate the researcher to higher standards of replicability, to learn from his own experience or mistakes, and to improve or even reconceptualize his own methods (see LTRP 2008: 6).

1.2 Introduction to the original study

Dictionary use in L2 writing has not received due attention from researchers. As indicated by Chen and Liu (2022), some efforts have been made to investigate how the availability of dictionary impacts on writing performance (e.g. Tall and Hurman 2002, East 2007, Qiao and Wang 2020, Lew 2016), how dictionary consultation affects lexical accuracy and lexical sophistication (e.g. Nesi and Meara 1994, Christianson 1997, East 2006, Qiao and Wang 2020), and what lookup patterns and strategies are employed by L2 writers (e.g. Boonmoh 2012, Chon 2009, Lai and Chen 2015), but scarce are endeavors to explore the use of electronic dictionary by Chinese EFL learners in writing.

A most recent study by Chen and Liu (2022) examined the effect of dictionary use on writing performance, lexical sophistication and the search patterns and strategies of English majors at a Chinese university. In the first week, the students familiarized themselves with Bing.dict, an online bilingual dictionary, and filled in a questionnaire on their dictionary consultation habits, preferences, perceptions on the role of dictionary in EFL writing, and needs for dictionary instruction. One week later, the students were asked to write, without any dictionary, a 200-word composition on a given topic on a word processor. In the following week, they were instructed to write on another given topic with Bing.dict. A screen-recorder was preinstalled in the computers to record how the students consulted the dictionary to assist their writing.

The study found that Bing.dict produced a significantly negative effect on the students' overall composition scores and the component scores for content and language use as well, although it did play a part in increasing the students' lexical richness. A variety of dictionary-based errors were committed in terms of lexicon, syntax and collocation due to the students' inadequate dictionary use skills and the unsatisfactory quality of the dictionary for language production. Screen recordings demonstrated that the students employed a range of poor strategies for dictionary consultation which brought about undesirable results.

1.3 Motivations for the present replication

The motivations to self replicate Chen and Liu's study were twofold. To our knowledge, the original study is the first one to reveal a substantially adverse

impact of dictionary use on writing performance. This result diverges from the conclusion from Tall and Hurman (2002), Lew (2016), East (2006, 2007), and Qiao and Wang (2020). Therefore, it is necessary to seek more evidence to test the original finding. In addition, Bing.dict is a commercially minded AED (alternative e-dictionary, Nesi 2012) which combines diverse resources such as dictionaries of different types and online resources which are not produced by lexicographers at all. Despite its wide popularity and high evaluation score according to Lew and Szarowska's (2017) Framework, it was not considered an ideal dictionary for language production due to its serious defects (Chen and Liu 2022: 486). The authors ascribed the negative role of dictionary use in writing partly to the unsatisfactory quality of Bing.dict, suggesting that users exercise caution when turning to AEDs for language encoding. This gives rise to an intriguing question: What would happen if other non-AEDs were utilized in a similar setting? Would that negative result hold true with another dictionary? To determine the generalizability of the original conclusion, replication seemed necessary.

In the next section, the design of the replication will be introduced, covering the research questions, the participants, the writing topics, especially the dictionary used for the study. Research methods and major procedure of the study will also be elaborated. Section 3 will report the results of the replication in terms of writing performance, lexical sophistication and dictionary look up behavior. It will also discuss and compare the outcomes of the replication with those of the original. Section 4 will contain a summary of the findings, making suggestions for improving dictionary compilation. Finally, section 5 will explain the limitations of the replication and propose some topics for future research.

2. Replication to be tried

2.1 Research questions

The replication to be tried seeks to address basically the same research questions as those in the original study. Specifically, the following questions are formulated.

- (1) Does the mobile application of *New Century English–Chinese Chinese–English Dictionary* (henceforth the APP) have a significant effect on the participants' writing scores?
- (2) To what extent does the APP contribute to increasing the participants' lexical sophistication as measured by lexical frequency profile (LFP)?
- (3) What differences occur in dictionary lookup patterns between the participants of the replication and the original study?

2.2 Participants and writing topics

To overcome one of the limitations of the original study, i.e. a relatively small number of participants, we increased the sample size from 34 to 62. These participants bore remarkable similarities with those original ones: they were English sophomores at the same university, shared the same linguistic and cultural background, and were about to take TEM4 (Test for English Majors, Band 4) in three months.

The writing topics from the original study were also taken. One was concerned with money saving, the other was regarding making friends online, on the prerequisite of the participants having no prior experience in writing on similar topics.

2.3 The dictionary used

We chose the APP for the replication out of several considerations. First, as demonstrated by the original research, Bing.dict has undeniable defects, so it is reasonable to try another dictionary to test whether a change of dictionaries would lead to different outcomes. Second, the APP is the first of its kind in China to integrate a prestigious L2–L1 dictionary with a quality L1–L2 dictionary, both produced by the same publisher. It would be interesting to gauge the effectiveness of this new type of dictionary application for language production. Third, unlike the online version of Bing.dict accessed via a computer, the APP is installed in the user's mobile phone. According to dictionary surveys (e.g. Li 2015, Fan 2018, Gao and Yao 2020), mobile dictionary applications have gained immense popularity among Chinese EFL learners. We believe that introducing such a dictionary to the replication can bring the participants closer to their daily dictionary use scenario.

Developed jointly by Foreign Language Teaching and Research Press (FLTRP) and Shanghai Haidi Digital Publishing Technology Co., Ltd., the APP was launched in 2018. It combines *New Century English–Chinese Dictionary* (2016) (ECD)² with *New Century Chinese–English Dictionary* (2nd edition, 2016) (CED)³, representing a new type of "two in one" application which enables users to search words bidirectionally via a "jump" facility between the two source dictionaries.

