English Language Teaching; Vol. 15, No. 1; 2022
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Establishing an Operational Model of Rating Scale Construction for English Writing Assessment

Xuefeng Wu[1]

[1] School of Foreign Studies, Nanjing Forestry University, Nanjing, China

Correspondence: Xuefeng Wu, School of Foreign Studies, Nanjing Forestry University, Nanjing, 210037, China.

Received: November 23, 2021 Accepted: December 12, 2021 Online Published: December 14, 2021

doi: 10.5539/elt.v15n1p16 URL: https://doi.org/10.5539/elt.v15n1p16

Sponsored by Humanities and Social Science Fund of Ministry of Education of China: A CSE-based Study on the Mechanism of Assessment for Learning in College English Writing (Grant number: 19YJC740091).

Abstract

Rating scales for writing assessment are critical in that they determine directly the quality and fairness of such performance tests. However, in many EFL contexts, rating scales are made, to certain extent, based on the intuition of teachers who strongly need a feasible and scientific route to guide their construction of rating scales. This study aims to design an operational model of rating scale construction with English summary writing as an example. Altogether 325 university English teachers, 4 experts in language assessment and 60 English majors in China participated in the study. 20 textual attributes were extracted, through text analysis, from China's Standards of English Language Ability (CSE), theoretical construct of summary writing, comments on sample summary writing essays from 8 English teachers and their personal judgement. The textual attributes were then investigated through a large-scale questionnaire survey. Exploratory factor analysis and expert judgement were employed to determine rating scale dimensions. Regression analysis and expert judgement were conducted to determine the weighting distribution across all dimensions. Based on such endeavors, a tentative operational model of rating scale construction was established, which can also be applied and adapted to develop rating scales in other writing assessment.

Keywords: CSE, operational model, rating scale, summary writing, writing assessment

1. Introduction

Summary writing is a common practice particularly in higher education where students usually need to grasp and digest the main ideas and the basic structure of a huge amount of information from books or teachers' lectures (Friend, 2000). Therefore, summary writing assessment tasks are considered of great authenticity (Li, 2016) and have been included in many language assessments worldwide, for instance TOEFL, China's National Matriculation English Test (NMET), Test for English Majors in China (TEM) etc.

Despite the prevalence of summary writing in language assessments and studies concerning how students should be trained for summary writing assessment tasks (Jin, 2016) and the construct of summary writing (Li, 2014; Yu, 2007, 2008, 2009), little attrition has been paid to investigate how summary writing should be rated or how its rating scales could be developed. In classroom teaching, teachers usually rely on three options to rate students' manuscripts including directly using an existing scale, adapting scales and creating a scale from scratch (Perlman, 2003). As a result, a valid rating scale appropriate for summary writing is urgently needed to guide the rating work.

Moreover, based on the document "*Deepening the Reforms on the Educational Exams and the Enrollment Systems*" issued by the State Council of China and supported by the Ministry of Education of China, the National Education Examinations Authority (NEEA) launched a research project to develop a proficiency scale of English, China's Standards of English Language Ability (CSE). The CSE aims to define the English proficiency of learners in China and provide references and guidelines for English learning, teaching and assessment (Liu, 2015). CSE consists of several subscales concerning skills of listening, reading, speaking, writing, translating, interpreting, etc., which could offer standards to be applied in summative and formative language assessment. It

is, therefore, expected that CSE writing scales can be employed to guide the construction of rating scales for summary writing assessment tasks.

The present study, basing itself on students' real performance in summary writing and writing scales of the CSE, aims to establish an operational model to construct a rating scale using both quantitative and qualitative methods. The rating scale to be constructed could be applied after appropriate modifications in either large-scale writing tests or classroom formative assessments, thus enhancing the scoring validity and test fairness of summary writing. More importantly, based on the construction process, it is possible to establish an operational model concerning rating scale construction in writing, which can have empirical value for the construction of rating scales for English writing tasks other than summary writing.

2. Literature Review

2.1 Construct of Summary Writing

Summary writing is a task in which students read to write (Stawiarska, 2016), aiming to convey information in an efficient manner so that readers can learn the main idea and essential details through a piece much shorter than the original (Frey, Fisher, & Hernandez, 2003). The statement reveals significant elements in summary writing, e.g., source text, main idea, essential details and has won many echoed voices (Yang, 2014). Therefore, summary writing is the result of integration and interaction of reading and writing skills.

Such interactive relationship reminds us that the accomplishment of summary writing lays a cognitive burden on summarizers when they make elaborate cognitive processing of source text information (Léonard, 2001). The cognitive burden never remain on the same level across all occasions but vary greatly due to many source-text-related factors including text quality (Hidi & Anderson, 1986), writing styles (Kobayashi, 2002; Yu, 2009), text length (Kirkland & Saunders, 1991), text availability, i.e. how much time summarizers could read the source text (Kirby & Pedwell, 1991; Stein & Kirby, 1992) and text structure (Lorch, 1989). Despite the aforementioned factors, more microscopic investigations have been conducted concerning cognitive loads from summary writing tasks. Asención (2004) compared ESL and EFL learners in cognitive processing in response to summary writing tasks. Think-aloud protocols revealed that monitoring and planning occurred most frequently while organizing, selecting and connecting were much less frequent. It could be argued that although cognitive loads of summary writing generally include identifying, analyzing and synthesizing (Yang, 2014), significance of such elements is by no means the same and might vary with different levels of proficiency of the target language.

2.2 Attributes of Rating Scales

Rating scales are defined as "rules that guide scoring" (Popham, 1997: 72) or "guidelines that clearly articulate performance expectations and proficiency levels" (Gezie et al., 2012: 422) or "a tool used in the process of assessing student work" (Dawson, 2017: 349). The above definitions are offered from the perspective of the function of rating scales. More specific definitions are offered, including "by using a number of descriptive bands for a particular skill, on a scale of competence…" (Kabir, 2012: 37). Representing constructs being tested (Knoch, 2011), rating scales play a significant role in terms of the validity of subjective writing assessments (Weigle, 2002), whose rating work should be conducted with rating scales to make the subjective rater decisions as fair and objective as possible.

There exist, generally speaking, two types of scales, i.e. holistic scales and analytic scales. To begin with, holistic scales, also known as "impressionistic scales" (Kabir, 2012: 36), refer to scales which offer "a single score to a script based on the overall impression of the script" (Weigle, 2002: 112). Analytic scales refer to scales with which raters "rate on several aspects of writing or criteria rather than give a single score" (Weigle, 2002: 114). The scores for each of the aspects are then added up to obtain a total score. Although there have been disputes concerning the appropriateness of the two kinds of scales (Kabir, 2012; Li & He, 2015; Perlman, 2003), analytic scales, compared with holistic scales, are more widely adopted in large-scale and high-stakes language tests, for example, TOEFL and IELTS, to name just a few.

2.3 Construction of Rating Scales

Previous studies generally adopted two methods of constructing rating scales, i.e., theory-based and performance-driven (Jeffrey, 2015; Plakans, 2013).

Constructing theory-based rating scales refers to the process of "starting with a construct or model of the skill being tested and create a scale to reflect this theory" (Plakans, 2013: 152). Rodriguez (2008) used the Narrative Theory as sources of what is to be included as textual attributes in rating scales. The ultimate fruit of the rating scale is long and specific, consisting of 10 sections including format, punctuation & grammar, language, setting,

etc. In addition, Zhao (2013) developed an analytic scale for measuring authorial voice strength in L2 argumentative writing. The scale is one adapted from Hyland (2008)'s interactional model of voice, but with more detailed descriptors of the six hierarchical levels in the scale added by the researcher. The construction process of theory-based rating scales, however, is rather subjective, based merely on designers' judgment, thus lacking persuasiveness. Other criticism holds that such scales may lead to reliability and validity problems (Brindley, 1998; Turner & Upshur, 2002), because scales based on general linguistic or assessment theories ignore specific and dynamic contexts of assessment tasks.

Knoch (2011) asserts that the ideal option is to base the construction on the psycholinguistic development process of test takers. Such scales are called performance-driven scales, which attach great importance to observations of language performance as the foundation of descriptors (Fulcher, Davidson, & Kemp, 2011). Students' performance samples are selected and reviewed by experienced raters, teachers or other specialists, after which samples, together with these people's verbal reports, are used to generate the verbal basic content of the scale (Jeffrey, 2015). In addition, since all descriptors come from analysis of students' essays in one particular assessment task, such scales are usually not used for scoring work in other tests, thus lacking generalizability.

