Li and Wang *Asian-Pacific Journal of Second and Foreign Language Education* (2021) 6:11
https://doi.org/10.1186/s40862-021-00113-6

Asian-Pacific Journal of Second and Foreign Language Education

ORIGINAL ARTICLE Open Access

Development and validation of a rating scale for summarization as an integrated task

Jiuliang Li[1*] and Qian Wang[2]

* Correspondence: lijiu@hotmail.com
[1]Beijing Institute of Fashion Technology, NO.2 Yinghua Road, Chaoyang District, Beijing 100029, PR China
Full list of author information is available at the end of the article

Abstract

Summary writing is essential for academic success, and has attracted renewed interest in academic research and large-scale language test. However, less attention has been paid to the development and evaluation of the scoring scales of summary writing. This study reports on the validation of a summary rubric that represented an approach to scale development with limited resources out of consideration for practicality. Participants were 83 students and three raters. Diagnostic evaluation of the scale components and categories was based on raters' perception of their use and the scores of students' summaries which were analyzed using multifaceted Rasch measurement (MFRM). Correlation analysis revealed significant relationships among the scoring components, but the coefficients among some of the components were over high. MFRM analysis provided evidence in support of the usefulness of the scoring rubric, but also suggested the need of a refinement of the components and categories. According to the raters, the rubric was ambiguous in addressing some crucial text features. This study has implications for summarization task design, scoring scale development and validation in particular.

Keywords: Summary writing, Scale development, Task design, Validation, Practicality

Introduction

The ability to summarize English articles has been emphasized in both secondary (Zhang, 2007) and tertiary education (Chen & Su, 2012) in China. Summarization skills have been taking on new importance as more and more Chinese college students seek further education in western universities where summarizing skills are long considered "essential to academic success" (Kirkland & Saunders, 1991, p.105). Undergraduates in these institutions are often required to summarize complex concepts and information in every subject they are studying. Surveys of academic tasks across disciplines reveal that this task is not only assigned in a variety of university classes but also plays an important role in more advanced, complex university writing assignments, such as article critiques and research papers (e.g., Carson, 2001; Hale et al., 1996). In addition, summarization and some other integrated tasks have come into some major international language testing programs, such as TOEFL (Yu, 2009). Therefore the need

Springer Open

© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

becomes urgent to sharpen students' summarization edge by incorporating the task into classroom assessment and large-scale language tests in China, where dramatic changes are to take place along with the recent launch of Chinese Standard of English Language Ability (CSE). However, summarization tasks are complex in nature (Cohen, 1993), and even more so when it comes to scoring (Yu, 2007). The task of developing summary criteria is extremely thorny for college instructors in general due to the expertise and efforts it requires. Rating scale constitutes an essential part in the process of summarization task design, the rating, score reporting and interpreting. The goal of the present study is to validate a rating scale developed for summary writing as an integrated task that is expected to be used in the assessment of English as foreign language (EFL) in China.

Review of related literature

Development of scoring criteria for summary writing

Determination of major points

Making appropriate choices as to what is important in the source material defines the major characteristic of the ability to summarize, according to some of the most comprehensive and extensively quoted definitions of summary (e.g., Hidi & Anderson, 1986; McAnulty, 1981). Thus it is desirable that the scoring scheme defines the major points of the original text in order to effectively assess the efficiency of students' summary writing (Yang, 2014).

The important ideas of a text can be determined by formal propositionalization of a source text based on Kintsch and van Dijk's representational situation model (Kintsch & van Dijk, 1978; van Dijk & Kintsch, 1977, 1983) and Meyer's (1975) structural content hierarchy system. Another common approach for test developers to follow involves the use of native speaker experts to rate the importance/priority of information in a source text, or to produce a summary of the source text (Yu, 2007).

However, neither of the above offers a practical solution to developing scoring criteria for EFL context use. Formal propositionalization involves enormous time commitment and expertise (Bernhardt, 1991: 202–203; Mills et al., 1993) and the models are found incapable to propositionalize extended texts (Schnotz, 1983). As Urquhart and Weir (1998) sighed, test constructors may spend a huge amount of time reading and rereading to stripe off deeper and deeper levels of meaning. These will impose a tremendous burden on college teachers who are not only occupied with heavy workload normally but are also likely lacking in the required expertise. The difficulty will be much more tremendous if we think of the situation that summarization tasks may be assigned frequently in classroom teaching. The second approach involving native speaker experts to help identify the important ideas from a source text is still impractical in an EFL context such as China. Teachers may have difficulty in obtaining help from a sufficient number of native experts. What's more, "even the experts did not fully agree on which ideas were essential to the construction of a meaningful summary" (Cohen, 1993: 137).

The consideration of a viable solution in this scenario pertains to practicality, an important aspect of test usefulness conceptualized by Bachman and Palmer (1996). Practicality entails determining the resources available in relation to the resources required to strike an optimum balance among, for instance, the test qualities, reliability and

construct validity. The consideration of practicality becomes a particular concern for the present study that seeks to develop summarization scoring scheme for tertiary level use under EFL context.

Development of rating scale

Rating scale can be developed through intuition-, theory-, and empirically-based methods (Knoch, 2009). Each of these is grounded on different types of knowledge, thus a mixed-method approach has been increasingly used to collect complementary information for rubrics development and validation (e.g., Cumming, Kantor, & Powers, 2001; Lim, 2012; Shaw & Weir, 2007). Intuitive methods include expert judgments, committee and experiential methods. An example of these methods is the *Experiential scale design*, which normally begins with expert judgment or committee design, then the scale is polished over time by its users. This is by far the "most common method of scale development" (Knoch, 2009, p.43). However, researchers argue that intuitively developed scales may invite subjectivity (Fulcher, 2012; Galaczi, French, Hubbard, & Green, 2011). Thus, it is desirable that development of rating scale should be guided by relevant theories.

McNamara (1996) and Weigle (2002) rightly pointed out, the rating scale that is used in assessing writing performance should embody the theoretical basis of a writing test, thus scale development need to ascertain that the scoring criteria should provide a clear and credible basis for scoring judgments, and for different levels of writing performance. Similarly, Xi (2008) suggested that scales "that do not reflect the relevant knowledge and skills could lead to erroneous scores" (p.183). In this regard, some classical summarization models help to identify the major mental operations involved in summary writing that could be incorporated into scoring scale. In the Kintsch and Van Dijk (1978) model, for instance, summary protocols operate at the global level according to three macrorules that transform the microstructure (propositions) of the text to produce a macrostructure, which can be considered as a summary:

1. Deletion: the disposal of unnecessary information;
2. Generalization: the coherent condensation of information, and.
3. Construction: the invention of global representations in place of sets of components, conditions or consequences.

Closely corresponding to the above mentioned macrorules, Brown and Day (1983) identified the following activities as essential to producing adequate summaries of lengthy texts: deletion of trivial and redundant information; replacement of more general, superordinate concepts for a list of specific items (e.g., vegetable for tomato, eggplant, and cucumber); and finally, selecting (if available) or making (if necessary) a topic sentence for each paragraph.