Take the L1 search word "词典" (*cídiǎn*, dictionary) for example. When users log into the interface of the APP (see Screenshot 1) and enter in the search box the Chinese word "词典", a guide-page (see Screenshot 2) instantly pops up showing three English equivalents to "词典", i.e. *dictionary*, *lexicon*, *wordbook* and other L1 words and phrases containing "词典" and their corresponding English translations. A tap on whichever of the three equivalents to "词典" will lead users to the information page in the L1–L2 source dictionary of the APP, i.e. CED as demonstrated in Screenshot 3 which includes information about the pronuncia-

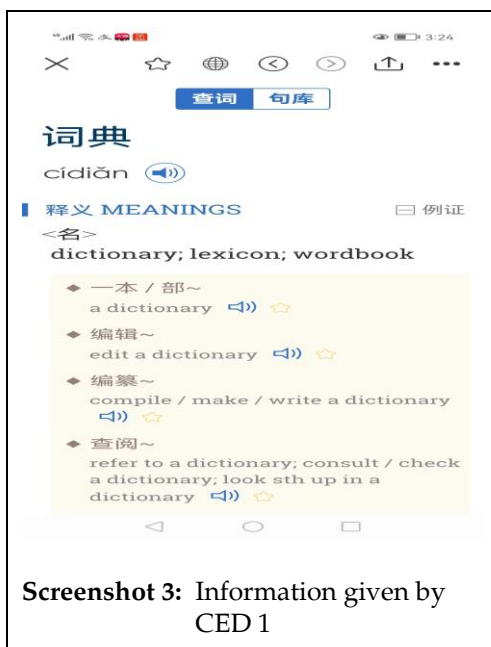
tion of "词典", its part of speech, meanings/equivalents, and auditory examples containing "词典". Users can further hit on each of the three equivalents to retrieve its specific information in the other source dictionary of the APP, i.e. ECD (L2-L1). For instance, if users tap on *dictionary*, the interface of the APP will "jump" from CED to ECD (see Screenshot 4), switching immediately from L1-L2 search to L2-L1 search (see Screenshot 5). ECD comprises phonetic information (both symbols and auditory), semantic information (both English explanation and Chinese equivalent/translation), and examples (both phrases and sentences) for the headword *dictionary*. It also gives lexical information on its part of speech and inflected form. Moreover, it includes frequency information, specifying that the headword *dictionary* belongs to the vocabulary for TEM 4, University Entrance Examinations and Graduate Admission Examinations, all being essential tests for different levels of students in China. Generally speaking, the design of the APP is simple and clear, offering convenient access routes and making dictionary consultation easy.⁴



Screenshot 1: Interface of the APP

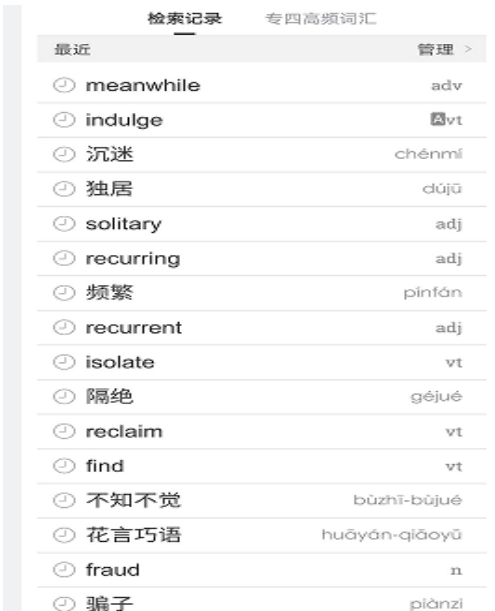


Screenshot 2: Guide-page for "词典"



2.4 Methods

Since users' search inputs and entry clicks are documented automatically in the Search Records, it is convenient to collect the data about what and how many words they retrieved and in what order. For example, it is evident from Screenshot 6 that the user looked up 16 words in the APP (9 English and 7 Chinese words) and that the search items were inclusive of both individual words and four-character Chinese idioms. By cross-checking the participants' Search Records and their compositions, we can learn about how they retrieved and applied dictionary information to their writing. In addition, a questionnaire survey was also undertaken to obtain feedback from the participants.



最近	管理 >
meanwhile	adv
indulge	vt
沉迷	chénmí
独居	dújū
solitary	adj
recurring	adj
频繁	pínfán
recurrent	adj
isolate	vt
隔绝	géjué
reclaim	vt
find	vt
不知不觉	bùzhī-bùjué
花言巧语	huāyán-qiǎoyǔ
fraud	n
骗子	piànzi

Screenshot 6: Search Records: Example

2.5 Procedure

The replication was implemented in the same experimental setting as the original study. In the first week, the students downloaded the APP into their mobile phones and received a brief training session about its structure, layout and usage⁵. Administered in the following week was Test 1 in which the students had 50 minutes to write a 200-word composition about money saving without any reference tool. One week later, the students were instructed, in Test 2, to write on another topic with the APP at hand (no other reference tool allowed).

After writing, the participants filled in a questionnaire about the usefulness, strengths and weaknesses of the APP and what they valued most for a good writing dictionary (see the Appendix).

2.6 Composition marking and analysis instruments

The compositions were sent to the two same evaluators in the original study and marked according to the same procedure and the same TEM 4 composition scoring rubric. Each composition was given an overall score (max = 20 points) together with separate scores for the three marking components, i.e. content (max = 10 points), structure (max = 3 points) and language use (max = 7 points). To check the inter-evaluator agreement, Pearson correlation coefficients were computed, as presented in Table 1.

Table 1: Correlation coefficient of scores on Test 1 and Test 2 between the two evaluators

	r (content)	r (language use)	r (structure)	r (overall)
Test 1	0.75	0.72	0.60	0.91
Test 2	0.75	0.81	0.62	0.92

Like in the original study, we employed RANGE 32, a lexical analysis tool developed by Heatley, Nation and Coxhead (2002) to compare the words in the compositions with the word lists for reference, including Base word 1 (approximately 1000 most-commonly-used English word families), Base word 2 (approximately 1000 second-commonly-used English word families), and Base word 3 (approximately 570 English word families). LFPs generated by RANGE contain information about (1) tokens, that is, all words in the composition; (2) types, that is, different words in the composition, and (3) families, that is, the base word, its inflections and its most common derivations (Laufer 2005), which can objectively reflect the students' choice and range of lexis. SPSS 20 software was also utilized for statistical processing.