An important branch of performance–driven scales requires collection of performance samples from test takers via identification of key performance attributes based on text analysis of students' essays (Fulcher, Davidson, & Kemp, 2011). What follows is the determination of the number levels in the scale through discriminant analysis. The text attributes and different levels then constitute the essential structure of the rating scale (Knoch, 2007). Such scale are conventional ones in that they are composed of various dimensions and descriptors reflecting a continuum ranging from what is poor performance to what is excellent.

Recent years, however, have witnessed more endeavors in developing "empirically based, boundary-driven (EBB)" scale (Plakans, 2013). Such scales are quite unique in that they are "composed of a hierarchical set of articulated binary questions or descriptions" (Hirai & Koizumi, 2013: 400). Raters need to repeatedly make choices between "Yes" and "No" through answering binary questions about a performance and are therefore led by the scale from one step to another until finally a total score is achieved (North, 2003). EBB scales are typical performance–driven because the content and descriptors are developed based on analysis of samples of students' test performance (Hirai & Koizumi, 2013; North, 2003; Plakans, 2013; Turner & Upshur, 2002). However, EBB scales have demonstrated weakness in rating efficiency in that raters have to make several rounds of "YES/NO" choices before making decisions. Therefore, such scales are rarely employed in large-scale tests with their feasibility suspected from time to time.

In a word, theory-based scales are formed with full reliance upon related theories and might be too general and not task-specific enough. In writing assessments, however, the change of tasks naturally means the change of the assessed construct (Brindley, 1998). In this sense, theory-based rating scales are weak in their promotion value. In contrast, descriptors of performance-driven scales are derived from the specific performance of test takers, thus guaranteeing the authenticity and suitability of rating scales for various writing tasks.

Now that theories-based and performance-driven approaches for constructing rating scale are powerful in their respective domains, neither of them should be excluded from rating scale construction. In addition, despite abundant studies concerning the CSE from the following aspects: (1) elaborations of theoretical foundations and basic principles in constructions of CSE as a whole and sub-scales of CSE (Liu, 2015; Liu & Han, 2018; Liu & Peng, 2017; Zeng & Fan, 2017); (2) Investigations and analysis of structures and content of CSE (He & Chen, 2017; Kong & Wu, 2019); (3) Validation of CSE scales (Fang & Yang, 2017); Application of the CSE in foreign language pedagogy and assessment (Liu, 2017; Liu, 2019), yet little has been done to explore its role in constructing rating scales. Therefore, the present study draws on the CSE and makes integrative use of theory and performance-driven approaches to construct a rating scale for summary writing tasks, aiming to answer the following questions:

(1) What are textual attributes to be included in the rating scale?

(2) What are the dimensions and their weighting of the rating scale?

(3) What is the operational model of rating scale construction for English writing assessment?

3. Methodology

3.1 Participants

Participants were 60 juniors (9 males, 51 females) majoring in English in a university in China. They have been learning English for at least 12 years with good English proficiency. They were invited to accomplish a summary

writing task, in which they summarized a source text of about 400 words using no more than 80 words. After writing, two summaries, which represented the intermediate level of proficiency of the group, were selected from all the 60 essays.

What's more, to determine appropriate source texts for summary writing, a pilot study was conducted in which another 4 juniors of English majors (not included in the 60) were invited to write summaries for all the 4 different source texts selected by the researchers. They were also asked to provide feedback on the quality and difficulty of the source texts so as to help determine which source text was the most appropriate for the present study.

In addition to students, two groups of teachers were enrolled in this study. The first group were English teachers (n=325) from 25 universities in China for a questionnaire survey concerning textual attributes of summary writing (See Table 1). The second group consisted of 8 English teachers randomly selected from the 325 teachers to read the two intermediate level samples of summary writing. After reading, the 8 teachers separately wrote their own comments on the quality of the two samples, which later served as one source of textual attributes of summary writing.

Table 1. Detailed information of teachers surveyed

Gender	Professional title	Academic diploma	Age	Length of teaching
Male	Teaching assistant	Bachelor	25-30 Y	1-3 Y
(72/ 22.2%)	*(34/10.5%)*	*(21/6.5%)*	*(29/8.9%)*	*(26/8%)*
Female	Lecturer	Master candidate	31-35 Y	4-6 Y
(253/77.8%)	*(211/64.9%)*	*(6/1.8%)*	*(109/33.5%)*	*(28/8.6%)*
	Associate professor	Master	36-40 Y	7-9 Y
	(71/21.8%)	*(200/61.5%)*	*(117/36%)*	*(76/23.4%)*
	Professor	Doctor candidate	41-45 Y	10-15 Y
	(9/2.8%)	*(50/15.4%)*	*(41/12.6%)*	*(108/33.2%)*
		Doctor	Over 46 Y	Over 16 Y
		(48/14.8%)	*(29/8.9%)*	*(87/26.8%)*

Note. Y=years old.

3.2 Data Collection & Analysis

3.2.1 Selecting Source Text

Considerations of selecting source texts for summary writing were made from two perspectives, one being the genre, the other linguistic complexity. To begin with, narratives are relatively easier to be summarized than expository or argumentative texts due mainly to people's more familiarity with them (Meyer & Freedle, 1984). Since student participants were English majors with good English proficiency, narrative texts, therefore, were excluded. "Narrative and expository texts are common in studies of summarization in education, linguistics, and psychology. Few have employed argumentative texts" (Yu, 2009: 118). With the above considerations, the present study chose expository texts as the source text.

Readability and text length are important factors that have strong impact on the products of summary writing (Kirkland & Saunders, 1991), making it necessary to consider linguistic complexity while selecting source texts. This study chose source texts from past English for Postgraduate Admission Examination (EPAE) to ensure text quality considering EPAE's widely acknowledged high validity. Four passages were selected for further comparison. They were, respectively, Text 2 in Reading Comprehension of EPAE-2014, Text 4 in Reading Comprehension of EPAE-2006, the passage in Translation of EPAE-2007 and the passage in Translation of EPAE-2014. What followed was to decide, based on textual analysis of the passages and students' performances in the pilot summary writing of the 4 passages, which passage would be the most appropriate for further use in the study. Table 2 presents the results of analysis of linguistic complexity of the 4 passages.

Table 2. Linguistic complexity of the 4 passages

Evaluation items		Text 1	Text 2	Text 3	Text 4
Readability	Flesch Kincaid Reading Ease	56.5	57.6	37.5	46.9
	Gunning Fog Scale	12.4	13.3	15.5	18.8
Lexical complexity	Total number of words	413	403	408	401
	Type-token ratio	57.11%	57.74%	50.37%	53.09%
	Percentage of hard words	10.29%	11.79%	12.96%	15.56%

The 4 students who wrote the summaries of the four texts reacted positively to Text 2, which was considered to be more suitable for them in terms of difficulty in comprehension. As a result, Text 2 was finally chosen as the source text for the study.

3.2.2 Composing Summaries

All 60 students composed summary writing essays respectively based on the two texts. According to National Matriculation English Test (NMET) of Shanghai and Zhejiang province in China, the lengths of source text of summary writing and summary essay are about 300 words and 60 words respectively (SMEEA, 2017; NEEA, 2015), the ratio being approximately 5:1. This study also adopts the same ratio and since the source texts are approximately 400 words in length, the limit for the summary length should be no more than 80 words. The time limit for the task was one hour, after which the 60 summary essays were collected and numbered from 1 to 60 for anonymous considerations.

3.2.3 Accumulating Textual Attributes

In order to make a CSE-based rating scale, the researchers investigated all CSE writing scales, from which descriptors related to summary writing were picked out. The descriptors were then further analyzed to extract elements directly related to the core of the construct of summary writing. Additionally, a comprehensive review of the construct of summary writing provided a theoretical base and a second source for textual attributes.

Now that the rating scale is designed to be performance-driven, 8 teachers from among the total 325 were invited to read and provide commentary feedbacks on the quality of the two sample essays. The comments were made through Think-aloud protocols (TAPs), i.e. the 8 teachers read the samples and wrote down whatever thoughts they had while reading, guided by no rating scales. After giving comments, the 8 teachers, based on their teaching experience and personal judgement, brainstormed as many textual attributes as possible, which they believed should be included in the evaluation standards for summary writing.

