Chapter 3
A Comparative Study of Chinese Test Takers' Writing Performance in Integrated and Discrete Tasks: Scores and Recurrent Word Combinations in PTE Academic

Yu-Hua Chen and Ying Zheng

Abstract Despite an increasing number of studies on L1 Chinese students' L2 English writing in recent years, little research is conducted on Chinese test takers' performance in high-stakes international tests, particularly in terms of how integrated and discrete assessment may impact on their writing. This book chapter, therefore, aims to fill this gap by comparing Chinese and non-Chinese test takers' responses in different writing tasks of PTE Academic in terms of scoring and use of recurrent word combinations. As an international test of academic English, PTE Academic includes both the traditional design of independent essay writing as well as two integrated tasks of summary writing from listening or reading input.

Written samples from 500 Chinese test takers and another 500 test takers of other L1s were randomly chosen from PTE Academic and analysed. The results indicate that Chinese test takers outperformed their non-Chinese peers in the tasks of read-to-summarise and essay writing. Chinese test takers also tended to write longer responses with more occurrences of recurrent word combinations, and they used more stance expressions in the essays compared with their peers. A closer examination of highly frequent word combinations, however, suggests a relationship between task type and the language elicited, regardless of test takers' L1.

Keywords Academic writing · Task type · Integrated writing · Recurrent word combinations · PTE Academic

Y.-H. Chen (✉)
Coventry University, Coventry, UK
e-mail: yu-hua.chen@coventry.ac.uk

Y. Zheng
University of Southampton, Southampton, UK
e-mail: ying.zheng@soton.ac.uk

L. Hamp-Lyons, Y. Jin (eds.), *Assessing the English Language Writing of Chinese Learners of English*, https://doi.org/10.1007/978-3-030-92762-2_3

3.1 Introduction

As a computer-based international test of academic English, Pearson Test of English Academic (hereafter PTE Academic) contains a variety of integrated writing tasks, including listen-to-write and read-to-write items which require test takers to produce a summary based on a given passage. Traditionally, a typical writing task involves only writing skills, hence often known as discrete or independent writing assessment, whereas an integrated task uses listening or reading input (or a combination of both as in the case of TOEFL iBT) as the source of information for language production, hence also known as source-based writing (e.g., Merkel, 2020). Although an increasing number of studies in recent years have investigated the linguistic features in Chinese students' L2 English writing (e.g., Chen & Baker, 2010, 2016; Leedham & Cai, 2013), very little research is conducted on Chinese test takers' performance in integrated writing assessment in relation to the traditional design of independent writing assessment. This study, therefore, aims to fill this gap by investigating Chinese test takers' writing performance in terms of scores and use of recurrent word combinations in PTE Academic, where a task of independent essay writing is used to measure writing ability as well as two integrated summary writing tasks that also evaluate reading or listening comprehension.

Recurrent word combinations are recurring word sequences extracted by computers with a set of selection criteria. This type of formulaic sequences has been identified as "building blocks" of discourse (Biber et al., 2004), and their occurrences and use in the written samples from Chinese and non-Chinese candidates will be compared between independent and discrete tasks in the current study.

One thousand test takers in total were randomly selected (500 each for Chinese and non-Chinese candidates) from PTE Academic, and four summaries (two each for the integrated tasks of read-to-summarise and listen-to-summarise) and one essay from each of the test takers in one single sitting of the test were extracted. With Chinese test takers as the focus of this study, the non-Chinese group was used as a reference to determine whether the patterns identified was unique to the Chinese group only. By comparing Chinese test takers' scores of those writing tasks and written samples with test takers from other L1 backgrounds, we will be able to identify distinctive features in L1 Chinese learners' L2 English writing between independent and integrated writing items. The results can then be used to help us better understand issues in Chinese learners' academic writing in different task types and provide insights into the teaching, learning or test development for the area of English for Academic Purposes (EAP).

3.2 Literature Review

3.2.1 Integrated vs. Independent Writing Assessment

Traditionally, writing ability is often assessed through a timed task independent of other skills such as reading or listening. Independent tasks are perhaps still the most popular approach in assessing L2 writing ability because its design is straightforward and marking also tends to be easier in comparison to integrated assessment which involves more than one skill. However, this 'snapshot' approach has been criticized because of its limited capacity (Hamp-Lyons & Kroll, 1996) as well as the lack of authenticity or validity (e.g., Cho, 2003; Read, 1990; Weigle, 2004). One of the greatest challenges confronted with independent writing assessment is its reliance on topic familiarity or creativity to a large extent. Since writing is notoriously time-consuming, a test often includes only one or two tasks which evaluate writing skills. If a test taker, however, is not interested in or not familiar with the given topic(s), this potentially place them at a disadvantage.

Compared to independent tasks such as timed essays, integrated tasks appear to have a higher degree of authenticity. Cumming et al. (2000) argue that real-life writing often involves writers drawing on various resources available during the writing process rather than just a given topic. Similarly, Weigle (2004) and Sawaki et al. (2013) point out that in real academic settings, English learners are expected to synthesize what they read or hear into written works. This seems to suggest that academic writing should be integrated with other skills. In relevant literature in Higher Education, this is sometimes referred to as source-based writing, and the focus typically surrounds the notion of how student writers engage with source materials (e.g., Hirvela & Du, 2013; Li & Casanave, 2012; Merkel, 2020). According to a meta-analysis conducted by Cumming et al. (2016), they compared 69 empirical studies on source-based academic writing and concluded that source-based writing often imposes a challenge for students, particularly for L2 students because of different prior knowledge and experience.