3. Results and discussion

Results of the replication are analyzed and discussed from three aspects. Firstly, the effect of dictionary use on writing performance is examined. Secondly, the impact of the APP on lexical sophistication is evaluated. Thirdly, the differences in dictionary lookup behavior between the participants of the replication and the original study are explained.

3.1 Dictionary use and writing performance

In this subsection, the participants' writing scores in the two tests are computed to ascertain whether the differences are statistically significant. Dictionary-based errors are discussed in relation to what was found in the original study. The questionnaire data are also interpreted regarding the participants' evaluation of the APP.

3.1.1 Analysis of scores

Statistics demonstrated that in Test 1, the lowest and highest overall scores were 11.5 and 18.5 points respectively while in Test 2, the overall scores ranged between 10.0 to 17.0 points. The highest percentage of students (27.4%) in both tests scored between 15.0–15.9 points (see Table 2).

Table 2: Distribution of students' overall scores in the two tests (Max=20 points, N=62)

Overall scores	10.0-11.9	12.0-12.9	13.0-13.9	14.0-14.9	15.0-15.9	16.0-16.9	17.0-17.9	18.0-18.9
Number (Test 1)	2	3	14	7	17	14	4	1
Percentage (Test 1)	3.2%	4.8%	22.6%	11.3%	27.4%	22.6%	6.5%	1.6%
Number (Test 2)	1	4	9	13	17	9	9	0
Percentage (Test 2)	1.6%	6.5%	14.5%	21.0%	27.4%	14.5%	14.5%	0.0%

As displayed in Figure 1, the largest distribution difference in the overall scores across the two tests lies in the area of 14.0–14.9 points, with Test 2 scoring higher than Test 1 by 9.7% (21.0%–11.3%). However, in the areas between both 13.0–13.9 points and 16.0–16.9 points, Test 1 surpassed Test 2 by 8.1% (22.6%–14.5%). It seems that some students performed better when using the APP whereas for others, this was not the case.

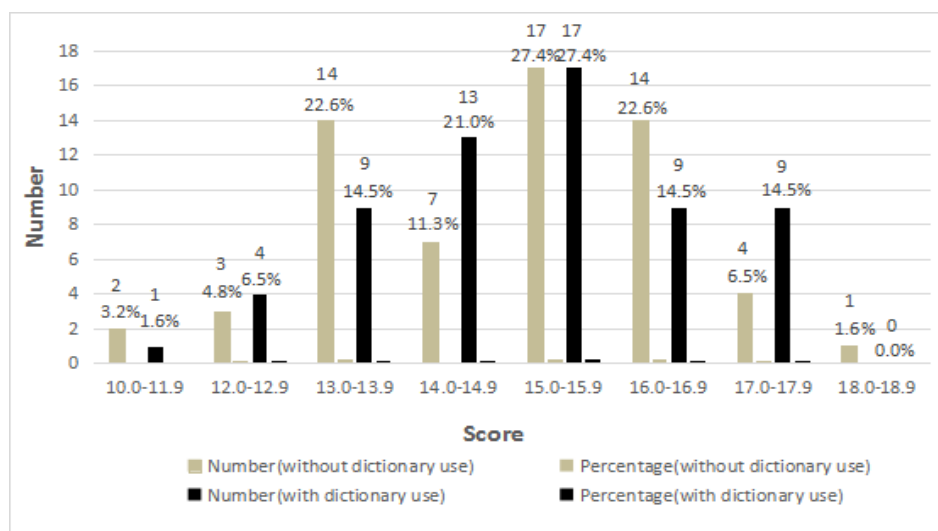


Figure 1: Distribution of the students' overall scores

Paired-Samples T-Tests were run to compute the students' scores (except structure scores) on Test 1 and Test 2. Table 3 indicates that the overall scores on Test 2 (M=14.95) are slightly higher than those on Test 1 (M=14.88), so are the scores for language use (M = 5.35 vs. M = 5.29), and the scores for content in the two tests look close (M=6.99 vs. M= 6.95).

Table 3: Descriptive statistics of scores on Test 1 and Test 2 (N=62)

	Mean	Std. Deviation	Std. Error Mean
Pair 1			
Content score (Test 1)	6.95	0.68	0.09
(Max= 10)			
Content score (Test 2)	6.99	0.67	0.08
Pair 2			
Language score (Test 1)	5.29	0.61	0.08
(Max= 7)			
Language score (Test 2)	5.35	0.64	0.08
Pair 3			
Overall score (Test 1)	14.88	1.49	0.19
(Max= 20)			
Overall score (Test 2)	14.95	1.50	0.19

Results of Paired-Samples T-Tests revealed that there was hardly any significant difference between the overall scores across the two tests [t (61)=-0.355,

$p=0.724$, two-tailed] (see Table 4). The difference between component scores on the two tests did not reach a significant level either for content [$t(61)=-0.354$, $p=0.725$, two-tailed] or for language use [$t(61)=-0.731$, $p=0.468$, two-tailed]. Apparently, the APP only made a marginal contribution to the students' writing performance.

Table 4: Paired-Samples T-Tests of scores on Test 1 and Test 2

	Paired Differences				t	df	Sig. (2-tailed)
	Mean	Std. Deviation	95% Confidence Interval of the Difference				
			Lower	Upper			
Pair 1 Content score (Test 1) - Content score (Test 2)	-0.322	0.718	-0.215	0.150	-0.354	61	0.725
Pair 2 Language score (Test 1) - Language score (Test 2)	-0.056	0.608	-0.211	0.098	-0.731	61	0.468
Pair 3 Overall score (Test 1) - Overall score (Test 2)	-0.645	1.433	-0.428	0.299	-0.355	61	0.724

In comparison with the original finding about a markedly negative impact of dictionary use on writing performance, the replication exhibited a more helpful role of dictionary use in writing. As will be illustrated, the APP induced very few dictionary-based errors and received favorable evaluation on its usefulness for writing. Nevertheless, it still failed to exert a significantly positive impact on writing scores. By performing Pearson Correlation analysis, we noticed a weak correlation between the students' overall scores on Test 2 and their frequency of APP consultation [$r=0.25$, $p=0.06 > 0.05$], implying that the higher scorers did not necessarily search more words.