3.2.4 Surveying with a Questionnaire

A questionnaire was designed to explore to what extent the textual attributes accumulated were appropriate for the rating scale of summary writing. It consisted of two sections, the first aiming to collect personal information about age, gender, professional title, rating experience etc. The other section displays all textual attributes to consult the 325 teachers for their attitudes towards whether the textual attributes should be included as descriptors in the rating scale. This section is presented in the form of 5 point Likert scale (1=completely disagree, 2=basically disagree, 3=uncertain, 4=basically agree, 5=completely agree).

To facilitate the survey process and ensure easier access to respondents, the questionnaire was presented using “Questionnaire Star”, a professional questionnaire online platform in China. The “Questionnaire Star” automatically collected all responses from the 325 teachers. However, the researchers examined the results and found that 14 respondents made the same choice for all items in the questionnaire. Therefore, these 14 copies were discarded and the number of copies of questionnaire for further research was 311. The Cronbach α of the questionnaire is 0.911, indicating very high internal reliability.

In order to determine and define the dimensions of the rating scale, exploratory factor analysis (EFA) was applied with SPSS 24.0 to categorize textual attributes into various dimensions. Meanwhile, qualitative expert judgement also played a significant role as supplement. The experts are 4 experienced university English teachers, each with a doctor’s degree and an academic title of full professor. Their research focuses are language assessment and second language acquisition.

3.2.5 Regression Analysis of Questionnaire Data

The purpose was to determine the weightings for various dimensions of the rating scale. Based on EFA and expert judgement, a preliminary rating scale was constructed with 5 specific dimensions, each with 5 levels on

the continuum of "Excellent—Good—Ordinary—Poor—Fail". For convenience considerations, the full mark of the summary writing task was 100 points and each dimension shared the same weighting, i.e. 20 points. Besides, the 20 points for each dimension was divided evenly across the 5 different levels. Table 3 presents an outline of the preliminary rating scale constructed.

Table 3. Outline of the preliminary rating scale

Levels	Points	D1	D2	D3	D4	D5
Level 1: Excellent	17-20 points	X	X	X	X	X
Level 2: Good	13-16 points	X	X	X	X	X
Level 3: Ordinary	9-12 points	X	X	X	X	X
Level 4: Poor	5-8 points	X	X	X	X	X
Level 5: Fail	0-4 points	X	X	X	X	X

Note. D=Dimension; X=descriptor.

To prepare statistics for regression analysis, the two researchers rated the 60 essays separately based on the preliminary rating scale. The 1st step was to determine at which level the essay was located for a particular dimension, i.e. a level score. They also needed to pick up an exact point from the range at the determined level, i.e. a specific score. The 2nd step was to repeat what had been done in the first step four times, one for each of the other 4 dimensions. The researchers initially rated 6 essays out of the 60, after which their results were compared to ensure inter-rater reliability. Correlation analysis revealed that the rating of the two raters was reliable for further studies ($r=0.92$, $p<0.05$). The two raters then separately rated the remaining 54 essays following the aforementioned 2 steps above. The average scores of the rating results of the two raters were used as the final scores for all the 60 summaries.

As a result, for all the 5 dimensions of an essay, there were respectively 5 level scores and 5 specific scores. These scores were then added up to obtain the final total scores for each of the 60 essays. Regression analysis was conducted with the 5 level-scores as independent variables and the total scores as the dependent variable. Standardized regression coefficients (β) and Unstandardized coefficients (B) were employed as indicators of different degrees of significance of the 5 variables, thus helping to determine the ratio of distribution of weightings among the 5 dimensions in the rating scale. The 4 experts were invited again to provide comments and feedbacks as a supplement for final decisions of weighting distribution.

4. Results

4.1 Accumulating Textual Attributes

Accumulation of textual attributes as descriptors of the rating scale was made with the CSE, construct of summary writing, teachers' comments through TAPs and personal judgement as major sources. From each source, we collected typical and representative attributes and then merged identical attributes into various independent attributes presented in Table 4. In order to adapt attributes for further large-scale questionnaire survey, attributes from across various sources were then merged to avoid repetition. For instance, B1, C3 and D9 focus on using paraphrased rather than copied language of the source text; A1, B2, C1, D1 all stress the content coverage of source texts, to name just a few. In addition, some attributes that contain over two aspects were split into several independent descriptors. For example, C5 can be further divided into such elements as "avoiding grammatical errors", "using flexible subordinate clauses" and "using appropriate words accurately", etc., which can then be revised and adapted to construct more specific descriptors.

After processing work of merging and splitting, 29 textual attributes were extracted, based on researchers inductive judgement (Table 5). As a result, such decisions were subjective and needed to be further evaluated. The 29 attributes were then sent to the 4 experts in language assessments for consultation. They suggested items Q39, Q17, Q19, Q44 be deleted. To be specific, Q39 was too easy for students; Q17 and Q19 overlapped with Q32; and Q44 overlapped with Q30.

Table 4. Selected textual attributes and commentary feedbacks (Excerpts)

CSE	Construct	TAP comments	Personal judgement
A1-Can extract important points or information from literature or references. **A2-**Can list key words or points to summarize articles. **A3-**Can use a topic sentence to highlight the main idea of a paragraph. **A4-**Can give a generally accurate and complete summary of the plot points of movies or dramas. **A5-**Can outline the characters, themes, or ideas of a text... **A6-**Can summarize factual and imaginative texts, highlighting the most important points.	**B1-**Restatement of the original text into writers' own words in showing only the author's main ideas (Doyle, 2012) **B2-**…convey correct information efficiently so that readers learn the main idea and essential details through a much shorter piece (Nancy et al., 2003) **B3-**A shortened form of a text giving main points from the original text and isolated from trivial details. (Benzer, et al., 2016)	**C1-**He gave a complete summary of the major content of the source text with clear structures. **C2-**Despite a few errors in vocabulary, overall it's satisfactory. **C3-**The writer could use his own words to summarize. **C4-**The summary is written with cohesion and connection between sentences. **C5-**The writer can use grammar and vocabulary correctly and also diversified & flexible sentence patterns. **C6-**The language isn't succinct. The writer has talked too much. **C7-**The writer put too much emphasis on the first part of the source text. **C8-**Punctuation is used properly.	**D1-**Accurate & reasonable summary of the main points, main idea and content **D2-**Excluding additional & irrelevant information interpretation or of the source text **D3-**Contains important facts & details **D4-**Accurate use of words **D5-**Vocabulary with complexity **D6-**Correct grammar & precise diction **D7-**Flexible grammatical usage such as clauses & inversion **D8-**No copy of the source text **D9-**Using their own highly generated language **D10-**Use conjunctions to outline text logic and a clear structure

As a result, 25 textual attributes remained, which were formally adopted for the questionnaire survey. Table 6 presents the descriptive statistics of the results. The means for most of the attributes are over 4 except Q34 (Means=3.93), Q35 (Mean=3.66), Q36 (Mean=3.60), Q41 (Mean=3.76) and Q43 (Mean=3.60), indicating that the 5 attributes failed to win large-scale support among the 325 respondents. For confirmation, consultation was made with the 4 experts. They supported deleting the 5 attributes including writing tone, idiom use, writing style, rhetoric devices and complicated grammatical structure, which had little to do with the construct of summary writing. As a result, the 5 attributes were removed, leaving 20 attributes that constituted a bank of textual attributes for the main body of the rating scale.

Table 5. Results of establishing the bank of attributes of summary writing

ID	Textual attributes	ID	Textual attributes
Q17	The theme of summary writing is the same as that of source texts	Q31	Write with smooth diction and fluent language
Q18	Write with a clear structure and distinct layers	Q32	Cover all the major points in source texts
Q19	Summarize source texts content briefly and comprehensively	Q33	No change of logical relations or opinions in source texts
Q20	Use conjunctions to make more natural paragraph transitions	Q34	Write with a tone similar to that in the source text
Q21	Use conjunctions to make more natural sentence transitions	Q35	Use idioms or slangs correctly when necessary
Q22	Write with succinct language without too lengthy and wordy expressions	Q36	Keep the style of the source texts in summary writing
Q23	Use subordinate clauses with flexibility to avoid sentences with loose structures	Q37	Use diversified vocabulary to avoid repetition and tediousness
Q24	Use various sentence patterns with frequent changes	Q38	Use some advanced vocabulary accurately and properly
Q25	Avoid using too long or too short sentences	Q39	Use capitalized letter accurately and properly
Q26	No inclusion of content unavailable in source texts	Q40	Neat handwriting and clean sheet
Q27	Accurate and appropriate use of punctuation	Q41	Use rhetoric devices properly to enhance expressive effects
Q28	Describe content of source texts with clear priorities	Q42	Write with clear diction without causing misunderstandings
Q29	Use words correctly and properly without spelling mistakes	Q43	Properly use complicated grammatical structures (e.g. subjunctive mood)
Q30	No appearance of grammatical errors	Q44	Use all kinds of tenses and voices accurately and properly
		Q45	Use standardized language without expressions too oral or internet words

Note. Q=questions in the questionnaire; Q1-Q16=questions about personal information, not included in the table.