Accordingly, the combination of different skills and source materials into writing tasks under the general term of integrated assessment reflects the reality of academic writing that test takers will encounter in their future studies. However, Buck (2001) points out the challenges of identifying the construct of integrated tasks. After reviewing a number of studies which report on integrated writing assessment, Cumming (2013) also concludes that integrated writing tasks "confound the measurement of writing ability with abilities to comprehend source materials". Although there has been an increasing interest in integrated writing assessment in recent years, not much research can be found regarding the inclusion of both read-to-write and listen-to-write tasks. Zheng and Mohammadi's study (2013), which explored the constructs of six writing item types in PTE Academic via the employment of exploratory factor analysis, is probably one of the few studies that have offered an

insight into not only read-to-write tasks but also listen-to-write tasks. Considering the high stakes of university admission tests such as PTE Academic, it is therefore important for us to better understand the relationship between various types of integrated writing and independent writing.

3.2.2 Research into L2 English of L1 Chinese Students

The motivation of focusing on Chinese students' L2 academic English writing is initiated by an increasing number of Chinese students who choose to study in English-speaking countries or EMI (English as Medium of Instruction) institutions. Accordingly, there is also growing literature which reports Chinese students' L2 English performance in comparison with other L1 groups. For example, previous studies have explored the potential link between test takers' first languages and their performance in English tests (Abbott, 2007; Chen & Henning, 1985; Kim, 2001), and one of the earliest attempts to investigate this relationship was a study conducted by Chen & Henning (1985). By comparing the responses from native speakers of Chinese and Spanish in an EFL test of five parts: listening, reading, grammar, vocabulary, and writing error correction, they identified vocabulary items which favoured the Spanish test takers. Also comparing the performance between Chinese and Spanish test takers, Sasaki (1991) found that vocabulary items with idiomatic expressions favoured the former group while English-Spanish cognates favoured the latter. The tradition of comparing a group of a specific L1 learners lays a foundation for the current study to compare L1 Chinese test takers with their peers of other L1s.

In relation to sourced-based writing discussed in the last section, Shi (2004) compared Chinese university students' L2 English writing with an L1 English student group. The written samples were collected from two types of tasks – writing an opinion essays and writing a summary, both of which were given a source text as the prompt. The findings indicate that students in general copied more words from the source text in the summary task than in the opinion essay task. Shi (ibid.) also found that Chinese students tended to borrow words from source text without proper citations when compared with the group of L1 English students.

The examples above illustrate that Chinese learners of English appear to perform their writing with certain linguistic features specific to their L1 background. Together with these distinguishing features, the increasing number of Chinese students pursuing a degree abroad in English, hence having to take high-stakes international tests of English (MOE, 2018), highlights the necessity for studies on Chinese test takers' written performance in these "gatekeeping" tests. The current study therefore addresses this call for research and hopes to provide some pedagogical implications for EAP courses catering for Chinese students.

3.2.3 *Recurrent Word Combinations in L2 Writing*

Functioning as "building blocks" of discourse (Biber et al., 2004), recurrent word combinations are also known as lexical bundles. Recurring word sequences are typically retrieved using a computer tool with specified frequency or dispersion thresholds. Recent research indicates that there are significant differences in terms of occurrences and discourse functions of recurrent clusters between genres (Biber et al., 1999, 2004; Biber & Barbieri, 2007), disciplines (Hyland, 2008), between L1 and L2 writers (e.g., Ädel & Erman, 2012; Chen & Baker, 2010), or across L2 proficiency levels (Chen & Baker, 2016).

Focusing on Chinese students' L2 English writing, Chen & Baker (2016) examined the use of lexical bundles and found that at lower levels, the writing discourse from Chinese learners shared more features with that of conversation, while the discourse of more proficient writing was more similar to that of academic prose. Similarly, Ruan (2016) looked at a corpus of academic texts written at four points of time between Year 1 and Year 4 at an EMI university in China to identify frequently used lexical bundles. The findings suggest that there is a developmental pattern of lexical bundle use in terms of both structures and discourse functions.

Staple et al.'s (2013) work is one of the few studies where the use of recurrent word combination was investigated and integrated tasks were included in the writing samples from a high-stakes English test. In their study, lexical bundle use was compared across three levels in the writing tasks of TOEFL iBT: high, intermediate and low. The findings indicate that test takers at lower levels overall used more bundles and also "borrowed" more bundles from the given prompts. Lexical bundles in this study were divided into two groups: prompt and non-prompt bundles, with the former referring to those "that appeared word for word and that are clearly related to the topic or task" (ibid.: 217). Based on our discussion about source-based writing earlier, it is reasonable to assume that test takers would "borrow" more words from the integrated tasks, hence more prompt bundles, because such tasks require test takers to use the source material to produce a summary. In Staple et al. study (2013), the written samples from the integrated and discrete writing tasks, however, were not distinguished when results were presented. It is therefore impossible to see whether there was any difference in terms of prompt bundles versus non-prompt bundles between integrated and independent writing tasks.