3.1.2 Dictionary-based errors

A cross-examination of the students' Test 2 compositions and their Search Records uncovered 35 dictionary-based errors, inclusive of 14 collocation errors, 10 lexical errors, 8 syntactic errors and 3 other errors (see Table 5)⁶.

Table 5: Distribution of dictionary-based errors

Dictionary-based errors (35)	Lexical errors	10	28%
	Syntactic errors	8	23%
	Collocation errors	14	40%
	Other errors	3	9%

(1) Lexical errors: Some students opted for inappropriate or incorrect English equivalents to express their ideas out of confusion or misunderstanding of the semantic difference between synonymous English equivalents listed in the entries. The following are some examples taken from their compositions.

*"Initially, online dating is easy to be cheated." (The student was confused about the difference between *initially* and *first* when s/he looked up "首先" [*shǒuxiān*, in the beginning] in the APP.)

*"We should enjoy the convenience and *sake* that online dating brings to us and at the same time keep vigilant." (The student failed to notice the difference between *sake* and *benefit* when they both appeared in the dictionary guide-page for "好处" [*hǎochù*, benefit].)

*"Recently, the debate on whether it is wise to make friends online has *thrashed out*." (The student misunderstood the meaning of *thrash out*.)

(2) Syntactic errors: A few students committed syntactic errors when applying the retrieved items to writing, ignoring the part of speech or syntactic properties of words. For example:

*"If you *addict* to the friends online, you will be disjointed with people around you."

*"... if both of us realize we don't *fit to* each other."

*"Sometimes we come across private problems in real life and the Internet is a good platform for us to *vent*."

*"We *are too immersed ourselves in* the virtual online dating world, which will affect our interpersonal relationships in real life."

(3) Collocation errors: Some students used the retrievals correctly in grammar, yet the combination of words sounded unnatural. Like the original study, this type of errors made up the majority of the total, for instance:

*"Others make the use of *vulpine communication skills* to earn their trust."

*"Some people will over *indulge in friends* in the network."

*"Not only can it expand our friends circle but it can also *relieve our anxiety anonymously when talking to people we are familiar with trickily.*"

*"Recent surveys unveil that a large number of young man are more ready to be solitary due to *indulging to the talking online.*"

*"We would encounter a variety of persons that could *embrace frauds.*"

(4) Other errors: A couple of students made use of words stylistically inappropriately, blind to the style annotations in the entry. To illustrate, one student wrote, "You'll *forlese* the ability of associating with others in reality." It seemed when the student checked the APP, she overlooked the style annotation for *forlese*, i.e. 〈废〉 (*fèi*, obsolete) which indicates that the word dropped out of use. Another case in point is the awkward sentence, *"When you feel *ennuied* and got nothing to do at home ...". The dictionary does proffer an annotation "〈文〉" (*wén*, literary) to specify the style of the word *ennuied*, but the student obviously didn't notice it.

From the analysis above, it can be observed that some of the errors could be attributed to the students' inadequate skills of dictionary use such as choosing equivalents without further looking for their semantic difference, unable to model on dictionary examples to produce natural collocations, and ignoring dictionary annotations or other useful information. Nevertheless, instances of such inappropriate strategies of dictionary use were comparatively rare, most probably thanks to the clear and user-friendly design of the APP. Without the distraction of a multitude of web-crawled lexicographical information, the students had easy access to the reliable information from the two source dictionaries, hence fewer errors. Some of the dictionary-based errors were related to the students' English proficiency, especially their shaky grammatical foundation or weak awareness of collocation.

It should be noted that only two errors were induced by the problems inherent in the APP itself, one of which was due to inaccurate lexical information offered by the APP. It translates *sonnetize* into "沉迷于 (*chénmí yú*, indulge, be addicted, be obsessed with); 把...写入十四行诗 (*bǎ...xiě rù shí sì háng shī*)", the former part being incorrect, thus misleading one student to write, "I used to *sonnetize* in chatting with congenial net friends" when she searched an equivalent for "沉迷于". The other error resulted from insufficient dictionary examples. The APP renders two translations for *venturesome*, i.e. "好冒险的 (*hǎo màoxiǎn de*, venturesome); 大胆的 (*dàdǎn de*, daring)" and "有风险的 (*yǒu fēngxiǎn de*, risky); 危险的 (*wēixiǎn de*, dangerous)", without any examples to support its detailed use. Consequently, one wrote *"... we are *venturesome* to make friends online."

Compared with the original study in which 34 participants made 106 dictionary-based errors, the replication reported more optimistic data, with only 35 errors from the 62 participants. In other words, the average error rate was 0.56 per person in the replication, much lower than the figure in the original (3.12). With two more reliable source dictionaries and a clearer interface design, the APP

serves users with more accurate lexical information and easier access to dictionary data than Bing.dict, thus causing much fewer errors. The compelling evidence presented by the replication points to the fact that a better dictionary leads to fewer consultation errors, which highlights the importance of dictionary quality.

3.1.3 Responses to the questionnaire

Seven students were excluded from analysis due to incomplete or self-contradictory feedback, leaving a sample size of 55 for the questionnaire survey. According to the survey, none of the students had been familiar with the APP before. Youdao and Eudict are the two applications used most frequently by about 49% and 45% of the sample respectively, followed by Powerword (four students), Baidu Translate (two students), and Collins (one student).