Johnson (1983) focused on six operations involved in writing adequate summaries. The first four activities, comprehending individual propositions, establishing links between them, identifying the structure of the text, and remembering the content, are identified as prerequisites for summarization. The other two processes, selecting the information to be placed in the summary and composing a concise and coherent verbal

representation, are seen as central to summarization. Johnson also suggested that in order to produce concise summaries, writers must carry out transformations on the information they identify as important, such as deletion of inferable ideas and substitution of segments by contracting original information.

These important mental operations have been addressed in the research on read-to-write tasks as discourse synthesis (Yang & Plakans, 2012), and more recently on the nature of integration of L2 reading and writing skill as shared process (Plakans, Liao, & Wang, 2019), as well as rating scale development in the integrated tasks involving reading and writing (Chan, Inoue, & Taylor, 2015). The findings can provide insight into our understanding of the complex processes linking reading and writing in a second language and into considerations about how best to represent both the reading and writing dimensions of test taker performance in the rubric descriptors.

The scoring scales used in previous studies could also help inform the scale development process in summarization assessment and research (e.g., Rivard, 2001; Sawaki, 2003; Yang, 2014; Yu, 2008). In this regard, a notable study was on testing French as L2 (Rivard, 2001), in which ten variables were selected for evaluation. Four variables with which to evaluate summary writers concerned issues related to the content of the summaries: the ability to identify main ideas, the ability to identify secondary ideas and supporting details, the ability to integrate ideas, and faithfulness to the text. Five variables related to the language of the summaries were also included in the study. Four of these were scored using analytic scales: organization, style, language usage, and objectivity. The fifth variable is an overall language score as rated holistically. The last variable, summarization efficiency, is a quantitative measure which examines both content and language that has been used in a number of studies on summary writing. This scale is deemed comprehensive to evaluate most of the skills required in the task. However, in the case of language assessment, where raters have to read a large number of written scripts, ten variables may be too much and may eventually affect reliability. Nevertheless, these criteria, as well as those in other summary studies, provide valuable information to the development of scoring criteria for the present study.

As for the type of scoring scale, the analytic schemes are preferred over holistic rubrics by many writing specialists on the ground that they "provide more detailed information about a test taker's performance in different aspects of writing" (Weigle, 2002, p.115), and hence more conducive to the evaluation of learners' writing development and more suitable for teachers' classroom instruction and assessment as well as learners' self-assessment. Thus it was decided for the present study to develop an analytic summary writing scale.

Studies of rating scale validation

Validation of rating scales is a necessary undertaking, because a rubric with well-defined score categories facilitates consistent scoring. The validity of rating scales for L2 writing assessment has been investigated in a number of studies, most of which focuses on large-scale, high-stakes assessments (e.g., see Chapelle, Enright, & Jamieson, 2008; Shaw & Weir, 2007; Weir, Vidakovic, & Galaczi, 2013). Some studies have examined the distinctness of the analytic dimensions using multifaceted Rasch measurement (MFRM) (e.g., Lallmamode, Daud, & Kassim, 2016; McNamara, 1996). Knoch (2009)

compared the performance of a theoretically-based and empirically-developed rating scale and a pre-existing scale. She employed questionnaires and interviews to elicit perception data from raters. She also employed FACETS analysis which included measures of discrimination of the rating scale, rater separation, rater reliability, variation in ratings, and scale step functionality. Results based on the above data showed the new scale worked better than the existing scale. Asención (2004) conducted a validation study for the rating scale of a summarization task. She performed correlation and FACETS analysis and found that the scoring components of the summary rubric were related aspects that described participants' summary performance. The analysis of the bands in the scoring categories revealed that the assumption that they were appropriately describing different levels of performance was partially met. In most of the categories, it was observed that overall the bands described different levels of performance.

Although research on tests that involve summary writing is on the rise (Yu, 2009), so far how summary writing as an integrated task should be scored has not been sufficiently addressed. Little attention is paid to the development of a rating scale specifically for summary writing in comparison to other types of writing, for example independent writing and response essay (but see Yu, 2007). Though regarded essentially as a writing task (Kim, 2001), summary writing differs from the average composing activity (Hidi & Anderson, 1986) as well as other types of read-to-write tasks, such as response essay (Asención Delaney, 2008). On the other hand, there is little research that investigates the validity of rating scales for summary writing, classroom-based assessments in particular, where moderate- to high-stakes decisions are often made against students' performance (e.g., course grade assignment, program advancement and/or exiting). Most of the summarization studies employed a scoring scheme without investigating the validity of that scheme leaving the issue of score interpretation sometimes questionable. Yu (2007) summarized the scantiness of research on rating scales of summary writing as the result of the challenges associated with developing an adequate scoring scheme and maintaining satisfactory scoring reliability. Cohen (1993, 1994) early has it that scoring summaries can be extremely knotty as it involves a risk of rendering the task potentially unreliable.

In order to enhance confidence and nuanced understanding of summary writing as an integrated task, a more focused effort must be taken to develop the rating scales and validate their use in L2 writing classrooms as well as large-scale tests. The present study attempts to perform diagnostic evaluation of a scoring scale that is designed for summarization tasks in both classroom setting as well as large-scale language assessment programs after appropriate adaptation. We first elaborate on the development of a rating scale with consideration given to test practicality, and then set out to validate the said scale.

The present study

Development of the scoring criteria

The scoring criteria consisted of an analytic rating scale and a model summary. This is believed to be able to improve accuracy and reliability as such a design encourages efforts of double checking, and it is expected to enhance rating efficiency. These serve both the needs of classroom assessment and large-scale testing.

The model summary

To cope with the above issues in relation to practicality, we finally decided to use the ready-made model summaries. The textbook for the participants' use in classroom instruction was developed by the Foreign Language Teaching and Research Press (FLTR P), who not only provided the texts to be summarized, which were considered fit in terms of topic and difficulty, but also the model summaries included in the accompanying material for teachers' use.

The quality of the model summaries were checked based on the definitions (Friend, 2002; McAnulty, 1981), rules (Kintsch & van Dijk, 1978), and procedures (Brown & Day, 1983; Friend, 2000, 2002; Johnson, 1983) of summary provided in existing literature. The model summaries proved to be good as they fit the above criteria. In spite of this, the e-version of the texts and the model summaries were sent to two native speakers of English who were asked to make comments and suggestions, and accordingly, some minor changes were made.

The scoring scale

Given the limited resources available in consideration for test practicality, the present study was in favor of an intuitive approach supplemented with a theory-based method for scale development. To be specific, we developed the scale based on existing rating criteria as well as judgment and opinions from experts of applied linguistics, language assessment in particular, then refined it over a period of time. This process is guided by models of summarization (e.g., Kintsch & Van Dijk, 1978) and theories of writing, Grabe and Kaplan's (1996) model of text construction in particular.