Another issue in traditional research on second language writing (or learner corpus research in recent years) is that L2 writing from learners of a certain L1 background is often compared with L1 writing (e.g., Granger, 1998; Granger et al., 2002; Ädel & Erman, 2012) rather than L2 learners' peers from other L1 backgrounds. This might lead to a potential misconception that Chinese students' writing, for example, seem to have plenty of issues such as overuse or underuse of certain language expressions, but without a comparison with other L2 learners, we cannot possibly determine whether those patterns are universal across different L2 students or unique in Chinese students' writing.

Taking into account relevant research discussed above, the current study will therefore aim to compare Chinese test takers' scores and use of recurrent word combinations (specifically prompt and non-prompt based ones) in the integrated and discrete writing tasks with those from non-Chinese test takers.

3.3 Research Questions

Focusing on Chinese students' writing performance in the integrated and discrete tasks in PTE Academic in relation to the non-Chinese group, the research questions are formulated as the following:

RQ1. How do Chinese test takers perform in the integrated and discrete writing tasks in comparison to non-Chinese test takers?
RQ2. To what extent are there differences of overall frequencies of recurrent word combinations between the integrated and discrete writing tasks in these two test taker groups?
RQ3. Are there differences, if any, of prompt and non-prompt recurrent word combinations between integrated and discrete writing tasks in these test taker groups?
RQ4. How are the discourse functions of non-prompt recurrent word combinations different or similar between integrated and discrete writing tasks in Chinese test takers' writing in relation to non-Chinese test takers?

3.4 Data and Methodology

3.4.1 Test Takers

Focusing on the comparison between independent writing and integrated writing tasks, 1,000 test takers were randomly selected from L1 Chinese and other L1 backgrounds (500 each), and their scores and written responses were used for comparison.

The demographic backgrounds of those chosen test takers can be found in Table 3.1. As those test takers are randomly selected, it is reasonable to assume that they represent the actual test taker population to a large extent. Among the non-Chinese test takers, the largest group comes from India, which accounts for over half of the international test takers (53.6%), while the remaining test takers primarily come from Southern or South-east Asia. It is also interesting to see that 21.4% of those international test takers speak English as their home language. Among those speakers of English, 43.9% again come from India while the others cover a range of other countries such as Malaysia, Pakistan, Philippines, Singapore, South Africa or U.K.

Table 3.1 Comparison of Chinese and non-Chinese test takers

	Chinese	non-Chinese (International)
Number	500	500
Nationality	China PRC 100%	India 53.6% Pakistan 7.2% Nepal 6.6% Philippines 4.2% South Korea 2.6% Vietnam 2.6% Others 23.2%
Home language	Mandarin 100%	English 21.4% (43.9% from India) Hindi 12.4% Punjabi 10.2% Urdu 8.2% Nepalese 5.6% Telugu 5.6% Others 36.6%

Table 3.2 An overview of the three writing item types in PTE Academic

Communicative skills	Item type code	Item type	Brief description	Enabling skills
Reading & writing	RW-SUMM	Summarize written text	Summarize written text in one sentence of no more than 75 words	Content; form; grammar; vocabulary
Listening & writing	LW-SUMM	Summarize spoken text	Summarize spoken text in 50–70 words	Content; form; grammar; vocabulary; spelling
Writing	WW-ESSA	Write essay	Write an argumentative essay (200–300 words) in response to a given topic	Content; form; development, structure and coherence; grammar; general linguistic range; vocabulary; spelling

3.4.2 Test Tasks

An overview of the three writing tasks from PTE Academic and the skills measured can be found in Table 3.2. Each of the writing task type is coded in a similar way: the first two letters in the first part indicate the skills assessed (e.g., RW for reading and writing), and the four letters in the second part refer to the task (e.g., SUMM for summary writing). The task type code will be used in the rest of this chapter when a specific task is discussed.

In addition to an overall score, the score report that a test taker receives after completing the test also contains communicative skill scores (for writing, speaking, listening and reading) and enabling skill scores (i.e. productive subskills such as content or vocabulary for writing and speaking tasks only). The automated scoring system used in PTE Academic is trained and calibrated on the trait scores of response

Table 3.3 Number of items used for each of the Chinese and non-Chinese groups in the current study

		Chinese	Non-Chinese
Skills	Item type code	Number of responses	Number of responses
Reading & writing	RW-SUMM	1000	1000
Listening & writing	LW-SUMM	1000	1000
Writing	WW-ESSA	500	500
Total		2500	2500

samples scored by human markers, and the traits measured in PTE academic include a number of enabling skills (Pearson, 2018). Enabling skills in Table 3.2 therefore provide valuable information about the linguistic features measured for the two summary writing and the independent writing tasks in PTE Academic. As can be seen, content is scored for each of the writing tasks, and language ability is evaluated differently between summary and discrete writing tasks as the latter covers a wider range of traits.