Table 6. Descriptive statistics of the questionnaire survey

Attributes	Min	Max	Mean	S.D.
Q18	1	5	4.57	0.692
Q20	1	5	4.47	0.717
Q21	1	5	4.46	0.708
Q22	1	5	4.43	0.733
Q23	2	5	4.27	0.764
Q24	2	5	4.27	0.790
Q25	1	5	4.11	0.785
Q26	1	5	4.39	0.779
Q27	2	5	4.47	0.621
Q28	2	5	4.51	0.736
Q29	2	5	4.50	0.722
Q30	2	5	4.45	0.716
Q31	2	5	4.58	0.677
Q32	2	5	4.27	0.727
Q33	2	5	4.44	0.688
Q34	**1**	**5**	**3.93**	**1.106**
Q35	**1**	**5**	**3.66**	**1.030**
Q36	**1**	**5**	**3.60**	**1.037**
Q37	1	5	4.29	0.705
Q38	1	5	4.08	0.867
Q40	1	5	4.55	0.645
Q41	**2**	**5**	**3.76**	**1.081**
Q42	2	5	4.56	0.587
Q43	**1**	**5**	**3.60**	**1.029**
Q45	2	5	4.37	0.664

4.2 Categorizing Rating Dimensions

The 20 attributes went through exploratory factor analysis (KMO=.857; $p<.05$) to help determine the classification of textual attributes and make tentative decisions of the rating scale dimensions.

With the results of exploratory factor analysis (Table 7), the attributes could therefore be safely categorized into 5 components as the dimensions of the potential rating scale. The categorization results are displayed in Table 8 with each potential dimension named by the researchers according to the shared features of attributes.

Table 7. Rotated Component Matrix

	Component				
	1	2	3	4	5
Q24	0.836				
Q23	0.816				
Q25	0.790				
Q38	0.769				
Q37	0.763				
Q20		0.902			
Q21		0.889			
Q18		0.848			
Q22		0.804			
Q33			0.853		
Q26			0.814		
Q32			0.807		
Q28			0.776		
Q30				0.926	
Q31				0.894	
Q29				0.894	
Q40					0.846
Q27					0.779
Q45					0.772
Q42					0.459

Based on the 20 textual attributes and the dimensions, a preliminary rating scale of summary writing was tentatively established (See Appendix 1). Each dimension consists of 5 levels discriminated by specific indicators including “no”, “less”, “comparatively”, “completely”, etc., which demonstrate to what extent summarizers accomplish the task.

Table 8. Suggested dimensions of the rating scale for summary writing

Dimension	Attributes	Reasons for the naming given to dimensions
Linguistic Complexity (LC)	Q23, 24, 25, 37, 38	The 5 textual attributes center on the use of advanced & diversified words, and diversified sentence structures as well as subordinate clauses to avoid repetition.
Coherence & cohesion (CC)	Q18, 20, 21, 22	The 4 textual attributes center on making connections between sentences, paragraphs to present summaries with clear structures.
Fidelity to source texts (FC)	Q26, 28, 32, 33	The 4 textual attributes stress the necessity of excluding new content in summary and complete coverage of points in source texts
Linguistic accuracy (LA)	Q29, 30, 31	The 3 attributes stress the significance of using language in a correct way in terms of vocabulary, grammar and language fluency.
Mechanism (MC)	Q27, 40, 42, 45	The 4 attributes stress the need to write with normalized punctuation, clear and neat handwriting & normalized use of language in authentic use.

4.3 Determining Weighting Distribution Across Dimensions

Multiple linear regression analysis supplemented by expert judgement was conducted to help determine weighting distribution across dimensions. Standardized coefficient β is used as an indicator of the influence of independent variables, i.e. the five dimensions, on the dependent variable, i.e. the “Writing scores” (Table 9). In

the descending order of β, the five dimensions are listed as follows: LA (β=.256, t=5.952, p<.05), CC (β=.222, t=6.468, p<.05), FS (β=.212, t=6.703, p<.05), MC (β=.204, t=6.419, p<.05) and LC (β=.195, t=4.996, p<.05). Therefore, LA could be tentatively allotted the highest weighting, i.e. approximately 25%, while FS, CC and MC can be tentatively allotted the same share, i.e. respectively 20% due to the very little difference in β value. The lowest weighting goes to LC, i.e. 15%.

Table 9. Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	S. E	β		
(Constant)	-11.032	1.190		-9.268	.000
LA	4.891	.822	.256	5.952	.000
LC	3.846	.770	.195	4.996	.000
FS	3.839	.573	.212	6.703	.000
CC	4.914	.760	.222	6.468	.000
NN	3.429	.534	.204	6.419	.000

For further confirmation, the preliminary scheme of weighting distribution was presented to the 4 experts. They, however, only partially agreed with the proposed distribution of the potential scale and strongly recommended that the weighting distribution be appropriately rearranged. Expert A made the following comments.

MC is simply *about punctuations, neatness of presence, etc. and is not closely connected with the summary work itself. So, I can't accept that MC shares with FS and CC the same weighting in the rating scale.*

Expert A's attitude was echoed by Expert C who held that "apparently, MC is far less important than the other four". Consequently, a decision was made that the weighting of MC should be decreased, which gained positive feedback from the other two experts.

Moreover, all experts agreed with the current weighting of LA, holding that a summary in EFL contexts should never be considered of high quality with the presence of many grammatical or lexical errors. This highlights the status of LA allotted with 25% of the total weighting as the highest among all dimensions. In the same sense, LC currently takes 15% of the weighting, which seems inadequate. This can be further supported by opinions from Expert C.

This rating scale is constructed targeting college students, rather than primary or middle school students. They have gained good English proficiency. The language they use in summaries should not only be accurate *but also complex, reflected by the use of advanced vocabulary, diversified sentence patterns, etc.*

Expert D held that "summary writing is in essence a member of the family of English writing tasks". He made the following comments concerning the issue.

To accomplish summary writing tasks, summarizers not only need to put together all major points of information from source texts, but also make smooth connections among all points. This is because it is a passage rather *than several isolated sentences summarizers write.*

Expert D elevated the status of CC and suggested increasing its weighting, which can be realized, according to Expert B, by "appropriately adding the weighting deducted from MC."

All the four experts unanimously expressed the same attitude towards FS, holding that this is a dimension typical of summary writing and might be absent in rating scales for other writing tasks in that "summary writing is a highly condensed version of its source text (Expert A)". Important as this dimension is, the 4 experts believed that the current weighting for FS—20%, is appropriate.

However, experts' suggestions only offered a general scheme of whether the weighting of a certain dimension should be increased, decreased or kept the same. There were no specific proposals concerning the extent to which dimension weightings are to be changed.

The preliminary scheme of weighting distribution based on the β coefficients could then be further revised via the B coefficients, which can also be used to compare significance of different variables with the same units as the precondition (Nardi, 2009). It is apparent that all variables in the study share the same unit of "points", indicating that the use of unstandardized B coefficients, including respectively B=4.891 (LA), B=3.846 (LC),

B=3.839 (FS), B=4.914 (CC) and B=3.429 (MC), is acceptable in helping to decide the weighting distribution. Based on the B coefficients, it can be tentatively concluded that LA and CC receive equal weighting, LC and FS receive lower but also equal weighting and MC seems to be the least important due to its comparatively much lower B coefficient.