As regards dictionary evaluation, Table 6 manifests that the APP was considered as very useful in rendering assistance for writing, as the mean score reached 9.3 out of a maximum of 10 points. It was also highly rated by a majority of the respondents (Mean=9.2) with respect to its convenience for dictionary search. In terms of the accuracy of dictionary information, the APP earned an average of 8.4 points. It seemed the APP was not as positively evaluated in terms of richness of dictionary information as in other evaluation dimensions, for it received a relatively low score (M=7.9). Generally speaking, the APP gained favorable recognition from the students.

Table 6: Evaluation scores of the APP (Max=10 points, Min = 0 points)

Evaluation score	Highest	Lowest	Mean
Accuracy of information	10	4	8.4
Richness of information	10	3	7.9
Convenience for word search	10	2	9.2
Usefulness for writing	10	2	9.3

This overall evaluation conforms to the responses from the students when asked about the strengths of the APP in comparison with other applications they have used. Half of the students agreed that the APP is more comprehensive in content, featuring rich ancillary learning resources. About 40% of the sample deemed it as more trustworthy due to its accurate lexical information, and some (28%) commented that it is more convenient in use, as it has a neat interface design.

According to over one third (35%) of the respondents, the major weakness of the APP lies in failing to fulfill their need for sentence translation. Some students complained of limited dictionary examples (30%) and lack of access route for phrase and collocation search (28%). Some (26%) responded that the lexical coverage is not wide enough and several students felt it to be expensive.

When it comes to how the APP can improve, 36% of the students hoped for more abundant information, especially on word disambiguation. Roughly one third (34%) expected a more user-friendly APP with more flexible search routes for multiword searches. About 28% desired more accessible examples and phrases. A few (15%) called for specialized columns like a writing guide.

When inquired about their ideas of an ideal writing dictionary, the students' opinions varied. The quality of lexical information was placed at the top by 31% of the sample, followed by a wide lexical coverage (20%), convenient access through keywords (17%) and dictionary brand (15%). A small number of students also maintained that a good writing dictionary should incorporate applicable sentence examples (7%), high quality sentence translations (5%) and should be free-downloadable, ad-free, and upgradable (2%).

3.2 Dictionary use and lexical sophistication

This subsection explores how the APP affected the participants in their choice of lexis during writing through an analysis of LFPs.

Like the original study, the two sets of compositions in the replication were also put into analysis, employing RANGE 32 to examine the students' choice of lexis in terms of tokens, types and families. As illustrated in Table 7, a majority of lexis were taken from Base word 1 with 84.02% for Test 1 and 80.25% for Test 2, indicating that the students mostly relied on the 1,000 most-commonly-used English word families. In Test 2, the tokens from the three Base words decreased respectively from 84.02% to 80.25%, from 5.86% to 4.41% and from 5.44% to 5.27%, whereas those from outside the lists boosted by 5.39% (from 4.68% to 10.07%). Likewise, in Test 2 the types from the three Base words presented a uniformed decreasing tendency compared with Test 1 while those from outside the lists went up from 20.81% to 27.75%. With regard to the families of lexis, except for Base word 2, there was an increase in both Base word 1 and Base word 3 by 38 and 15 families respectively. All this suggests that the students made use of less basic words and preferred more academic ones when accessible to the APP. By cross examining the students' Search Records and the results of LFP analysis for outside the lists in Test 2, we made a list of words retrieved from the APP like "abyss", "alienate", "authenticity", "celebrity", "detrimental", "harassment", "intangible", and "recap" etc. An overwhelming proportion of the words were searched only once and except for several misspelt words, there were no unattested or non-existent words.

Table 7: LFPs of the two sets of compositions

Profiles	Base word 1		Base word 2		Base word 3		Not in the lists	
	Test1	Test 2	Test 1	Test 2	Test 1	Test 2	Test 1	Test 2
Tokens/%	13153/ 84.02	13350/ 80.25	917/ 5.86	733/ 4.41	852/ 5.44	877/ 5.27	732/ 4.68	1675/ 10.07
Types/%	977/ 49.00	1025/ 45.07	299/ 14.99	290/ 12.75	303/ 15.02	328/ 14.42	415/ 20.81	631/ 27.75
Families	559	597	223	209	204	219	?	?

As Independent-Samples T-Tests (see Table 8) revealed, the differences in both tokens and types from "not in the lists" reached a statistically significant level ($p < 0.01$), implying that the students tended to use more advanced and sophisticated words when the APP was available. In addition, the tokens from Base word 2 between the two sets of lexis also differed remarkably from each other ($p < 0.01$). Although no substantial difference was found between the two sets of lexis from Base word 1 and Base word 3, there was an observable fall in the use of high-frequency words and a noticeable increase in more complex ones.

Table 8: Independent-Samples T-Tests on LFPs

	Base word 1	Base word 2	Base word 3	Not in the lists
Tokens				
t	-0.961	2.672	-0.579	-6.728
df	122	122	122	122
p	0.338	0.009**	0.564	0.000**
Types				
t	-1.892	0.492	-0.590	-2.593
df	122	122	122	122
p	0.061	0.623	0.556	0.004**
Families				
t	-1.204	0.267	-0.843	
df	122	122	122	
p	0.231	0.790	0.401	

** $p < .01$ (two-tailed).

The replication confirmed the original finding about the impact of dictionary use on lexical sophistication. However, despite the enhanced lexical range, scores on Test 2 were only marginally higher than those on Test 1 (see Table 3). In other words, the extent of richer lexis was not large enough to have a significant effect on the scores.

3.3 Dictionary lookup behavior

This subsection identifies the differences in dictionary lookup behavior between the participants of the replication and the original study, looking into the questions about who consulted the dictionary more frequently, what the preferred language input was, L1 or L2, what kind of lexical items were searched most often, and how the participants made use of the source dictionaries differently.

3.3.1 Frequency of dictionary searches

The data about the frequency of dictionary consultation was gathered from the Search Records which encompassed the words entered in the search bar and the items in the guide-page or the entries tapped for further or cross-reference. As displayed in Table 9, the students looked up 884 lexical items altogether with an average of 14.3 per person. The frequency of dictionary lookup varied from 2 to 47. Evidently, the students in the replication turned to the APP more frequently than those in the original study.