For the group of 500 Chinese test takers, 1000 item responses for each of the integrated task types (two items each test taker for the RW-SUMM and LW-SUMM tasks) were extracted. Because each test paper only had one essay item (WW-ESSA), only 500 of them were available. This means two listen-to-write, two read-to-write summary tasks as well as one essay task for each of the test takers, amounting to 2500 responses in total (see Table 3.3). The same procedure was completed for the non-Chinese group. It has to be noted that because the test paper delivered to a test taker on a computer is randomly assigned and each paper consists of different items, it is impossible to control the prompts between the Chinese and non-Chinese test takers, which might have an impact on the statistical and linguistic analysis to some extent. Yet since the data was randomly selected from a large pool (as will be discussed later), we believe such an impact should be quite minimal.

3.4.3 Procedures

The data including item responses and scores (from automated scoring) were provided by Pearson. To answer RQ1, MS Excel and SPSS were used for the calculation of score averages, standard deviations and correlations among the three task types. Independent samples T-tests were also run to test significant average score differences between the two groups. For RQs 2–4, the corpus tool AntConc 3.5.8 (Anthony, 2019) was utilised to extract recurrent word sequences, and any continuous word sequences with minimum three occurrences were retrieved. Note that "type" and "token" are distinguished in this study: the former refers to different word combinations (e.g., *on the other hand* and *at the same time* counted as two types) while the latter refers to the number of occurrences (e.g., *on the other hand* occurring twice counted as two tokens).

To investigate the relationship between prompt and non-prompt word combinations in integrated and discrete writing tasks in RQ3, because of the huge amount of data, i.e. thousands of recurrent 4-word combinations with tens of thousands of instances extracted this way, only the most frequent 20 word combinations (with occurrences ranging between 10 and 40 times) from each of the task types were chosen for further analysis. This is considered comparable with Staple et al.'s (2013) study, where a cut-off point of 25 occurrences was used to retrieve lexical bundles from the writing responses of 480 test takers in TOEFL iBT but discrete and integrated tasks were not separated.

In addition to a cut-off frequency threshold, the dispersion requirement, i.e. how many texts a word sequence occurs in (typically no less than 3–5 texts), is also often adopted when determining a recurrent word combination to guard against individual idiosyncrasies (e.g., Ädel & Erman, 2012; Biber et al., 2004; Biber & Barbieri, 2007; Chen & Baker, 2016; Hyland 2008; Staples et al., 2013). A scrutiny of highly recurring clusters in the current study revealed that all of them occurred across different item responses. This was probably because test takers in PTE Academic only needed to summarise the input in one or two sentences in an integrated task (see Table 3.2), and the shorter lengths of texts plus a higher frequency threshold adopted in this study warrants that it is unlikely a word combination would occur multiple times in one single text.

In terms of RQ 4, only non-prompt word combinations were further examined because they represented test takers' writing skills in using their own words to present arguments or organize ideas rather than "borrowing" chunks of text as in prompt-based word combinations. The primary reason for excluding prompt bundles here is the concern of data sensitivity as our data comes from live test content, and any information regarding prompts needs to stay confidential to maintain the test integrity.

After removing all the prompt-based word sequences, the remaining highly frequent non-prompt clusters were then classified according to the discourse functions of referential, stance, or discourse organising, a widely used framework developed by Biber et al. (2004; Biber & Barbieri, 2007).

3.5 Results

3.5.1 Comparing Writing Task Scores

To answer RQ1, the average item scores and standard deviations (SD) from the three writing tasks and overall scores for both Chinese and non-Chinese test takers were calculated, and the results can be found in Table 3.4. The overall score is based on all four skills and all the test items from one single test sitting per test taker. Despite a lower average overall score (61.7), Chinese test takers outperformed their peers (with an overall score of 64.3) in two of the three writing tasks: WW-ESSA (7.13 vs. 6.35) and RW-SUMM (2.20 vs. 1.91). In contrast, both groups achieved

Table 3.4 Descriptive statistics of Chinese and non-Chinese groups' scores in the writing tasks of PTE Academic

		Integrated		Discrete	
Skill		Read-to-write	Listen-to-write	Writing only	
Item type		RW-SUMM Summary	LW-SUMM Summary	WW-ESSA Essay writing	Overall Score
Max. score		4	7	10	90
Chinese	Average	2.20	3.85	7.13	61.68
	SD	*0.77*	*1.64*	*1.97*	*12.59*
Non-Chinese	Average	1.91	3.87	6.35	64.33
	SD	*0.77*	*1.75*	*2.38*	*15.84*

a very similar score for the LW-SUMM task. This suggests that the Chinese students in general performed relatively well in the writing tasks in comparison with the average test takers. Based on our relevant teaching experience, Chinese students tend to be weaker in listening comprehension in comparison with reading, and it is possible that because of this reason the listen-to-write summary task (LW-SUMM) might be more challenging for Chinese test takers than the essay or read-to-write tasks when compared with non-Chinese test takers. Correlations among the three writing tasks indicate that the average scores of the three task types are all significantly correlated ($p < 0.01$), with the WW-ESSA score correlated more highly with LW-SUMM ($r = .509$), followed by the correlation between WW-ESSA and RW-SUMM ($r = .353$) and the correlation between RW-SUMM and LW-SUMM ($r = .336$).