With the above considerations, Table 10 presents the revised scheme of weighting distribution with symbols "↑", "↓" and "—" respectively meaning "adding", "decreasing" and "no change". For further confirmation and comments, the new scheme was sent to the 4 experts, all of whom offered positive feedback. Expert C, however, was a bit more cautious, suggesting that "although the new scheme seems more acceptable than the previous one, it is still a tentative decision and needs further validation to judge its appropriateness". In practical use, while scoring each dimension, raters are expected to initially locate summarizers' performance into a particular level of that dimension and then decide which specific score to be given within the score range of each of the levels.

Table 10. Comparison between the old and new scheme

	LA	LC	FS	CC	MC	Total
Weighting-original	25%	15%	20%	20%	20%	100%
Adjustment method	—	↑	—	↑	↓	—
Weighting-revised	25%	20%	20%	25%	10%	100%

5. Discussion & Conclusion

We now discuss the findings by returning to the research questions of the present study.

(1) What are textual attributes to be included in the rating scale?

Rating scale, as an instrument of scoring performance-based assessments like English writing, can serve as the representation of the construct assessed through various textual attributes (Turner & Upshur, 2002). The process of the accumulation of textual attributes of summary writing in this study could be synthesized into a "3D Principle", respectively meaning "Diversified coverage" of the scope of textual attributes, "Diversified sources" of textual attributes and "Diversified extraction approaches" in accumulation. As for diversified coverage, results show that the attributes collected cover a variety of aspects of the construct of summary writing. This is in line with the argument of Yu (2013) that attributes or descriptors collected for rating scales need to be concrete and diversified in content and format. The bank of textual attributes of summary writing cover a wide range, for instance, vocabulary use, sentence structure organization, cohesion, summary-source text relationship etc.

The diversified sources of attributes pertain to endeavors of collection from sources like CSE, construct of summary writing, teachers' commentary feedback and personal judgement. Apparently, CSE and the construct provide some macroscopic attributes concerning the overall summary writing ability, represented for example, by key words "main idea", "most important points", "complete summary", "essential details" etc. In contrast, TAP comments and personal judgement offer more microscopic perspectives reflected by more specific definition of summary writing ability, for example, "using his own words (C3)", "vocabulary with complexity (D5)", etc. Therefore, macroscopic and microscopic perspectives can be supplements for each other to guarantee appropriate coverage of attributes. Extracting textual attributes from teachers' TAP comments on sample summaries, as the second source, has echoes in many studies (Chen & Liu, 2016; Jeffrey, 2015; Turner & Upshur, 1996) because individual teachers view students' summary writing from different perspectives, some of which might overlap but others of which could be put together, thus broadening the coverage of attributes. Jeffrey (2015) proposed the value of teachers' verbal comments on students' performance rather than the written ones in the present study. The difference, however, reminds us that both written and verbal feedbacks or comments can be used for exploration of attributes of writing tasks, which could expand the scope of extraction to avoid any possible omission. The final source of attributes is teachers' personal judgement, which is in line with suggestions by Perlman (2003) concerning rating scales development. Apparently this source resembles, to some extent, teachers' TAP comments, highlighting the important role played by teachers in developing rating scales. Teachers' voices are of more authenticity because textual attributes from such sources are based on students' actual writing performance.

Finally, diversified approaches for extracting textual attributes include analyzing quantitative questionnaire results and the qualitative researchers' extraction work and expert judgment. The researchers, by comparing, merging and splitting attributes, made preliminary processing work, leading to a tentative version of the bank of textual attributes. To justify and ensure the appropriateness of decisions concerning whether to remove or keep certain attributes, it is of great necessity to enroll expert judgement as a qualitative approach. Similar opinions

could be found in Plakans (2013), who argued that data and statistical analysis can't be perfect without analysis of language experts including teachers and researchers.

(2) What are the rating dimensions and their weighting of the rating scale?

The 5 dimensions of the rating scale are respectively LA, LC, CC, FS and MC, which are in consistency with the construct of summary writing. Summary writing involves the integration of reading and writing skills (Stawiarska, 2016). The first step to write a summary is reading for an accurate and comprehensive understanding of source texts (van Dijk & Kinstch, 1983), which does not simply lie in the anticipation of equating the content of summary writing with that of the source text, but more demanding and challenging requirements concerning the complicated mental processing of source texts. Summarizers need to extract from source texts specific points of information divided into major and secondary ones, the former of which should all be covered (Q32) while the latter of which is to be abandoned or at least integrated and deducted (Q28). Such attributes were put into the FS dimension in that they all concern the matching relationship between summaries and source texts. This is in line with previous assertions about cognitive loads on summarizers, including selecting essential ideas across original paragraphs (Brown, Day, & Jones, 1983), selecting the information to be represented in a summary (Johnson, 1983) and working out text thesis and major ideas (Li, 2016). Moreover, summary writing in this study was conducted in EFL contexts, which stresses the involvement of language use as well as logical relationship among sentences or paragraphs in rating scales. Therefore, traditional dimensions like LC, CC, LA, & MC are included in the rating scale and present a more comprehensive and complete coverage of the construct of summary writing. Together with FS, the 5 dimensions further confirmed that summary writing is a discourse in its own right, a discourse that requires evaluation criteria different from any other integrated writing tasks (Yu, 2013).

The present study, based on β coefficients in regression analysis, made a preliminary arrangement of weighting distribution (LA=25%, LC=15%, FS=CC=MC=20%). It was then revised through decreasing the weighting of MC to 10% and increasing that of LC to 25% based on B coefficients and expert judgement. Quantitative regression analysis is advocated as an effective and scientific method for "assigning appropriate weights to component parts of a rating scale (Henning, 2001) since it can indicate the different significance of each dimension vividly through statistics. However, quantitative statistics themselves are insufficient to guarantee the persuasiveness of the decisions. Therefore, this study turned to, after the regression analysis, expert judgement for confirmation. Such an approach of double-check can ensure that there is no obvious deficiency.

In addition, more attention should also be paid to general guiding principles that direct our endeavors. The weighting in this scale is not evenly distributed but with emphasis on certain dimensions such as LA (25%) and LC (25%) and with the least share of weighting given to MC (10%). There have been many attempts different from the scheme in this study. Sasaki & Hirose (1999) constructed a rating scale for Japanese university students' expository writing, arguing that all dimensions should receive equal share of weighting because "the explanatory power of each criterion can vary from composition to composition". Such equal weighting distribution roots itself in early discoveries represented by Hamp-Lyons (1991). However, equal distribution has not gained many supportive voices in the academic world, where most people hold that proper rather than arbitrarily equal distribution of weighting is required in constructing rating scales (Zou, 2011).

The issue of weighting distribution is closely connected with the construct of the test task and, therefore, neither equal nor arbitrary distribution is encouraged. The scheme of weighting distribution in the present study is not inflexible at all. Instead, the current scheme needs to be adapted according to factors like test takers' English proficiency, specific types or tests, etc. One typical example is NMET Shanghai, whose test takers are mostly graduates of senior high schools. They are quite different from summarizers in the present study in terms of language proficiency and such difference has been taken into consideration for weighting distribution in the scales. The rating scale in NMET Shanghai, with a full mark of 15 points, consists simply of two dimensions, namely content and language, whose weightings respectively are 10 points and 5 points (SMEEA, 2017). However, summarizers in the present study are juniors majoring in English. As a result, the scheme of weighting distribution for NMET version is by no means suitable for the present study, where more refined divisions should be made with diversified weighting distributions. As an EFL writing task, more emphasis laid upon LA and LC is justified because language proficiency takes priority in EFL contexts, which can be confirmed with Expert A's following words: "the current scheme of weighting distribution conforms to general cognition and has won wide recognition in the field of language teaching, learning or testing".

(3) What is the operational model of rating scale construction for English writing assessment?

The approach for rating scale construction in this study is a mixed one, taking into account the CSE, construct of summary writing and students' actual performance, because only one single approach for rating scale development is by no mean sufficient (Knoch, 2011). All approaches (CSE-based, construct-based & performance-based) have their own strengths and weakness and should be put into synthesized use rather than separated from each other. Figure 1 presents an operational model for constructing the rating scale, which can also offer suggestions for the development of other rating scales in writing assessment.