Table 9: Frequency of dictionary searches

	Total number	The average	The maximum	The minimum
The original (n=34)	405	11.9	31	0
The replication (n=64)	884	14.3	47	2

As evidenced in Table 10, the students performed more L2–L1 than L1–L2 consultation (542 vs. 342). A dominant number (92%) of L2–L1 searches were individual English words with one student consulting as many as 41 items. L2 multiple-word combinations accounted for no more than 8%, mostly phrasal verbs such as *bear upon*, *wear down*, and *fan out*, etc. Among the L1–L2 searches, individual Chinese words constituted the bulk (85%) with a maximum of 17 words, followed by 14% of four-character Chinese idioms and phrases such as "不知所措" (*bùzhīsuǒcuò*, all at sea), "喜怒哀乐" (*xǐnùāilè*, joy, anger, sorrow

and happiness), "难言之隐" (*nányánzhīyǐn*, *painful secret*), "随时随地" (*suíshísuídì*, *anytime and anywhere*), "网上聊天" (*wǎngshàng liáotiān*, *cyber chat*), "主旋律" (*zhǔ xuánlǜ*, *theme*) etc. There were only three search cases for Chinese sayings like "不怕一万, 只怕万一" (*bù pà yī wàn, zhǐ pà wàn yī*, *be prepared for the one risk in a million*). Most students switched between L1–L2 and L2–L1 searches, with six performing L2–L1 consultation exclusively.

Table 10: Descriptive statistics of the students' Search Records (N=62)

Total searches	L1–L2 searches			L2–L1 searches	
884	342			542	
Mean=14.3	L1 words	L1 idioms and phrases	L1 sayings	L2 words	L2 multiple-word combinations
	290/85%	49/14%	3/1%	500/92%	42/8%
	Mean=4.68	Mean=0.79	Mean=0.05	Mean=8.06	Mean=0.68
	Max=17	Max=7	Max=1	Max=41	Max=5
	Min=0	Min=0	Min=0	Min=0	Min=0

It is noteworthy that only two students applied all the words and phrases they retrieved from the APP to their writing. Of the 884 search items, 541 were actually put into the compositions. The average use of lexicographical information was about 61%. Most students utilized 40–80% of their retrievals, with the exception of three who used only about a tenth of their lookups.

3.3.2 Preference for language search

In the original study, the participants mostly attempted to obtain L2 equivalents or translations with L1 inputs dominating the scene. No one carried out L2–L1 searches exclusively. In contrast, the replication reported that the frequency of L2–L1 searches was 1.6 times that of L1–L2 searches (see Table 10). Six students conducted L2–L1 searches solely without a single L1 input. This difference might result from data collection method. The APP only records the lexical items that appear in the headwords or examples in the two source dictionaries. Consequently, inputs of some Chinese phrases, collocations and sentence fragments such as "潮湿的空气" (*chāoshī de kōngqì*, *moist air*), "开展活动" (*kāizhǎn huódòng*, *carry out activities*), "谨防上当受骗" (*jǐnfáng shàngdāng shòupiàn*, *beware of being cheated*) cannot be documented in the Search Records unless the translations of such items are included in the APP either as headwords or as entry examples. In other words, the APP cannot keep track of the Chinese phrases,

collocations or other multiword expressions which go beyond the coverage of the source dictionaries.

3.3.3 Preference for search items

One notable lookup pattern discovered by the original study was that many participants tended to seek English translation for Chinese sentences. This formed a striking contrast with the replication where the searches of individual words made up the bulk. Bing.dict features a multitude of web-crawled lexicographical information as well as automatic machine translation, which renders sentence translation possible, but in many cases the translation is of poor quality, if not ridiculous, and only misleads users. By comparison, since the source dictionaries in the APP are essentially the electronic versions of the original print dictionaries without fundamental changes in content and structure, they are unable to cater for users who try to look for sentence translation. Moreover, the link directing users to Google Translate when the lookup items go beyond the lexical coverage of the APP is currently inaccessible in Mainland China.

Another difference consists in the consultation of basic words. Some participants in the original study looked up high frequency Chinese words like "通常" (*tōngcháng*, usually), "其次" (*qícì*, next), and "第二" (*dìèr*, secondly). However, the replication identified only two such instances. This divergence might arise from the participants' overall English proficiency. Despite the similar linguistic proficiency of the participants, students in the replication achieved a higher, though not significantly, average score on Test 2 than those in the original study ($M = 14.9$ vs. $M = 12.8$), which implied that they can use English more competently and may not feel the need to look up high frequency words.

3.3.4 Use of source dictionaries

The original participants mostly depended on Internet-generated lexicographical information, neglecting the source dictionaries in Bing.dict. Only 9 out of 32 participants further clicked on them for cross-reference and no one ever hit on the tabs in the bilingualized source dictionary to read examples. The replication showed a different picture, for as many as 54 students switched between L1-L2 and L2-L1 searches, meaning the majority of the students made use of both source dictionaries.

This difference can be ascribed to the interface design of the dictionaries involved. In Bing.dict, the Internet-generated translations and web-crawled sentence examples are posted in a conspicuous spot, dwarfing the source dictionaries on that score. In contrast, the APP is based on only two source dictionaries without accessible links to extra lexicographical resources. Its interface looks clean and clear, making easy the "jump" from one dictionary to the other. Without access to extra lexicographical information from the Internet, users have no alternative but to focus on the two source dictionaries.

4. Conclusion

Three findings emerge from the replication. Firstly, the use of the APP has a non-significant effect on the participants' writing performance, distinct from the original conclusion about the negative role of dictionary in EFL writing. The APP proves to be more helpful for encoding, as it gave rise to a much smaller number of dictionary-based errors than the original dictionary, suggesting that the better a dictionary is, the fewer consultation errors it will cause.