Overall, the non-Chinese group seems to have slightly larger standard deviations (except for the RW-ESSA task type), indicating a wider spread in performances on the LW-SUMM and WW-ESSA tasks. This is unsurprising, considering that those test takers came from a much wider range of different backgrounds (cf. Table 3.1).

In Table 3.5, results from Levene's test shows that equal variances between the two test taker groups can be assumed on their RW-SUMM performance, and the two average group scores are significantly different (2.20 vs. 1.91). While on LW-SUMM and WW-ESSA writing, equal variances cannot be assumed, and the two group average scores are not significantly different for LW-SUMM (3.85 vs. 3.87), but significantly different on WW-ESSA (7.13 vs. 6.35).

3.5.2 Comparing Overall Frequencies of Recurrent Word Combinations

As mentioned earlier, recurrent word combination with minimum frequency of three times were first retrieved across the writing tasks from Chinese and non-Chinese test takers, and the type and token counts of those 4-word combinations are presented in Table 3.6. As can be seen, Chinese test takers overall used significantly more word

Table 3.5 Results of the independent samples test in comparing Chinese and non-Chinese groups' scores in the writing tasks of PTE Academic

Independent samples test									
		Levene's test for equality of variances		t-test for equality of means					
								95% confidence interval of the difference	
		F	Sig.	t	df	Sig.	Mean Difference	Lower	Upper
RW-SUMM	Equal variances assumed	0.472	0.492	6.016	998.000	0.000*	0.292	0.197	0.387
LW-SUMM	Equal variances not assumed	4.389	0.036*	−0.149	994.366	0.882	−0.016	−0.227	0.195
WW-ESSA	Equal variances not assumed	23.645	0.000*	5.645	963.920	0.000*	0.780	0.509	1.051

Note: * significant at $p < 0.05$

Table 3.6 Recurrent word combinations with no less than three occurrences

Test takers	Type/token	RW-SUMM	LW-SUMM	WW-ESSA
Chinese	Type	2374	938	2479
	Token	14,695	4555	16,689
	Ratio per 100 tokens	26.53	7.32	13.61
Non-Chinese	Type	1067	820	1528
	Token	4640	3953	9069
	Ratio per 100 tokens	12.54	6.33	7.58

Table 3.7 Total and average lengths of task responses in the Chinese and non-Chinese groups

Test takers	Lengths	RW-SUMM	LW-SUMM	WW-ESSA
Chinese	Total	55,383	62,219	122,629
	Average	55.4	62.4	245.7
Non-Chinese	Total	36,998	62,483	119,706
	Average	37.0	62.6	239.9

combinations in all of the writing tasks in terms of both types and tokens. Note that at this stage prompt and non-prompt clusters were not distinguished. Because there are various length requirements for each task (cf. Table 3.2), the occurrences are also standardised to the ratios per 100 tokens for comparison.

However, note that overall Chinese test takers appear to have produced longer responses in the RW-SUMM and WW-ESSA tasks (see Table 3.7), and it is therefore perhaps not surprising to see that more recurrent word combinations were identified in the Chinese test taker group although this is also true in terms of the ratio per 100 words (cf. Table 3.6).

3.5.3 Comparing Prompt and Non-prompt Recurrent Word Combinations

To compare the use of prompt and non-prompt recurrent word combinations, as mentioned earlier, only the most frequent word combinations (i.e. 20 for each of the tasks) were examined and divided into prompt and non-prompt categories (cf. Staples et al., 2013). The results can be found in Table 3.8. As can be seen, there is a striking difference among writing tasks in terms of the distribution of prompt and non-prompt word combinations, and the patterns of use is consistent for both the Chinese and non-Chinese test takers' groups (Figs. 3.1 and 3.2). The RW-SUMM tasks have the highest numbers of prompt-based word combinations in terms of both type and token whereas the highest numbers of non-prompt word combinations are found in the tasks of LW-SUMM or WW-ESSA.

Table 3.8 Most frequent 20 recurrent word combinations divided into prompt and non-prompt groups

Test takers	Non-prompts/ Prompts	Type/ token	RW-SUMM	LW-SUMM	WW-ESSA
Chinese	Non-prompts	Type	1	19	13
		Token	35	633	692
	Prompts	Type	19	1	7
		Token	479	18	319
Non-Chinese	Non-prompts	Type	2	19	10
		Token	27	494	392
	Prompts	Type	18	1	10
		Token	225	19	316

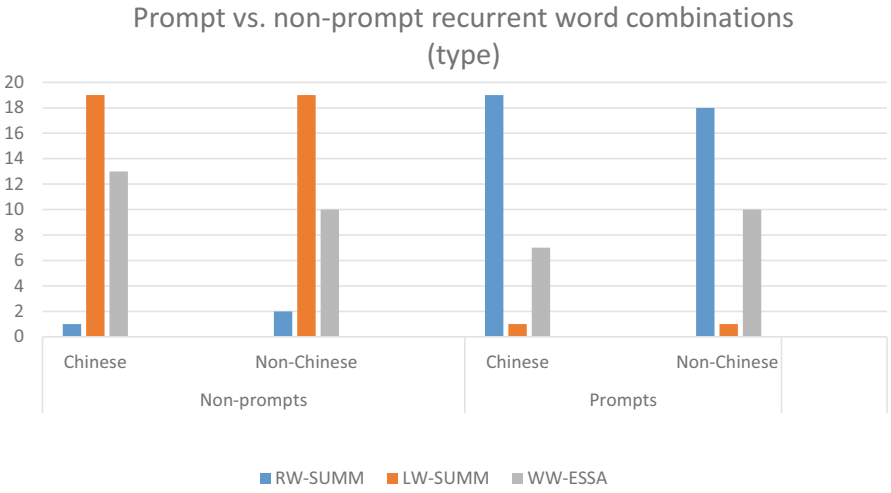

Fig. 3.1 Most frequent 20 word combinations (type) divided into prompt and non-prompt categories