The resultant analytic scoring scale originally contains five components: Main Idea Coverage (MIC), Faithfulness (FAIT), Integration (INT), Language Use (LU) and Source Use (SU). Each of the components can be scored on a 0–5 scale with each followed by a descriptor, the ratings of each components were averaged to produce a final score for a given participant.

The first component focuses on the number of main ideas included in the written summary, which is considered the central concern of a good summary. FAIT deals with the factual inaccuracies, additions or embellishments in the written summary, such as false ideas, errors, generalizations, interpretations, evaluations and exaggerations (Rivard, 2001). However, after consulting an expert in language assessment, the component FAIT was removed on the ground that it might be confused with MIC on the part of raters, because if a main idea of the source text is not faithfully conveyed, it should also not be considered creditable in the component of MIC. So basically they are about the same thing. On the other hand, if eliminated, the number of subscales would be reduced from five to four, which would be certainly more convenient and operational for raters to use, and hence would in turn improve rating efficiency and task practicality.

INT examines the extent to which the information in the text is presented succinctly by using strategies such as deleting unnecessary information, combining and condensing information across sentences and paragraphs, reordering information in text, and by displaying smart use of connectives. LU is also considered essential as a review of previous scoring criteria for integrated writing tests shows that language use is one of the three major features frequently assessed in these tests, the other two being content

and organization. In the component of language use, grammar, syntactic variety and vocabulary are the major criteria for evaluation.

In the component of SU, the evaluation is performed in terms of the accurate use and verbatim use of source information. Yang (2009) wrote that "the source use in essays should be evaluated because an appropriate use of source materials is expected in all academic writing contexts" (p.41). Source use was given more attention in the scale than those in other summary studies. Research indicates that patchwriting, interwoven with sentences or phrases copied from original sources characterize L2 composing by university students as indicated by their summary writing. The use of SU component is meant to draw teachers and students' attention to appropriate use of source text so as to avoid and enhance awareness of plagiarism, which is regarded as dishonesty and cheating (Leask, 2006; Pecorari, 2001; Yamada, 2003), and is deemed a more serious problem among L2 students as observed by researchers (e.g., Currie, 1998; LoCastro & Masuko, 1997; Matalene, 1985; Myers, 1998; Pennycook, 1996). For details about the analytical scale and the descriptors of the categories please see the Appendix.

Scale validation

For the validation analysis, we attempt to answer the following two research questions: (1) Does the rating scheme give appropriate assessment for the summaries at different levels of performance? (2) What are the raters' perceptions of the usefulness of the rating scale?

Methods

Test takers

A sample of 83 EFL learners was drawn from an undergraduate program in a Chinese university. All the participants were in their early 20s and had been learning English for at least 6 years. Generally speaking, the sample was at the intermediate level of English proficiency according to their NMET (Chinese national matriculation English test) scores (mean = 92.6 on a 0–150 scale), which is mostly aligned with CEFR level B2 (Papageorgiou, Wu, Hsieh, Tannenbaum, & Cheng, 2019). One month before data collection, the participants were provided instruction on summary writing and were given the opportunities to practice writing summaries both as in- and after-class assignment.

The summarization task

Two source texts accompanied with model summaries were chosen for use in the summarization tasks, which were given to the participants within a 5-day interval. The texts, one narrative and the other expository in genre, were taken from a college English textbook developed by FLTRP so that they were fit for the test in terms of topic and difficulty. The task instructions stated that the students should read the text first, and then write an English summary for about 130 words without copying the source. To make the scoring more operational and improve accuracy and reliability, the model summaries were divided into idea units based on Kroll's (1977) definition. In this study, a statement was a loosely defined idea unit in the form of a complete clause or sentence (Yu, 2007). These idea units were put into a table, and a certain range of number of main ideas was correspondingly allocated to the five bands of the MIC component.

This information constituted a frame of reference for raters and was expected to ease the rating process and improve accuracy and efficiency.

Raters

Three native Chinese researchers (with the pseudonyms of Leo, Cathie, and Zalia) acted as raters for the study, including the researcher himself (Leo). The other two researchers were postgraduate students, one majoring in language testing and the other in second language acquisition. All the raters have experience in rating essays for large-scale tests. Rating occurred in two sessions. The first session focused on the narrative summaries, and the second focused on the expository summaries. Before the first rating session, the three raters underwent a training session that familiarized them with the test tasks, the source texts, and the scoring criteria. To facilitate the induction process the training included a pilot rating session in which the summary scripts of three participants were used. The scores of these three participants were excluded from the follow-up statistical analysis. The results showed that the reliability estimates of the ratings were fairly high for both narrative ($\alpha = .804$) and expository ($\alpha = .802$) summarization tasks.

Interview

Raters (Cathie and Zalia) attended an interview immediately after they had completed the second rating session to talk about (1) what features they attended to in the summaries; (2) how they made judgment about test taker's summarization ability; (3) what factors affected their rating; and (4) how they made use of the scoring scale and how the scoring scale functioned. This is supposed to facilitate an investigation into the validity issue associated with the scoring scheme of the summarization tasks.

MFRM analysis

For quantitative analysis of participants' summary scores, MFRM was used with the FACETS 3.58 (Linacre, 2005). Four facets were included: examinees, tasks, raters, and rubric components. The examinee facet included 83 elements. The task facet consisted of two tasks using the above two texts. The three raters served as judges in the rater facet. The rubric component facet included the four components in the scoring scale. FACETS calibrates the examinees, raters, tasks, and the rubric components onto the same equal-interval scale (i.e., the logit scale), where higher Rasch measures mean examinees hold greater ability, raters are more lenient, and tasks are more difficult.

Results

Functioning of components and categories

Functioning of components

Functioning of the components is evaluated based on correlation analysis and FACETS statistics for the calibrated scores in the components. In this study, the four components of summarizing ability were expected to show certain degree of overlap, as they were assumed to represent different aspects of unidimensional ability. Table 1 summarizes the relationships between rubric components which were explored with correlation analysis using the average scores given by the three raters for the two tasks for

Table 1 Correlation between rating components (Spearman's rho)

	MIC	INT	LU	SU
MIC	1.000			
INT	.731(**)	1.000		
LU	.659(**)	.880(**)	1.000	
SU	.578(**)	.884(**)	.821(**)	1.000

** Correlation is significant at the 0.01 level (2-tailed)

each examinee. For the two summarization tasks as a whole, the correlations ranged from 0.578 to 0.884, with the lowest being that between MIC (Main Idea Coverage) and SU (Source Use) and the highest being that between INT (Integration) and LU (Language Use). All correlations were significant at the 0.01 level.

Table 2 presents the four components of the analytical scale in difficulty order, from 0.47 logits (SE = .06) for Language Use, the hardest, to − 0.31 logits (SE = .06) for Source Use, the least hard, encompassing a 0.78-logit span. The average scoring component difficulty is 0.00 with a corresponding measurement error of 0.06. The reliability index is 0.97, which suggests that the components are thus reliably distinguished across different levels of difficulty. The difference between the difficulty of these four components is statistically significant ($\chi^2 = 109.0$, $df = 3$, $p < .01$).