Secondly, dictionary use did enhance the participants' lexical sophistication, although this advantage was not significant enough to make a marked difference in writing scores. This conforms to the original conclusion.

Thirdly, some differences in dictionary lookup behavior were detected between the participants of the replication and the original study. Students in this study consulted the APP more frequently. They entered or tapped on more L2 items than L1 ones, mainly looking up individual words and paying more attention to the source dictionaries than the original participants. Moreover, they committed far fewer dictionary-based errors, chiefly owing to the more authoritative source dictionaries and the well-designed dictionary interfaces.

The study shows that the APP, as a pioneering "two in one" product at the Chinese lexicographic market, is more effective than Bing.dict for EFL writing, yet it did not exert a significantly positive effect on writing performance. The questionnaire survey reflected that the APP was highly ranked in terms of usefulness for writing, accuracy of dictionary information and convenience in use. However, it was also perceived to have some weaknesses such as lack of search function for some phrases, collocations and sentences, limited lexical coverage, and insufficient dictionary examples, etc. The participants expressed their hopes for more useful lexical information such as word disambiguation, a wider vocabulary coverage, richer dictionary examples, and easier access routes to multiword search. To better satisfy users' needs and expectations, improvements should be made to the APP with regards to the above-mentioned areas.

According to the questionnaire survey, the participants held a variety of opinions concerning the criteria for an ideal electronic dictionary for EFL writing. It seemed the quality of dictionary information was prioritized by most students, followed by multiple and convenient access routes. Practical information categories such as derivatives, synonym disambiguation, collocations, sentence examples, and a writing guide were also among the list. Conceivably, the participants' feedback can serve as useful advice for dictionary optimization.

ECD and CED are rated among the best bilingual dictionaries in China, receiving positive recognition from lexicographical experts and users (Wang et al. 2019). However, it should also be pointed out that they are general linguistic dictionaries in nature, different from learners' dictionaries in at least four major aspects, i.e. the target users, lexical coverage, sense arrangement and the amount of lexical information crucial for language production such as collocations and examples. Take lexical coverage for example. The APP covers

low-frequency words and archaic or even obsolete words, which may lead users to select unfamiliar words due to the misconception that the rarer the word, the better. By comparing collocations and examples in the APP with those of LDOCE (*Longman Dictionary of Contemporary English*)⁷, one of the most well-known English learners' dictionaries in the world, we believe that the former leaves much to be desired in these aspects. The APP could have a more beneficial effect on writing if it had been equipped with a better encoding function. It is a pity that despite the remarkable progress in China's practical lexicography, there is still a long way to go in the compilation of production-oriented L1–L2 learners' dictionaries.

5. Limitations of the study and suggestions for future research

The study is not without limitations. Since the Search Records can only keep track of search items which fall within the lexical coverage of the two source dictionaries in the APP, we were unable to know what and how many invalid searches were performed by the participants. Such a problem could have been avoided if we had relied on screen recording for data collection as we did in the original study. Besides, due to the restriction of the research method, we could not learn about the cognitive aspects of dictionary consultation such as what prompted a particular search, how the participants chose among equivalents, how they dealt with lexical issues when failing to retrieve needed information from the APP, and why many L2 lookups were not used in writing. Think-aloud protocols or a follow-up interview would help to elicit some interesting information about the students' cognitive processes and strategy use.

The dictionary per se is a crucial variable when testing the effect of dictionary use. To gain more insights into the impact of dictionary use on writing, it is advisable to carry out more replications. Future research may try another type of dictionary, especially production-oriented dictionaries with user-centered design. Dictionary use competence constitutes another important factor influencing the outcomes of dictionary consultation. Therefore, it is necessary to involve participants of different proficiency levels such as English-majoring MA students or skilled dictionary users in further studies. In addition, other types of writing tasks (e.g. free-topic writing), or other forms of language production (e.g. L1–L2 translation), are also considerable. Finally, more explorations can be attempted to develop writing assistants and check their effectiveness for language production.

Endnotes

1. Approximate (also known as partial or systematic) replication involves repeating the original study exactly in most respects, but changing one of the non-major variables so as to allow for comparability between the original and replication study (Abbuhl 2012: 298).

2. Based on the Collins Corpus, ECD is a general linguistic dictionary with a coverage of over 250,000 words and 350,000 senses. It encompasses concise and accurate definitions, rich and comprehensive information and a wide coverage of new words and senses, representing the status quo of the English language.
3. CED is hailed as the first work of the fourth generation Chinese–English dictionaries. It covers more than 150,000 headwords, highlighting linguistic information and including encyclopedic information. It is an official dictionary used for China Accreditation Test of Translators and Interpreters. Due to its innovation in terms of lexical coverage, definition, examples, translation, and part-of-speech tagging, it has won several national awards.
4. The APP has remarkable user-friendly features such as a multitude of functions for customization in language learning and abundant learning resources including the Chinese ideological and cultural terminology, special Chinese–English columns (such as Chinese four-character idioms, proverbs and particularized sayings), and special English–Chinese columns (such as English phrases and idioms, usage notes, cultural columns and collocations). Since those features were irrelevant to the replication, they are mentioned briefly here.
5. Thanks to the general support of FLTRP, all students in the experiment had a three-month free access to the APP.
6. Due to the complexity and challenges involved in the classification of errors, some of the errors might fall into more than one category, some might be borderline cases and some might be hard to categorize.
7. *Longman Dictionary of Contemporary English Online* can be accessed via <https://www.ldoceonline.com/>.

Acknowledgements

This research is part of the project "Constructing an Evaluation Framework for Bilingual Dictionary APPs in the Digital Era" (No. FJ2023B031) funded by the Fujian Social Science Foundation. We would like to extend our sincere gratitude to the anonymous reviewers who dedicated their precious time to reading our paper and making insightful comments and suggestions. We are also grateful to those students who participated in the experiment. In addition, heartfelt thanks also go to our colleagues, Professor Hou and Ms. He, who helped to proof-read our paper.