3.5.4 Non-prompt Recurrent Word Combinations

In the last section, we have seen the writing task types appear to have an impact on the ratio of prompt and non-prompt word combinations, and prompt bundles account for a significantly large portion of written text in the RW-SUMM tasks, at least in the most frequent word clusters that we examined. As mentioned in Sect. 3.4.3, only non-prompt word combinations were further qualitatively analysed. The highly frequent non-prompt word combinations from Table 3.8 were categorised on the basis of three discourse functions: making reference, expressing stance or organising the discourse (cf. Biber et al., 2004; Biber & Barbieri, 2007). In a more qualitative

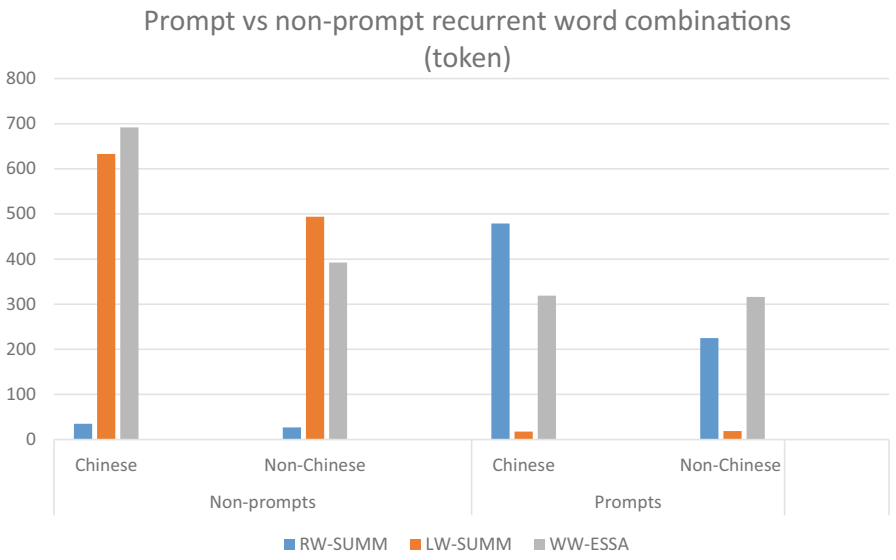

Fig. 3.2 Most frequent 20 word combinations (token) divided into prompt and non-prompt categories

examination of the data, it was then discovered that quite a few 4-word combinations were part of a longer phraseological unit. For example, the three sequences of "*this lecture mainly talks*", "*lecture mainly talks about*", and "*mainly talks about the*" actually form a longer 6-word expression of "*this lecture mainly talks about the*". To guard against inflated counts of word combination types, those clusters were therefore combined with brackets indicating the extensions such as "*this lecture mainly talks* (*about the)*" (cf. Chen & Baker, 2010). The results are presented in Table 3.9.

As can be seen, there is a striking difference between discourse functions and task types. In the RW-SUMM tasks, only "*at the same time*" and "*on the other hand*" were identified, and those are typical discourse markers found in the literature (e.g., Chen & Baker, 2010, 2016). In terms of the LW-SUMM tasks, almost all the word combinations fall into the discourse function of discourse organising, which incorporated the phrases of "*the lecture*" or "*the speaker*" to introduce main ideas of a summary. Interestingly, this type of introductory expressions were not found in the other summary writing task RW-SUMM with reading input. For the WW-ESSA tasks, this is the only task type that stance expressions occurred in the recurrent word combinations retrieved here, and Chinese test takers appeared to have used more stance bundles than their peers to present an argument or express their stance. In terms of discourse organizers in this task type, again, typical bundles that can be found in the literature such as "*on the other hand*" or "*is one of the most*" were identified.

Table 3.9 Discourse functions of non-prompt recurrent word combinations in the three writing tasks