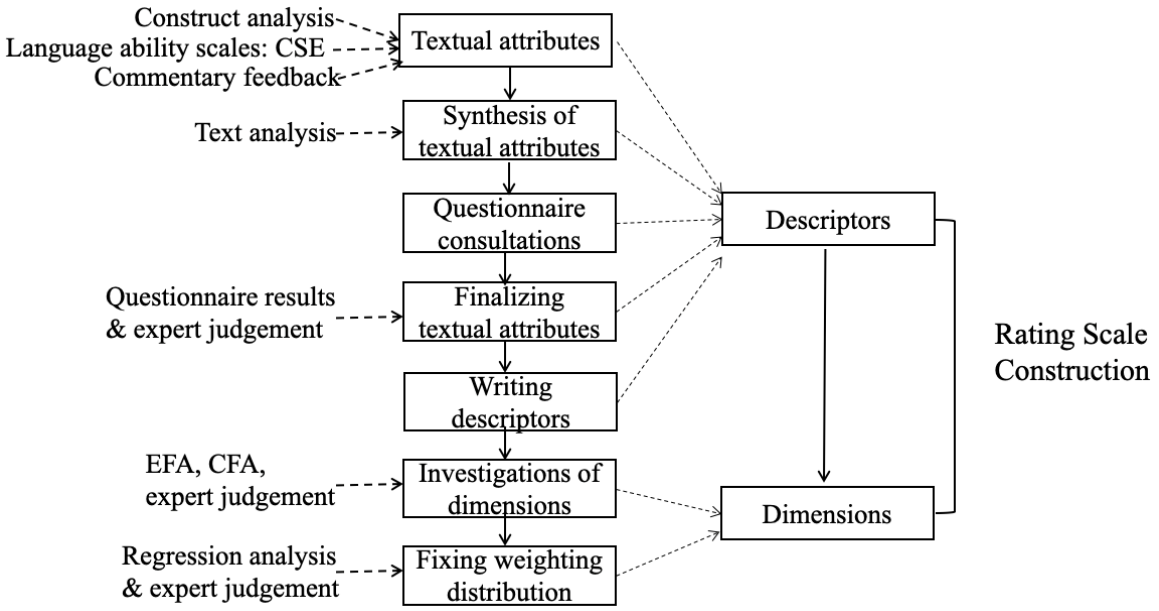

Figure 1. An operational model of constructing rating scales for English Writing Assessment

Performance-based approach has long been advocated to construct rating scales (Fulcher, Davidson, & Kemp, 2011; Plakans, 2013). Unlike prototypical performance data-based approaches which collect performance samples and identify key features through textual analysis, this study relies on teachers' commentary feedback on students' summaries together with all potential attributes based on their personal judgement. The 8 raters chosen to give commentary feedbacks worked separately without referring to any rating scales so as to elicit potential textual attributes of summary writing. This process could also be regarded as a "rater workshop" (Shaw & Weir, 2007), a method for construction of rating scales adopted also by Educational Testing Service (Cumming, Kantor, & Powers, 2001) and is in line with the study of Jeffery (2015), in which teachers' summative comments on their students' performance were recorded to be used for extractions of potential descriptors of rating scales. The accumulation of textual attributes from teachers' commentary feedbacks should be flexible and dynamic, in order to enlarge the coverage and enrich the diversity of attributes.

Rating scales in writing assessment have long been developed, as a tradition, based on intuitive judgement of experts, lacking the support of test takers' authentic data. As a remedy, the rating scale constructed in the present study can be either directly applied by teachers and raters, or be modified to suit various assessment contexts. Raters and teachers should be encouraged to strengthen their awareness that rating scales for writing assessment need to be developed based not on subjective and intuitive judgment, but on the integrative use of resources available, including investigations of the construct of the assessment task, teachers' written or verbal commentary feedbacks (Jeffrey, 2015) on students' authentic performances of the assessment task. This is to guarantee that rating scales are more targeted for specific users since they are constructed in a way deeply rooted in real situations. Moreover, since the construction of rating scales for writing assessment is an iterative and ongoing process (Hirai & Koizumi, 2013; Weigle, 2002), full of "missteps and multiple revisions" (Plakans, 2013:161), it should be kept in mind that rating scales, after construction, have to undergo continuous rounds of validation and investigations for potential improvement.

6. Limitations and Future Directions

The first limitation is that the study covered only a limited sample of data which involved only 350 teachers and 60 students. Therefore, perceptions and attitudes of a limited number of people undermine the plausibility of the

rating scale constructed. A larger and more varied sample would be desirable for greater generalizability. Furthermore, based on the CSE, the construct, TAP-based commentary feedback and personal judgment, generalization of textual attributes was made by researchers themselves. This process is completely intuitive and there might be some attributes neglected or misinterpreted. To avoid this potential problem, more experts could be invited to jointly work on text comparison and analysis.

As for future directions, the present study constructed a rating scale of summary writing based on data collected among juniors of English majors whose English proficiency is relatively quite high, indicating that the rating scale may not be suitable for summary writing tasks of other proficiency levels. Therefore, the study could be duplicated targeting students of various proficiency levels including, for instance, non-English majors, or middle school students. Further comparisons can also be made among rating scales of different proficiency levels for investigations of similarities and differences.

In addition, the present study focuses only on the construction of the rating scale, which requires comprehensive validation, thus offering suggestions for its improvement. Moreover, when involved in rating their or peers' performance, students, with rating scales at hand, might greatly improve their writing performance since they have opportunities to explore their own strengths and weakness based on full comprehension of the criteria in rating scales (Andrade, 2005; Becker, 2016). As a result, a rating scale for students could also be constructed with full consideration of students' language proficiency and cognitive characteristics.

Declaration of Interest

None.

References

Andrade, H. G. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27-30. https://doi.org/10.3200/CTCH.53.1.27-31

Asención, D. Y. (2004). *Validation of reading-to-write assessment tasks performed by second language learners* (Unpublished doctoral dissertation). Northern Arizona University, Arizona, USA.

Bacha, N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? *System, 29*(3), 371-383. https://doi.org/10.1016/S0346-251X(01)00025-2

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: a mixed-method study. *Assessing Writing, 12*(2), 86-107. https://doi.org/10.1016/j.asw.2007.07.001

Becker, A. (2016). Student-generated scoring rubrics: examining their formative value for improving ESL students' writing performance. *Assessing Writing, 29*, 15-24. https://doi.org/10.1016/j.asw.2016.05.002

Benzer, A., Sefer, A., Ören, Z., & Konuk, S. (2016). A student-focused study: strategy of text summary writing and assessment rubric. *Education and Sciences, 41*(186), 163-183. https://doi.org/10.15390/EB.2016.4603

Brindley, G. (1998). Describing language development? Rating scale language acquisition. In L. F. Bachman & A. D. Cohen (Eds.), *Interface between SLA and language testing research* (pp. 112-114). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139524711.007

Brown, A. L., Day, J. D., & Jones, R. S. (1983). The development of plans for summarizing texts. *Child Development, 54*(4), 968-979. https://doi.org/10.2307/1129901

Chen, Y., & Liu, J. (2016). Constructing a scale to assess L2 written speech act performance: WDTC and e-mail tasks. *Language Assessment Quarterly, 13*(3), 231-250. https://doi.org/10.1080/15434303.2016.1213844

Cumming, A., Kantor, R., & Powers, D. E. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (unpublished manuscript). Princeton: Educational Testing Service.

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education, 42*(3), 347-360. https://doi.org/10.1080/02602938.2015.1111294

Fang, X. J., & Yang, H. Z. (2017). Validity and Validation of Language Proficiency Scales. *Journal of Foreign Languages, 40*(4), 2-14.

Frey, N., Fisher, D., & Hernandez, T. (2003). "What's the gist?" Summary writing for struggling adolescent writers. *Voices from the Middle, 11*(2), 43-49.

Friend, R. (2000). Teaching summarization as a content area reading strategy. *Journal of Adolescent and Adult Literary, 44*(4), 320-329.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: an advanced resource book*. New York, NY: Routledge. https://doi.org/10.4324/9780203449066

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: performance decision trees. *Language Testing, 28*(1), 5-29. https://doi.org/10.1177/0265532209359514

Gezie, A., Khaja, K., Chang, V. N., Adamek, M. E., & Johnsen, M. B. (2012). Rubrics as a tool for learning and assessment: What do Baccalaureate students think? *Journal of Teaching in Social Work, 32*(4), 421-437. https://doi.org/10.1080/08841233.2012.705240

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In Hamp-Lyons, L. (Ed.), *Assessing second language writing in academic contexts* (pp. 241-276). Norwood: Ablex.