The fit indices for the four components of the scoring scale (MIC, INT, LU and SU) are within the range of good fit as proposed by McNamara (1996). They closely cluster around the expected value of 1 within a range of 0.06. This indicates that 1) the rating patterns for each of the four scoring components are very close to those expected by the FACETS model; 2) in terms of the measurement dimension constructed by the analysis, it makes sense to add the scores from the different components together; and 3) scores in the components MIC, INT, LU and SU are making independent contributions to the underlying measurement dimension; in that sense the components can be said to have been validated (McNamara, 1996).

Functioning of categories

Following the guideline proposed by Bond and Fox (2007), several measures were used to diagnose the rating categories: category frequencies and average measures, threshold estimates, probability curves, and category fit. They are, as Bond and Fox stressed, "very useful in pointing out where we might begin to revise the rating scale to increase the reliability and validity of the measure" (p. 226).

Table 2 Components measurement report

Component	Obsvd count	Obsvd average	Fair-M average	Measure	Model S.E.	Infit MnSq	ZStd
LU	492	2.1	2.05	.47	.06	.94	−1.0
INT	492	1.9	1.84	.02	.06	.96	−.5
MIC	492	2.2	2.16	−.18	.06	1.06	.9
SU	492	2.1	2.02	−.31	.06	1.05	.8
Mean	492.0	2.1	2.02	.00	.06	1.00	.0
S.D.	.0	.1	.13	.34	.00	.06	1.0

Separation 5.77; Reliability .97; Fixed chi-square: 109.0, d.f.: 3, significance: .00

Category frequencies and average measures

The simplest way to evaluate category functioning is to look at category use statistics (i.e., category frequencies and average measures) for each response option (Andrich, 1996; Linacre, 1999, as cited in Bond & Fox, 2007). These category frequencies present the distribution of the responses across all categories, allowing for a very quick and basic analysis of rating scale use. Shape distribution is an essential feature in the category frequencies, and regular distributions such as unimodal distribution is preferable to those that are irregular. Average measures are defined as the average of the ability estimates for all persons in the sample with the average calculated across all observations in the category. These average measures are expected to increase in size as the variable increases. A monotonic increase indicates that on average, candidates with higher ability are placed in the higher categories.

Table 3 shows the FACETS output for the rating scale by rubric components. Not all categories were used by raters, who did not endorse Category 5 for three (MIC, INT and SU) out of four components. MIC, INT, LU, and SU are all unimodal (i.e., possessing a unique mode) in terms of shape distribution of category frequencies (i.e., the observed count). Average measures (in logit) appear in the next column. For all components, the average examinee ability measures increased in magnitude as the rubric categories increased. This suggests that examinees with higher ratings on a particular component were indeed more able than examinees with lower ratings on the same component. For instance, the average measure for Category 1 of MIC is −.02, meaning that the average ability estimate for persons being scored 1 is −.02 logits. For the persons who were scored 2, the average ability estimate is .50. (i.e., these persons are more able on average than the persons who are scored 1). It can be seen that these average measures across the components functioned as expected (i.e., they increase monotonically across the rating scale of the four components). This means the categories of the rating scale performed normally according to the diagnosis of the above two measures.

Threshold and category fit

In addition to category frequency and the monotonicity of average measures, other pertinent rating scale characteristics include thresholds, or step calibrations, and category fit statistics (Lopez, 1996; Wright & Master, 1982, as cited in Bond & Fox, 2007). Step calibrations are the difficulties measured for being scored one category over another (e.g., how difficult it is to obtain a '4' over '3') (Bond & Fox, 2007). Like the average

Table 3 FACETS output for rating scale by rubric components

	MIC		INT		LU		SU	
Category label	Observed count	Average measure	Observed count	Average measure	Observed count	Average measure	Observed count	Average measure
0	4	−.91	5	−1.42	2	−1.87	4	−1.17
1	109	−.02	194	−.15	151	−.74	165	.16
2	202	.50	150	.53	172	−.05	152	.75
3	146	1.14	111	1.10	114	.51	108	1.20
4	31	1.87	32	1.67	48	1.07	63	1.89
5					5	2.20		

Keys: *MIC* Main Idea Coverage, *INT* Integration, *LU* Language Use, *SU* Source Use

measures, thresholds (step calibrations) should increase monotonically. Thresholds that do not increase monotonically across the rating scale are deemed disordered. Table 4 reveals that the thresholds across all components functioned well (i.e., they increase monotonically across the components of the rating scale).

Another helpful indicator of rating scale functionality is the outfit mean-square statistic and this statistic is calculated for each rating scale category by FACETS. Mean-squares have an expectation of 1.0. The INFIT MnSq is not reported because it "approximates the OUTFIT MnSq when the data are stratified by category" (Linacre, 2011, p.186). As can be seen in Table 4, the outfit mean-square indices for the categories of rubric components range from 0.7 to 1.2 which are near to the expected value of 1.0. The largest distance from the expectation of 1.0 is the outfit statistic for Category 5 of Language Use (0.7) which is also the only component where the full rating scale (0–5) is used. This is hardly surprising because extreme categories have greater opportunity for unexpected mean-squares than central categories (Linacre, 2011). Overall the findings suggest that all components were functioning as expected by the model.

Probability curves

Probability curves provide another type of information for evaluating the quality of rating scales. It shows the probability of endorsing a given rating scale category for every agreeability-endorsability difference estimate. Each category should have a distinct peak in the probability curve graph, illustrating that each is indeed the most probable category for some portion of the measured variable (Bond & Fox, 2007; Wiseman, 2008). Below are the probability curve graphs for the rating scale categories of the four rubric components.

As these figures displayed similar patterns, they are discussed in aggregate. The figures show each category of the four components has a distinct peak, which means that they met the above-mentioned criterion. However, category 2 and 3 seemed a little problematic as they defined much less wide intervals on the latent variables than the other categories in Figs. 2, 3, and 4. This means that the definitions (wording) of these categories may need rewording so that they could define wider intervals.

By and large, the above analysis using the diagnostic measures suggested by Bond and Fox (2007) shows the rating scale meet the relevant criteria. All the four rubric components possess a unique mode in shape distribution of category frequencies. In terms of average measures, for all components, the average examinee ability measures

Table 4 Threshold and category fit

	MIC		INT		LU		SU	
Category label	**Threshold**	**OUTFIT MnSq**	**Threshold**	**OUTFIT MnSq**	**Threshold**	**OUTFIT MnSq**	**Threshold**	**OUTFIT MnSq**
0	None	1.0	None	1.0	None	1.0	None	1.0
1	−3.82	1.1	−4.17	.9	−5.41	.9	−3.89	1.0
2	−.40	1.1	.45	1.0	−.50	.9	.51	1.2
3	1.18	1.0	1.10	.9	.63	1.0	1.35	1.1
4	3.04	.9	2.63	1.1	1.66	1.0	2.11	1.0
5					3.62	.7		

increase in magnitude as the rubric categories increased. This is also true for threshold estimates which increased monotonically. In terms of category fit, the outfit mean-square indices for the categories of rubric components are all near to the expected value of 1.0. And the probability curve graphs showed that each category has a distinct peak.