Task	Function	Chinese	Non-Chinese
RW-SUMM	Referential	1. *at the same time* (35)[a]	1. *at the same time* (15)
	Discourse organising	–	2. *on the other hand* (12)
LW-SUMM	Referential	–	–
	Discourse organising	Referring to "*lecture*" 1. *this lecture talks about* (67) 2. *the lecture talks about (the)* (52) 3. *in this lecture the (42)* 4. *(this) lecture mainly talks about (the) (33)* 5. *the lecture is about (31)* 6. *this lecture is about (30)* 7. *lecture is talking about (28)* 8. *this lecture is mainly (19)* *Referring to "speaker"* 9. *the speaker talks about (the) (61)* 10. *the speaker mentioned that (33)* 11. *the speaker mentions that (19)* 12. *the speaker said that (17)* *Referring to both* 13. *this lecture the speaker (talks) (32)* 14. *the lecture the speaker (talks)* (23)	Referring to "*lecture*" 1. *the lecture was about (the)* (49) 2. *the lecture is about (the) (33)* 3. *in this specific lecture (17)* 4. *lecture and it comprises (that) (16)* *Referring to "speaker"* 5. *the speaker was discussing (about) (44)* 6. *according to the speaker (29)* 7. *the speaker talks about (19)* 8. *the speaker talked about (17)* 9. *speaker was talking about* (16) *Others* 10. *the talk delineates the (significance of)* (41)
WW-ESSA	Stance	1. *(while) others hold the view (that)* (74) 2. *some people believe that* (71) 3. *as far as I (am concerned)* (62) 4. *in my opinion I* (46) 5. *it has been argued (that)* (40) 6. *my point of view* (44)	1. *I would like to* (43) 2. *(above) one can conclude that* (31)
	Discourse organising	7. *on the other hand* (69) 8. *first and foremost it* (48) 9. *is the most important* (39)	3. *(is) one of the (most) (52)* 4. *on the other hand (49)* 5. *in this essay I (39)* 6. *at the outset there (are) (36)* 7. *this essay will discuss (31)*

[a]Frequency is added at the end of each word combination in brackets

3.6 Discussion and Conclusion

This study set out with a view to reviewing and comparing PTE Academic writing tasks, both integrated and independent, to compare the writing scores and use of recurrent word combinations between integrated and discrete writing tasks in the group of Chinese test takers in comparison with test takers of different L1s. Integrated writing assessment, or source-based writing, is still a fast-growing area which deserves further research. In this study, it can be seen that Chinese test takers outperformed their non-Chinese peers in the read-to-summarise and essay writing tasks despite a lower overall score in PTE Academic, and they also tended to produce longer responses in both of these two task types. This is probably somewhat unexpected for many because in the traditional second language writing research where Chinese students' writing is compared with native standards of English, usually only issues such as non-native language use are reported. This is therefore perhaps encouraging for Chinese students, particularly considering the fact that over 20% of the non-Chinese test takers in PTE Academic actually use English as their home language (see Table 3.3). For the listen-to-summarise tasks, interestingly, Chinese and non-Chinese test takers performed similarly in relation to the item scores and response lengths.

In terms of recurrent word combination use, higher numbers of occurrences in Chinese test takers' writing were identified in all of the writing tasks when compared with the non-Chinese group. This is interesting because Staples et al. (2013) reported that candidates at lower level in TOEFL iBT overall used more bundles in comparison with candidates at higher level, but in the current study, Chinese test takers overall performed significantly better than the other group in the writing tasks (except for LW-SUMM tasks) while also using more recurrent word combinations. It is possible that more content was covered in Chinese students' writing, considering that they tended to produce longer responses, and content is one of the enabling skills evaluated in PTE Academic (see Table 3.2).

It also has to be noted that there appears to be a relationship between task type and language use represented by recurrent word combinations. In the current study, prompt-based bundles dominated almost all of the most frequent word combinations in the read-to-summarise tasks, but this feature was not found in the listen-to-summarise or the essay writing tasks. This patterns also seems to hold true for both Chinese and non-Chinese groups. It is likely that for the read-to summarise tasks, where the source text can be easily accessed, test takers can just easily copy and paste chunks of text to the summary they are producing. This may be a concern for a test of academic English because discrete and integrated writing tasks contain different constructs, or labelled "task representation" by Plakans (2010). For academic writing, the ability to paraphrase as well as proper citation are both very important skills, and failing to do so can lead to plagiarism, a serious offense in Higher Education. However, this aspect of integrated writing assessment seems to have been overlooked in the design of integrated writing assessment in existing international tests of academic English, and it is perhaps an area that should be researched further.

In terms of integrated writing assessment, there appears to be mixed attitudes towards the implementation of integrated tasks in high-stake English language tests (Wei & Zheng, 2017), and one possible reason may be the lack of comprehensive L2 read-to-write or listen-to-write constructs (Plakans, 2008; Sawaki et al., 2013; Zheng & Mohammadi, 2013). However, instead of avoiding "the necessary interdependence of writing performance on reading and/or listening performance", it should be the time for language assessment researchers to take a step forward and redefine the writing construct for academic purposes, not only for the benefits of integrated tasks but also of independent tasks (Cumming, 2013: 5). We also argue that future research should look at both skills and sub-skills, for example, how to engage with the source in academic writing, which will allow us to better interpret test results. The interpretation will enable test developers to improve the tests and provide pedagogical implications to better support test takers in preparation for their tests. In terms of limitations, there are a couple of methodological issues in the current study. First of all, because our data were extracted from live test content, unfortunately we are unable to reveal much information about task prompts, but great efforts were made to ensure that the analysis still generated some meaningful results despite the lack of prompts for readers. In addition, in the qualitative investigation of recurrent word combinations, as a result of huge amounts of data, only the most frequent items were examined from each of the writing task types, and the generalizability of the results were therefore affected to some extent.