He, L., & Chen, D. (2017). Developing common listening ability scales for Chinese learners of English. *Language Testing in Asia, 7*(4), 1-12. https://doi.org/10.1186/s40468-017-0033-4

Henning, G. (2001). *A Guide to Language Testing: Development, Evaluation and Research*. Beijing: Foreign Language Teaching & Research Press.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, instruction. *Review of Educational Research, 56*(4), 473-493. https://doi.org/10.3102/00346543056004473

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a story retelling speaking test. *Language Assessment Quarterly, 10*(4), 398-422. https://doi.org/10.1080/15434303.2013.824973

Hyland, K. (2008). Disciplinary voices: Interactions in research writing. *English Text Construction, 1*(1), 5-22. https://doi.org/10.1075/etc.1.1.03hyl

Jeffrey, R. (2015). Using feedback comments to develop a rating scale for a written coursework assessment. *Journal of English for Academic Purposes, 18*, 51-63. https://doi.org/10.1016/j.jeap.2015.03.002

Jin, Y. (2016). A study of summary writing in middle schools: Problems and solutions. *Foreign Language Testing and Teaching*, (4), 38-42.

Johnson, N. S. (1983). What do you do if you can't tell the whole story? The development of summarization skills. In Keith E. Nelson (Ed.), *Children's Language* (pp. 315-383). NJ: Hillsdale.

Kabir, M. H. (2012). Necessity of initiating rating scale for more reliable assessment of writing skill at HSC level: a case study. *Liuc Studies*, (6), 35-52. https://doi.org/10.3329/iiucs.v6i0.12247

Kirby, J. R., & Pedwell, D. (1991). Students' approaches to summarization. *Educational Psychology, 11*(4), 297-307. https://doi.org/10.1080/0144341910110306

Kirkland, M. R., & Saunders, M. A. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly, 25*(1), 105-121. https://doi.org/10.2307/3587030

Knoch, U. (2007). *Diagnostic writing assessment: the development and validation of a rating scale* (Unpublished doctoral dissertation). The University of Auckland, New Zealand.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: what should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81-96. https://doi.org/10.1016/j.asw.2011.02.003

Kobayashi, M. (2002). Method effects on reading comprehension test performance: Text organization and response format. *Language Testing, 19*(2), 193-220. https://doi.org/10.1191/0265532202lt227oa

Kong, J. F., & Wu, X. F. (2019). Investigating the structure of the writing scales of China's Standards of English Language Ability. *Foreign Language Testing & Teaching, 1*, 16-26.

Léonard, P. R. (2001). Summary Writing: A Multi-Grade Study of French Immersion and Francophone Secondary Students. *Language, Culture & Curriculum, 14*(2), 171-186. https://doi.org/10.1080/07908310108666620

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. *Language Assessment Quarterly, 12*(2), 178-212. https://doi.org/10.1080/15434303.2015.1011738

Li, J. L. (2014). *Validity investigations of summary writing* (Unpublished doctoral dissertation). Guangdong University of Foreign Studies, China.

Li, J. L. (2016). Modeling the process of summary writing of Chinese learners of English as a foreign language. *Irish Educational Studies, 35*(1), 1-28. https://doi.org/10.1080/03323315.2016.1146154

Liu, J. D. (2015). Some thoughts on developing China Common Framework for English language proficiency. *China Examinations*, (2), 7-15.

Liu, J. D. (2017). China's Standards of English and its application in English Learning. *Foreign Languages in China, 14*(6), 4-11.

Liu, J. D. (2019). China's Standards of English Language Ability and English teaching. *Foreign Language World*, (3), 7-14.

Liu, J. D., & Han, B. C. (2015). Theoretical considerations for developing use-oriented China's Standards of English. *Modern Foreign Languages*, (1), 78-90.

Liu, J. D., & Peng, C. (2017). Developing Scientific China's Standards of English. *Foreign Language World*, (2), 2-9.

Lorch, R. F. (1989). Text-signaling devices and their effects on reading and memory processes. *Educational Psychology Review, 1*(3), 209-234. https://doi.org/10.1007/BF01320135

Meyer, B. J., & Freedle, R. O. (1984). Effects of discourse type on recall. *American Educational Research Journal, 21*(1), 121-143. https://doi.org/10.3102/00028312021001121

NEEA. (2015). *An introduction to NMET: A trial version for provinces with renovation of college entrance examinations*. Beijing: Higher Education Press.

North, B. (2003). *Scales for rating language performance: Descriptive models, formulation styles, and presentation formats*. TOEFL Monograph, 24.

Perlman, C. C. (2003). Performance assessment: designing appropriate performance tasks and scoring rubrics. *Ability*, (12), 497-506.

Plakans, L. (2013). Writing scale development and use within a language program. *TESOL Journal, 4*(1), 151-163. https://doi.org/10.1002/tesj.66

Popham, W. J. (1997). What's Wrong and What's Right with Rubrics. *Educational Leadership, 55*(2), 72-75.

Rodriguez, A. (2008). The 'problem' of creative writing: using grading rubrics based on narrative theory as solution. *New Writing, 5*(3), 167-177. https://doi.org/10.1080/14790720802209963

Sasaki, M., & Hirose, K. (1999). Development of an analytical rating scale for Japanese 11 writing. *Language Testing, 16*(4), 457-478. https://doi.org/10.1177/026553229901600403

Shaw, S. D., & Weir, C. J. (2007). *Examining Writing: Research and Practice in Assessing Second Language Writing*. Cambridge: Cambridge University Press.

SMEEA. (2017). *A handbook for NMET Shanghai version*. Shanghai: Shanghai Ancient Books Press.

Stawiarska, L. (2016). The influence of summary writing on the development of reading skills in a foreign language. *System, 59*, 90-99. https://doi.org/10.1016/j.system.2016.04.006

Stein, B. L., & Kirby, J. R. (1992). The effects of text absent and text present conditions on text. *Journal of Reading Behavior, 24*(2), 217-232. https://doi.org/10.1080/10862969209547773

Turner, C. E., & Upshur, J. A. (1996). Developing rating scales for the assessment of second language performance. *Australian Review of Applied Linguistics*, (13), 55-79. https://doi.org/10.1075/aralss.13.04tur

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49-70. https://doi.org/10.2307/3588360

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732997

Westhoff, G. (2007). Challenges and opportunities of the CEFR for reimagining foreign language pedagogy. *The Modern Language Journal, 91*(4), 676-679. https://doi.org/10.1111/j.1540-4781.2007.00627_9.x

Yang, H. C. (2014). Toward a model of strategies and summary writing performance. *Language Assessment Quarterly, 11*(4), 403-431. https://doi.org/10.1080/15434303.2014.957381

Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing, 24*(4), 539-572. https://doi.org/10.1177/0265532207080780

Yu, G. (2008). Reading to summarize in English and Chinese: a tale of two languages? *Language Testing, 25*(4), 521-551. https://doi.org/10.1177/0265532208094275

Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing, 14*(2), 116-137. https://doi.org/10.1016/j.asw.2009.04.002

Yu. (2013). The use of summarization tasks: some lexical and conceptual analyses. *Language Assessment Quarterly, 10*(1), 96-109. https://doi.org/10.1080/15434303.2012.750659

Zeng, Y., & Fan, T. (2017). Developing reading proficiency scales for EFL learners in China. *Language Testing in Asia, 7*(8), 1-15. https://doi.org/10.1186/s40468-017-0039-y

Zhao, C. G. (2013). Measuring authorial voice strength in L2 argumentative writing: the development and validation of an analytic rubric. *Language Testing, 30*(2), 201-230. https://doi.org/10.1177/0265532212456965

Zou, S. (2011). *A course book of English testing* (2nd ed.). Higher Education Press.

Appendix 1

The Questionnaire of Textual Attributes of Summary Writing

Dear teachers,

Thank you so much for attending our surveys.

Summary writing requires students to describe succinctly, with their own words, about the given source texts. This questionnaire focuses on to what extent certain textual attributes, i.e. features a good summary writing should have, can be included in the rating scale of summary writing. The questionnaire consists of two parts, the first of which collects your basic information while the second deals with your views or attitudes towards the textual attributes. Please make judgement based on your teaching experience.

The information you are to provide is very important to us. It will only be applied in my research and kept confidential to others.