Raters' perception of the functioning of scoring rubric

The scoring rubric played a key role in the rating process as reflected in follow-up raters' interviews. Overall the raters thought the rubric was rational and helpful. However, they also identified some problems and uncertainties encountered in using the scoring rubric, mostly about ambiguity of the rubric in addressing crucial text features.

Main idea coverage: raters, particularly Zalia, experienced some difficulties in applying the criteria stipulated in this component:

> How to define 'main idea point'? When an idea is just mentioned, should it be taken as one or a half point? How about only the key words of one main idea are written? A harsher rater would not accept them. This issue calls for careful thinking and reasoning. Zalia

Source use: Raters expressed similar concerns over the vagueness in rating this component:

> In the beginning, I was not quite sure to what extent is the use of original text can be defined as a copy or a paraphrase. Then through discussion, it was clearer to me. Zalia

> In the beginning, I was not quite sure about the scale of Source use. It is difficult sometimes to determine whether a sentence is written in the writer's own language. In the first glance, you may be thinking many sentences in a summary are copied from the text. But through careful examination, you changed your mind. There are just some words or expressions that are similar to the source. On the whole, the writer used his/her own sentence structure to combine several ideas from the text. Sometimes I became confused so I went back to the text. Cathie

Integration: With regard to this component raters' view converged to some extent. Cathie mentioned the needs to include her own criteria in rating this component which can be taken as an indication of the vagueness of the rubric in addressing this issue:

> I think the scale of Integration requires raters to see whether the essay is logical. Sometimes it was difficult to make judgment. Cathie

Zalia struggled at the boundaries between score levels:

> In scoring the component of Integration, it was sometimes difficult to distinguish between category 3 and 4, i.e. Good and Very Good. Category 3 requires writers

> 'displays moderate examples', while category 4 requires 'displays good examples of integration'. I think the distinction between 'moderate' and 'good' is different to different people. Zalia

Where the dilemma was difficult to solve, Zalia resorted to her impression of other components:

> Oftentimes, when it was difficult to distinguish between 3 and 4, I turned to the candidate's general linguistic ability. I saw whether the writer performed well in Language use, used his/her own language most of the times, and presented sufficient information. If yes, I was inclined to give a 4, otherwise, a 3. Zalia

Zalia added it would be helpful to provide raters with exemplar essays to illustrate how to distinguish the levels of integration.

Language use: Zalia offered little account of her experience in rating language use and she seemed to be doing smoothly in this respect. On the contrary, Cathie encountered much difficulty in this respect, wavering between 0 and 1, for instance:

> I found it hard to make decisions in scoring Language Use; you see, s/he wrote many things after all, though a little disordered sometimes. I was wavering between 0 and 1 in the rating scale. Cathie

Discussion

This study represents an attempt to construct and validate an intuition and theory-based summary writing scale. Despite the criticisms directed at intuitively developed scales, they are still widely utilized in assessments across the world (Knoch, 2009), either solely (e.g., Lallmamode et al., 2016) or combined with other approaches to create new or retrofit existing rating criteria (e.g., Deygers & Van Gorp, 2015; Hawkey & Barker, 2004). Intuitive approach to scale development is considered appropriate particularly where resources are confined, such as the present study, and has been found to demonstrate the effectiveness and practicality as opposed to the empirically developed data-based scales (Lallmamode et al., 2016).

To seek answers to the two research questions which were formulated to examine whether if the newly developed rating scale was built with well-defined score components that could facilitate consistent scoring, the present study performed analyses with both qualitative and quantitative methods. For the latter, various measures were taken to collect information about the scoring components which includes the correlations (Table 1), and FACETS statistics for the calibrated scores in each component (Table 2). The correlation coefficients among the components (from .578 to .884) showed that all correlations were significant at the 0.01 level, indicating that the components measured related aspects of the ability to write a summary. The INT, LU and SU components were the aspects of the summary performance more highly related (from .821 to .884) than the comparisons in which the MIC component was involved. This is perhaps due to the fact that these components (INT, LU and SU) were all measuring abilities more related to the dimension of writing than reading and therefore were expected to show some degree of overlap. The highest correlation coefficient was found between INT

and SU which was somewhat unexpected in the first glance. However, this relationship was understandable upon reasoning, as these two components were more or less predicated on the same premise to the extent that integrating different chunks of source material requires textual operation across sentences and paragraphs to represent the global meaning of the text, and the process of this representation entails students' use of own words and sentence structures.

Meanwhile, these high correlations might also be due to factors at work in the rating process which involves raters' personal belief and knowledge. Some researchers (e.g., Weigle, 2002) have shown that when raters use analytical rating scales they often display a halo effect, because their overall impression of a writing script (or the impression of one aspect of the writing script) guides their rating of each of the traits (e.g., see Knoch, 2011). During the rater interview Zalia, for instance, expressed a tendency to rely on LU and SU for judging INT where she encountered difficulty in making a decision. Apart from the halo effect, there might be other causes leading to this tendency. For one thing, it may be because raters are not clear of the cognitive operations that are involved in INT (integration); in this case, better training should be conducted. For another, the criteria in this part may not be clear enough to raters who then had to resort to other component for help. In this case, these components need to be refined, such as redesigning, rewording, or merger between categories, so as to better differentiate performances representing different aspects of the construct.

The scoring rubric was also examined with a many-faceted Rasch model built by FACETS which yielded results largely in favor of the performance of the scoring components. The reliability index is 0.97 (Table 2), which suggests that the components are thus reliably distinguished across different levels of difficulty. The difference between the difficulties of these four components is statistically significant, which indicated that the components were measuring different aspects of the ability. The scoring component mean-square infit statistics were close to 1, showing that there was no unexpected variation among the component scores. Therefore, these aspects were working together in the measure of the summarizing ability. This result was congruent with the correlation analysis performed on the summary rubric components. These suggest that these scoring criteria could provide the information necessary to place students in the appropriate L2 levels.

Evidence was then sought to determine whether the categories of the scoring components in the rubric described different levels of participants' summarizing performance, because if the measure does not increase with each higher category, then "doubt is cast on the idea that larger response scores correspond to 'more' of the variable" (Linacre, 2011, p.186). To this end, information on calibrated scores provided by the FACETS program was obtained. These included average measures (Table 3), threshold estimates (Table 4), and probability curves (Figures 1, 2, 3 and 4), and they largely showed evidence of good performance of the categories in the components. However, category 2 and 3 in the component of INT, LU, and SU seemed problematic as they defined much less wide intervals on the latent variables than the other categories in these components. The finding revealed a concern that has been encountered by previous studies that examined the process of scale development. Asención (2004), for instance, found that the bands of the rubric could not clearly differentiate all levels of performance.