Considering the high stakes of academic English tests which are often used for visa or admission purposes, we urge that in the future more research should be conducted in the area of integrated writing assessment in relation to the traditional task of essay writing. For example, test takers' strategies, types of source input (e.g., Rukthong & Brunfaut, 2020) or students' perceptions and practices about plagiarism in source-based writing (Merkel, 2020) should be researched together with scoring and textual analysis of test taker responses to better inform test development.

Acknowledgements We are grateful for Pearson plc who kindly provided the data for us to use in this study.

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing, 24*(1), 7–36.

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes., 31*(2), 81–92.

Anthony, L. (2019). AntConc 3.5.8 (Windows) [Computer Software]. Waseda University. Available from http://www.laurenceanthony.net/

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263–286.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at*: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405.
Buck, G. (2001). *Assessing listening*. Cambridge University Press.
Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30–49.
Chen, Y. H., & Baker, P. (2016). Investigating criterial discourse features across second language development: lexical bundles in rated learner essays. CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849–880.
Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2), 155–163.
Cho, Y. (2003). Assessing writing: Are we bound by only one method? *Assessing Writing, 8*(3), 165–191.
Cumming, A. (2013). Assessing integrated writing tasks for academic purposes: Promises and perils. *Language Assessment Quarterly, 10*(1), 1–8.
Cumming, A., Kantor, R., Powers, D., Santos, T., & Taylor, C. (2000). *TOEFL 2000 writing framework: A working paper* (TOEFL monograph series no.18). Educational Testing Services.
Cumming, A., Lai, C., & Cho, H. (2016). Students' writing from sources for academic purposes: A synthesis of recent research. *Journal of English for Academic Purposes, 23*, 47–58. https://doi.org/10.1016/j.jeap.2016.06.002
Granger, S. (Ed.). (1998). *Learner English on computer*. Longman.
Granger, S., Hung, J., & Petch-Tyson, S. (Eds.). (2002). *Computer learner corpora, second language acquisition, and foreign language teaching*. John Benjamins.
Hamp-Lyons, L., & Kroll, B. (1996). Issues in ESL writing assessment: An overview. *College ESL, 6*(1), 51–72.
Hirvela, A., & Du, Q. (2013). "Why am I paraphrasing?" Undergraduate ESL writers' engagement with source-based academic writing and reading. *Journal of English for Academic Purposes, 12*, 87–98. https://doi.org/10.1016/j.jeap.2012.11.005
Hyland, K. (2008). *As can be seen*: lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4–21.
Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*, 89–114.
Leedham, M., & Cai, G. (2013). *Besides. . . on the other hand*: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials. *Journal of Second Language Writing, 22*(4), 374–389.
Li, Y., & Casanave, C. P. (2012). Two first-year students' strategies for writing from sources: Patchwriting or plagiarism? *Journal of Second Language Writing, 21*(2), 165–180. https://doi.org/10.1016/j.jslw.2012.03.002
Merkel, W. (2020). A case study of undergraduate L2 writers' concerns with source-based writing and plagiarism. *TESOL Quarterly,11(3), e00503*. https://doi.org/10.1002/tesj.503
Ministry of Education of the People's Republic of China. (2018). *2017 sees increase in number of Chinese students studying abroad and returning after overseas studies*. Retrieved from http://www.moe.gov.cn
Pearson. (2018). *Pearson test of English academic score Guide*. Available on https://pearsonpte.com/organizations/why-pte-academic/understand-our-scores/
Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*, 111–129.
Plakans, L. (2010). Independent vs. integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185–194.
Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes, 9*(2), 109–121.
Ruan, Z. (2016). Lexical bundles in Chinese undergraduate academic writing at an English medium university. *RELC, 48*(3), 327–340.

Rukthong, A., & Brunfaut, T. (2020). Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing, 37*(1), 31–53.
Sasaki, M. (1991). A comparison of two methods for detecting differential I item functioning in an ESL placement test. *Language Testing, 8*(2), 95–111.
Sawaki, Y., Quinlan, T., & Lee, Y.-W. (2013). Understanding learner strength and weaknesses: Assessing performance on an integrated writing tasks. *Language Assessment Quarterly, 10*(1), 73–95.
Shi, L. (2004). Textual borrowing in second-language writing. *Written Communication, 21*, 171–200.
Staples, S., Egber, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP development: Lexical bundles in the TOEFL iBT writing section. *English for Specific Purposes, 12*(3), 214–225.
Wei, W., & Zheng, Y. (2017). An investigation of integrative and independent listening test tasks in a computerised academic English test. *Computer Assisted Language Learning, 30*(8), 864–883.
Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27–55.
Zheng, Y., & Mohammadi, S. (2013). An investigation into the writing construct(s) measured in Pearson Test of English Academic. *Dutch Journal of Applied Linguistics, 2*(1), 108–125.

Yu-Hua Chen is interested in the use of Corpus Approaches in SLA and Language Testing. Her research has been published in international journals such as Applied Linguistics. She co-developed the Academic Collocation List (ACL) and also led on the development of the CAWSE Corpus and Transcribear (an online transcription tool).

Ying Zheng is Associate Professor from Faculty of Arts and Humanities at the University of Southampton, UK. Ying received her PhD in Cognitive Studies from Queen's University Canada, specialising in Second Language Testing and Assessment.