Part I. Basic information

1. Your gender:

A. Male B. Female

2. Your age:

A. 25-30 B. 31-35 C. 36-40 D. 41-45 E. over 46

3. Your professional title:

A. teaching assistant B. Lecturer C. Associate professor D. Professor

4. The university where you work:

A. "985" Universities B. "211" Universities

C. ordinary universities D. junior colleges

5. Your academic title:

A. bachelor B. master student C. master D. Ph.D. student E. Ph.D.

6. Your research focus:

A. literature B. theoretical linguistics C. applied linguistics D. translation

7. The length of your English teaching:

A. 1-3 years B. 4-6 years C. 7-9 years D. 10-15 years E. over 15 years

8. The location where you work:

A. Northern China B. Southern China C. Eastern China D. Western China E. Central China

9. Your involvement in summary writing:

A. Never B. Seldom C. often D. always

10. Your involvement in studies on summary writing:

A. Never B. Seldom C. often D. always

11. The length of your teaching English writing:

A. N/A B. 1-2 years C. 3-5 years D. 6-10 years E. over 10 years

12. Do you think it necessary to score writing with rating scales?

A. completely disagree B. basically disagree C. not sure D. basically agree E. completely agree

13. Your experiences of rating summary writing in large-scale tests?

A. N/A B. 1-2 times C. 3-4 times D. 5-6 times E. over 7 times

14. Have you asked your students to do summary writing exercises and provide your feedback?

A. Yes (Answer Questions 13-14); B. No (No need to answer Questions 13-14)

15. How many times do you ask your students to do summary writing exercises and provide your feedback each academic year on average?

A. 1-2 times B. 3-5 times C. 6-8 times D. 9-12 times E. ≥13 times

16. What criteria do you employ to judge the quality of summary writing by your students?

A. self-made rating scales B. direct use of existing rating scales from various tests C. adaptation of existing rating scales from various tests

Part II. Your attitude towards the inclusion of the following textual attributes into the rating scale of summary writing.

17. The theme of summary writing is the same as that of source texts

18. Write with a clear structure and distinct layers

19. Summarize source texts content briefly and comprehensively

20. Use conjunctions to make more natural paragraph transitions

21. Use conjunctions to make more natural sentence transitions

22. Write with succinct language without too lengthy and wordy expressions

23. Use subordinate clauses with flexibility to avoid sentences with loose structures

24. Use various sentence patterns with frequent changes

25. Avoid using too long or too short sentences

26. No inclusion of content unavailable in source texts

27. Accurate and appropriate use of punctuation

28. Describe content of source texts with clear priorities

29. Use words correctly and properly without spelling mistakes

30. No appearance of grammatical errors

31. Write with smooth diction and fluent language

32. Cover all the major points in source texts

33. No change of logical relations or opinions in source texts

34. No opinions or comments on source texts

35. Use idioms or slangs correctly when necessary

36. Keep the style of the source texts in summary writing

37. Use diversified vocabulary to avoid repetition and tediousness

38. Use some advanced vocabulary accurately and properly

39. Use capitalized letter accurately and properly

40. Neat handwriting and clean sheet

41. Use rhetoric devices properly to enhance expressive effects

42. Write with clear diction without causing misunderstandings

43. Properly use complicated grammatical structures (e.g. subjunctive mood)

44. Use all kinds of tenses and voices accurately and properly

45. Use standardized language without expressions too oral or internet words

Appendix B

A finalized version of rating scale of summary writing

Linguistic Accuracy (25%)	Linguistic Complexity (20%)	Coherence & Cohesion (25%)	Fidelity to Source Texts (20%)	Mechanism (10%)
▲No spelling mistakes ▲Write with no grammatical mistakes ▲Use vocabulary accurately and appropriately ▲Write with smooth and fluent language	▲Use frequently lots of advanced vocabulary ▲Use rich vocabulary ▲Use diversified sentence patters ▲Use diversified clauses with flexibility ▲Use lots of complicated sentence patterns	▲Very clear structures and hierarchical levels ▲Skillfully use various devices to make natural connections between sentences and between paragraphs ▲Succinct language without lengthy expressions	▲Tell apart completely major and minor information in source texts and write with appropriateness of length accordingly ▲Add no additional information unavailable from source texts ▲Cover all major points in source texts ▲Objectively present source texts without any change	▲Use punctuation completely accurately ▲Write with completely clear handwriting and neat presentation ▲Write with normalized language without any ambiguity ▲Write with normalized language without oral or internet–based expressions
▲A few spelling mistakes ▲Write comparatively accurately with just a few grammatical mistakes ▲Write comparatively correctly with just a few words used inappropriately ▲Write with comparatively smooth and fluent language	▲Use comparatively many advanced vocabulary ▲Use comparatively rich vocabulary ▲Use comparatively diversified sentence patters ▲Use comparatively many clauses ▲Use comparatively many complicated sentence patterns	▲Write with comparatively clear structures and hierarchical levels ▲Comparatively skillfully use various devices to make comparatively natural connections between sentences and between paragraphs ▲Write with comparatively succinct language with very few lengthy expressions	▲ Tell apart comparatively accurately major and minor information in source texts and write with comparatively appropriateness of length accordingly ▲Add a little additional information unavailable in source texts ▲Cover comparatively completely major points in source texts, leaving out just a few ▲Comparatively objectively present source texts with just a few changes	▲Most punctuations are used accurately ▲Write with comparatively clear handwriting and neat presentation ▲Write with comparatively normalized language with just a little ambiguity ▲Write with comparatively normalized language with just a few oral or internet –based expressions
▲Write with some mistakes in spelling ▲Write basically accurately with some grammatical mistakes ▲Write basically correctly with some words used inappropriately ▲Write with basically smooth and fluent language	▲Use limited number of advanced vocabulary ▲Vocabulary lacks diversity & seems a bit monotonous ▲Sentence patters lack diversity & seems a bit monotonous ▲Use limited number of clauses ▲Use limited number of complicated sentence patterns	▲Write with basically clear structures and hierarchical levels ▲Use basic devices to make a degree of natural connections between sentences and between paragraphs ▲Write with basically succinct language with some lengthy expressions	▲ Tell apart basically accurately major and minor information in source texts and write with basically appropriateness of length accordingly ▲Add some additional information unavailable in source texts ▲Cover basically completely major points in source texts, leaving out some of them ▲Basically objectively present source texts content with some changes	▲ Punctuations are used generally and basically accurately ▲Write with comparatively clear handwriting and neat presentation ▲Write with basically normalized language with just some ambiguity ▲Write with basically normalized language with just some oral or internet –based expressions
▲Write with lots of mistakes in spelling ▲Write inaccurately with lots of grammatical mistakes ▲Write incorrectly with lots of words used inappropriately ▲Write with unsmooth language	▲Use a few advanced vocabularies ▲Vocabulary not diversified but rather monotonous ▲Sentence patters not diversified but rather monotonous ▲Use only a few clauses ▲Use only a few complicated sentence patterns	▲Write with rather poor structures and hierarchical levels ▲Use only a few basic devices to make unnatural connections between sentences and between paragraphs ▲Write with language of poor succinctness and many lengthy expressions	▲Tell apart inaccurately major and minor information in source texts and write with inappropriateness of length accordingly ▲Add much additional information unavailable in source texts ▲Leave out many of the major points in source texts ▲Much content of the source texts is changed	▲Make lots of mistakes in using punctuations ▲Write casually in an unclear manner ▲Write with less normalized language with much ambiguity ▲Write with lots of oral or internet –based expressions that are not normalized enough
▲Write with most words wrongly spelled ▲Write inaccurately with grammatical mistakes throughout the whole essay ▲Write incorrectly with	▲Use few or no advanced vocabulary ▲Vocabulary extremely not diversified but extremely monotonous ▲Sentence patters extremely monotonous ▲Use few or no	▲Write with completely chaotic structures and hierarchical levels ▲Fail to use devices to make connections between sentences and between	▲Fail completely to tell apart major and minor information in source texts and write with complete inappropriateness of length accordingly ▲Most of what is written is not related to what is in	▲Make mistakes in most cases in using punctuations ▲Write casually in an extremely unclear manner and are hard for identification ▲Write with

most of the words used inappropriately ▲ Write with language not smooth at all	clauses ▲Use very short and simple sentence patterns	paragraphs ▲Write with language of no succinctness	source texts ▲Leave out most of the major points in source texts ▲Most content of the source texts is changed	less normalized language with much ambiguity ▲Write with oral or internet –based expressions almost throughout the essay