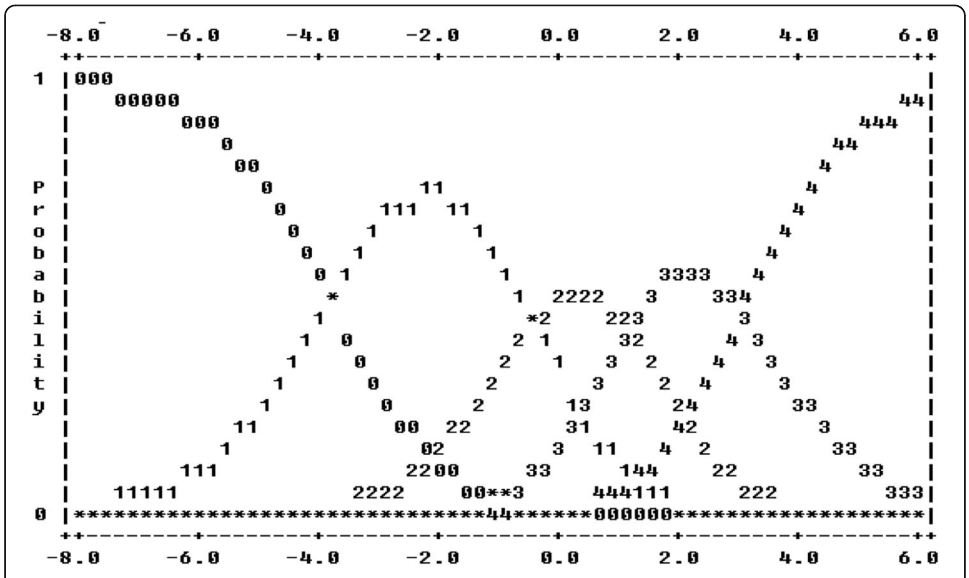

Fig. 1 Probability curves for Main Idea Coverage

All the components in the rubric have six categories (i.e., 0–5), but three of them (MIC, INT, SU) functioned with five categories. The little use of category 5 in the scoring components could be explained by the fact that the sample was at the intermediate level of English proficiency which has restricted the variability of scores that ideally should be reflected in the scoring categories. Another plausible explanation is that the raters were very cautious of giving the components full scores which denotes that a summary was free of error and fully met the criteria, which is seldom the case with most EFL learners unfortunately.

The interview provided important information with respect to the raters' perception of the functionality of the scoring components, revealing the concerns over the usefulness of the criteria in differentiating ability at different levels. Overall less criticism is leveled at Main idea coverage and Language use. For the former, perhaps the table of

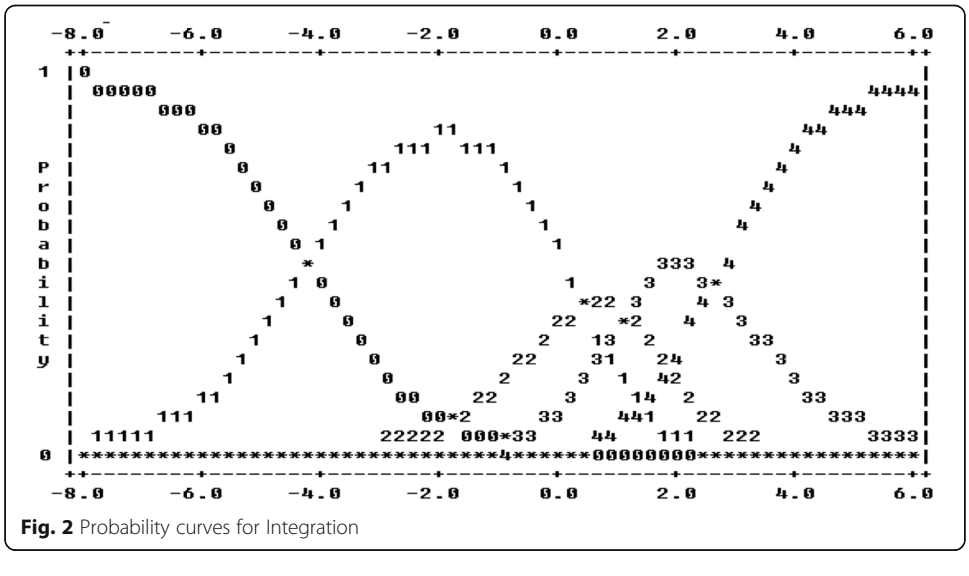

Fig. 2 Probability curves for Integration

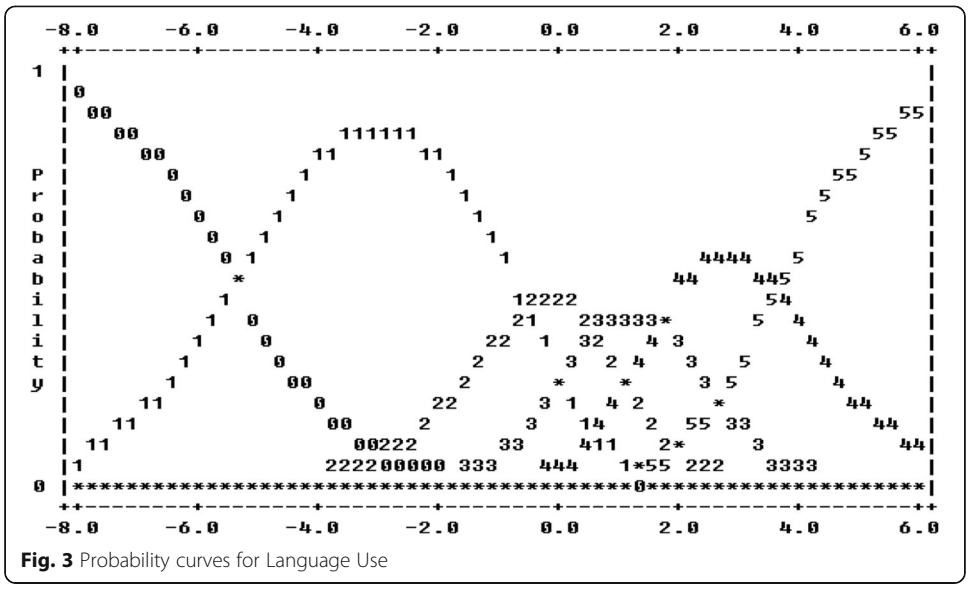

Fig. 3 Probability curves for Language Use

idea units constructed based on the model summary contributed to the relative ease in using the component. For the latter, raters may be less familiar to the construct of the other components than to language use, which is stressed in virtually all rating scales of EFL writing. In contrast, the raters showed less confidence in using Integration, expressing difficulty in distinguishing the categories. The categories may need to be redesigned or couch them in terms that better distinguish performance at different levels.

As for Source use, raters raised concerns about the vagueness of the term "copy" and "paraphrase" used in the scale, and mentioned the trouble of frequently shuttling between the text and the summaries to assess the extent to which sentences in the scripts were formulated with students' own vocabulary and structures. It is suggested the task of tackling source use be coped with automatic detection technology, i.e. those described in Mandin, Lemaire, and Dessus's (2007) report, to ease the burden and improve efficiency and accuracy.