References

- Abbuhl, R.** 2012. Why, When, and How to Replicate Research. Mackey, A. and S.M. Gass (Eds.). 2012. *Research Methods in Second Language Acquisition: A Practical Guide*: 296-312. Oxford: Blackwell Publishing.
- Boonmoh, A.** 2012. E-dictionary Use under the Spotlight: Students' Use of Pocket Electronic Dictionaries for Writing. *Lexikos* 22: 43-68.
- Chen, Y.Z. and S.P. Liu.** 2022. Exploring the Use of an Online Bilingual Dictionary in EFL Writing. *International Journal of Lexicography* 35(4): 468-490.

- Chon, Y.V.** 2009. The Electronic Dictionary for Writing: A Solution or a Problem? *International Journal of Lexicography* 22(1): 23-54.
- Christianson, K.** 1997. Dictionary Use by EFL Writers: What Really Happens? *Journal of Second Language Writing* 6(1): 23-43.
- Dziemianko, A.** 2010. Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. *International Journal of Lexicography* 23(3): 257-273.
- Dziemianko, A.** 2011. Does Dictionary Form Really Matter? Akasu, K. and S. Uchida (Eds.). 2011. *ASIALEX 2011 Proceedings. Lexicography: Theoretical and Practical Perspectives*: 92-101. Kyoto: Asian Association for Lexicography.
- Dziemianko, A.** 2012. Why One and Two Do Not Make Three: Dictionary Form Revisited. *Lexikos* 22: 195-216.
- Dziemianko, A.** 2017. Dictionary Form in Decoding, Encoding and Retention: Further Insights. *ReCALL* 29(3): 335-356.
- East, M.** 2006. The Impact of Bilingual Dictionaries on Lexical Sophistication and Lexical Accuracy in Tests of L2 Writing Proficiency: A Quantitative Analysis. *Assessing Writing* 11(3): 179-197.
- East, M.** 2007. Bilingual Dictionaries in Tests of L2 Writing Proficiency: Do They Make a Difference? *Language Testing* 24(3): 331-353.
- Fan, K.** 2018. An Investigation of English Major Students' Use of English Dictionaries. *Lexicographical Studies* 6: 24-32.
- Gao, X. and K.X. Yao.** 2020. Survey and Prediction of the Usage of College Students' Mobile Phone Dictionary. *Jiang Su Technology and Information* 37(19): 63-65.
- Gast, D.L. (Ed.).** 2009. *Single Subject Research Methodology in Behavioral Sciences*. New York: Routledge.
- Heatley, A., P. Nation and A. Coxhead.** 2002. RANGE and FREQUENCY Programs [EB/OL]. <http://www.vuw.ac.nz/lals/staff/Paul-Nation>. [2021-03-20]
- Lai, S.-L. and H.-J. Howard Chen.** 2015. Dictionaries vs Concordancers: Actual Practice of the Two Different Tools in EFL Writing. *Computer Assisted Language Learning* 28(4): 341-363.
- Language Teaching Review Panel (LTRP).** 2008. Replication Studies in Language: Learning and Teaching: Questions and Answers. *Language Teaching* 41: 1-14.
- Laufer, B.** 2005. Lexical Frequency Profiles: From Monte Carlo to the Real Word. A Response to Meara. *Applied Linguistics* 26(4): 582-588.
- Lew, R.** 2016. Can a Dictionary Help You Write Better? A User Study of an Active Bilingual Dictionary for Polish Learners of English. *International Journal of Lexicography* 29(3): 353-366.
- Lew, R. and A. Szarowska.** 2017. Evaluating Online Bilingual Dictionaries: The Case of Popular Free English-Polish Dictionaries. *ReCall* 29(2):138-159.
- Li, J.Y.** 2015. A Survey of Mobile-phone Dictionary Use by University Students in China's Remote Minority Regions. *Lexicographical Studies* 3: 38-47.
- Nesi, H.** 2012. Alternative e-Dictionaries: Uncovering Dark Practices. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 369-378. Oxford: OUP.
- Nesi, H. and P. Meara.** 1994. Patterns of Misinterpretation in the Productive Use of EFL Dictionary Definitions. *System* 22(1): 1-15.
- Park, C.** 2004. What Is the Value of Replicating Other Studies? *Research Evaluation* 13(3): 189-195.

- Qiao, L.T. and W.Y. Wang.** 2020. An Empirical Study of the Impact of Dictionary Use on Learners' Lexical Performance in L2 Writing: Chinese L2 Learners as an Example. *Foreign Languages Research* 183(5): 36-42.
- Tall, G. and J. Hurman.** 2002. Using Dictionaries in Modern Language GCSE Examinations. *Educational Review* 54(3): 205-217.
- Wang, R.Q., Z.Z. Huo and J. Deng.** 2019. A Study of the Representation Strategies of Heterosemy in *A New Century Chinese-English Dictionary* (2016). *Foreign Languages and Literature* 35(2): 11-22.

Appendix

A questionnaire on the use of the application of *New Century English–Chinese and Chinese–English Dictionary (the APP)*

Instructions: Please write an answer to or put a tick at the answer of the following questions.

1. Was it the first time you used this APP? Yes. No.
2. Please write the name of the mobile phone APP you use most often.

3. Please rate your satisfaction of the APP (from low to high, the full score is 10) according to your dictionary use experience.
 - ☆ the accuracy of dictionary information: 1 2 3 4 5 6 7 8 9 10
 - ☆ the richness of dictionary information: 1 2 3 4 5 6 7 8 9 10
 - ☆ the convenience for dictionary research: 1 2 3 4 5 6 7 8 9 10
 - ☆ the usefulness of the APP for writing: 1 2 3 4 5 6 7 8 9 10
4. Compared with other electronic dictionaries you use, what do you think are the advantages of the APP?
5. In your opinion, what are the weaknesses of the APP?
6. In what ways can the APP improve?
7. Do you think the APP is affordable or not at a price of 138 RMB?
8. What do you think are the features of an ideal electronic dictionary for writing?