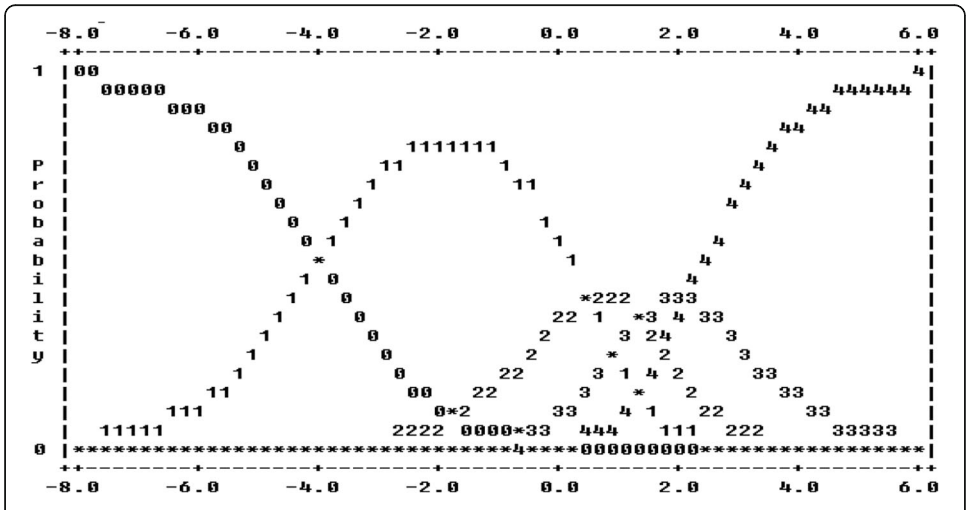

Fig. 4 Probability curves for Source Use

Conclusion

The present study aimed to conduct validation of an analytic scale that was developed based on intuition and theory of summary writing for use in classroom assessment and large-scale testing as well. The scoring scale played a key role in the rating process with the template as an aid for the raters. MFRM analysis, the diagnostic measures in particular as suggested by Bond and Fox (2007), was carried out to see if the scoring components were related aspects of the ability to write a summary and if the categories discriminated among different levels of performance. Examination of the scoring components and their categories provided evidence in support of the use of the scoring rubric, but also suggested, and is confirmed by rater interviews, the need of refinement of the components and categories to better describe the differing levels of summarization performance. The high correlation coefficients among some of the components are still in need of a plausible interpretation. Perhaps there are some relations between these significant correlations and the narrow intervals defined in some of the categories as revealed in the MFRM analysis. These are yet to be found out and dealt with in future research.

This research confirms what has been proposed by other related studies that the Many facet Rasch Model provides better evidence of validity in the assessment of scoring rubrics. In particular, the measures proposed by Bond and Fox (2007) are useful in giving diagnostic evaluation of an analytical scale. This information helps to identify the possible weakness to which remedial efforts could be prescribed so that the scale could yield useful information about students' summarizing abilities. An adequate scoring scheme would help to achieve satisfactory reliability and reduce subjectivity in scoring.

This study has implications for the design of summarization tasks, particularly the scoring rubric, which embodies "what underlying abilities are being measured by the test" (Knoch, 2009, p.60). The validity of a scoring scale may be appropriately examined from at least two perspectives: one is related to the scale itself, and the other is to raters who use the scale. The present research is in keeping with previous studies that a good knowledge of the construct of summarization tasks is needed and should constitute the basis for building an analytical scale, so that each of its components represents a distinct aspect of the summarizing ability. Rubric with clear and sound constructs are crucial for performance evaluation, because evidence supporting the evaluation inference is based on, among other things, the "care with which the scoring rubrics are developed and applied" (Xi, 2008, p.182). With such knowledge in mind, careful wording of the criteria should be performed to avoid vagueness and confusion. These could be identified by expert judgment and statistical analysis using MFRM. In the scoring process, it is desirable that the principles underlying the development of the summary scale be efficiently communicated to the raters, because the ways in which rating scales and rating criteria are constructed and interpreted by raters act as the de facto test construct (McNamara, 2002; Turner, 2000).

The study takes a practical approach that is believed to be able to strike an optimum balance among the resources available in constructing a rating scale for summary writing. Hopefully, the approach could offer a viable solution for teachers to use summarization as an efficient tool to foster learner development and evaluate language proficiency in both formative and summative assessment practices in the college-level foreign language education in China.

With the proposed approach, as well as the research findings made, the present study also holds implications for the design and scoring of integrated tasks in large-scale national English tests, such as college English test (CET), which currently do not take

summary writing tasks. With an on-going in-depth study and understanding of the relevant theories, decision could be made about how best to formulate and represent the constructs of the tasks. Diagnostic evaluation and improvement of the rating scale would help enhance scoring reliability, which has been a major concern for integrated test tasks. The inference made about test-takers' summarization ability would be more accurate and constructive, which would, in turn, enhance the consequential validity of the tasks. Given the fact that A growing number of language tests (e.g., The Internet-based Test of English as a Foreign Language, Canadian Academic English Language Assessment, General English Proficiency Test, and Georgia State Test of English Proficiency) have incorporated tasks that involve summarization in their assessment batteries (Yang, 2014), it is suggested that the task be given serious consideration in major domestic tests.

Appendix

Rating scale

Main idea coverage

5 EXCELLENT: A response has complete coverage of main ideas

4 VERY GOOD: A response has coverage of most main ideas

3 GOOD: A response has moderate coverage of main ideas

2 MODERATE: A response has some coverage of main ideas

1 POOR: A response has coverage of very few ideas

0 NO: A response has no coverage of main ideas

Integration

5 EXCELLENT: A response rearranges the order of the statements logically, displays excellent examples of integration and connectives, and demonstrates global interpretation of the source text

4 VERY GOOD: A response rearranges the order of the statements logically, displays good examples of integration and connectives, and demonstrates global interpretation of the source text

3 GOOD: A response rearranges the order of the statements logically, displays moderate examples of integration and connectives, and demonstrates global interpretation of the source text

2 MODERATE: A response basically follows the order of source text with few cases of re-ordering and integration, and is not global in the interpretation of the source text

1 POOR: A response follows the original order of the statements in the source text, shows rare instance of proper integration and connectives, and is not global in their interpretation of the source text

0 NO: A response has no instances of integration or connectives at all

Language use

5 EXCELLENT: A response displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice; it is within the word limit as required

4 VERY GOOD: A response displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, or word form that do not interfere with meaning; it is basically within the word limit

3 GOOD: A response demonstrates inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning; and/or it exceeds the word limit to a noticeable degree

2 MODERATE: A response has a noticeably inappropriate choice of words or word forms, an accumulation of errors in sentence structure and/or usage; and/or it exceeds the word limit to a large degree

1 POOR: A response has serious and frequent errors in sentence structure or usage, the text shows a lack of control of vocabulary and/or grammar; and/or it exceeds the word limit to a large degree

0 NO: A response is totally incomprehensible due to language errors, or because the response is left blank

Source use

5 EXCELLENT: A response is predominantly in the summarizers' own words and sentence structures, in addition to the accurate use of the information from the source text

4 VERY GOOD: A response is mostly in the summarizers' own words and sentence structures, in addition to the accurate use of the information from the source text

3 GOOD: A response is basically in the summarizers' own words and sentence structures, in addition to appropriate use of information from the source text

2 MODERATE: A response has some use of the summarizers' own words and sentence structures, in addition to the adequate use of the information from the source text

1 POOR: A response is predominately verbatim copying the source text

0 NO: A response demonstrates completely verbatim copying from the source text

Abbreviations

CEFR: Common European framework of reference for languages; CET: College English test; CSE: Chinese Standard of English Language Ability; EFL: English as foreign language; FAIT: Faithfulness; FLTRP: Foreign Language Teaching and Research Press; MFRM: Multifaceted Rasch measurement; MIC: Main Idea Coverage; INT: Integration; LU: Language Use; NMET: Chinese national matriculation English test; SU: Source Use; TOEFL: Test of English as a foreign language

Acknowledgements

Not applicable.

Authors' contributions

Jiuliang Li carried out research design and the writing of the paper. Qian Wang carried out data collection and analysis.

Funding

National Education Examinations Authority of China & British Council. Award number: EARG2020002.
Beijing Institute of Fashion Technology. Award number: JG-1923.

Availability of data and materials

The data will not be shared with a reason.

Declarations

Ethics approval and consent to participate

The need for approval was waived.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Beijing Institute of Fashion Technology, NO.2 Yinghua Road, Chaoyang District, Beijing 100029, PR China. [2]North China Electric Power University, No.2, Bei Nong Road, Changping District, Beijing 102206, PR China.

Received: 24 September 2020 Accepted: 3 May 2021
Published online: 01 July 2021

References

Andrich, D. (1996). Measurement criteria for choosing among models with graded responses. In A. von Eye, & C. C. Clogg (Eds.), Categorical variables in developmental research: Methods of analysis, (pp. 3–35). San Diego, CA: Academic Press.
Asención Delaney, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes*, *7*(3), 140–150. https://doi.org/10.1016/j.jeap.2008.04.001.
Asención, Y. (2004). *Validation of reading-to-write assessment tasks performed by second language learners*. Unpublished PhD thesis, Northern Arizona University.
Bachman, L. F., & Palmer, A. (1996). *Language testing in practice*. Oxford: Oxford University Press.
Bernhardt, E.B. (1991). *Reading development in a second language: Theoretical, empirical, and classroom perspectives*. Norwood, NJ: Ablex.
Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences*. Mahwah: Lawrence Erlbaum.
Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: the development of expertise. *Journal of Verbal Learning and Verbal Behavior*, *22*(1), 1–14. https://doi.org/10.1016/S0022-5371(83)80002-4.
Carson, J. (2001). A task analysis of reading and writing in academic contexts. In D. Belcher, & A. Hirvela (Eds.), *Linking literacies: perspectives on L2 reading-writing connection*, (pp. 48–83). Ann Arbor: The University of Michigan Press.
Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skills: A case study. *Assessing Writing*, *26*, 20–37. https://doi.org/10.1016/j.asw.2015.07.004.
Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the test of English as a foreign language.™*. New York: Routledge.
Chen, Y. S., & Su, S. W. (2012). A genre-based approach to teaching EFL summary writing. *ELT Journal*, *66*(2), 184–192. https://doi.org/10.1093/elt/ccr061.
Cohen, A. D. (1993). The role of instructions in testing summarizing ability. In D. Douglas, & C. Chapelle (Eds.), *A new decade of language testing research*, (pp. 132–159). Washington, DC: TESOL.
Cohen, A. D. (1994). English for academic purposes in Brazil: the use of summary tasks. In C. Hill, & K. Parry (Eds.), *From testing to assessment: English as an international language*, (pp. 174–204). London: Longman.
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. TOEFL monograph series 22*. Princeton: Educational Testing Service.
Currie, P. (1998). Staying out of trouble: apparent plagiarism and academic survival. *Journal of Second Language Writing*, *7*(1), 1–18. https://doi.org/10.1016/S1060-3743(98)90003-0.
Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, *32*(4), 521–541. https://doi.org/10.1177/0265532215575626.
Friend, R. (2000). Teaching summarization as a content area reading strategy. *Journal of Adolescent & Adult Literacy*, *44*(4), 320–329.
Friend, R. (2002). Summing it up – teaching summary writing to enhance science learning. *The Science Teacher*, *69*(4), 40–43.
Fulcher, G. (2012). Scoring performance tests. In G. Fulcher, & F. Davidson (Eds.), *The Routledge handbook of language testing*, (pp. 378–392). London: Routledge.
Galaczi, E., French, A., Hubbard, C., & Green, A. (2011). Developing assessment scales for largescale speaking tests: a multiple-method approach. *Assessment in Education: Principles, Policy & Practice*, *18*(3), 217–237.
Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing*. New York: Longman.
Hale, G., Taylor, C., Bridgeman, B., Carson, J., Kroll, B., & Kantor, R. (1996). *A study of writing tasks assigned in academic degree programs*. Princeton: Educational Testing Service.
Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing*, *9*(2), 122–159. https://doi.org/10.1016/j.asw.2004.06.001.
Hidi, S., & Anderson, V. (1986). Producing written summaries: task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, *56*(4), 473–493. https://doi.org/10.3102/00346543056004473.
Johnson, N. S. (1983). What do you do if you can't tell the whole story? The development of summarization skills. In K. E. Nelson (Ed.), *Children's language*, (vol. 4, pp. 315–383).
Kim, S. A. (2001). Characteristics of EFL readers' summary writing: A study with Korean University students. *Foreign Language Annals*, *34*(6), 569–581. https://doi.org/10.1111/j.1944-9720.2001.tb02104.x.
Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*(5), 363–394. https://doi.org/10.1037/0033-295X.85.5.363.
Kirkland, M. R., & Saunders, M. A. P. (1991). Maximizing student performance in summary writing: Managing cognitive load. *TESOL Quarterly*, *25*(1), 105–121. https://doi.org/10.2307/3587030.
Knoch, U. (2009). *Diagnostic writing assessment: the development and validation of a rating scale*. Frankfurt: Peter Lang.
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.
Kroll, B. (1977). *Combining ideas in written and spoken English: a look at subordination and coordination*. Discourse across time and space, 5.
Lallmamode, S. P., Daud, N. M., & Kassim, N. L. A. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing*, *30*, 44–62. https://doi.org/10.1016/j.asw.2016.06.001.
Leask, B. (2006). Plagiarism, cultural diversity and metaphor—Implications for academic staff development. *Assessment & Valuation in Higher Education*, *31*(2), 183–199. https://doi.org/10.1080/02602930500262486.
Lim, G. (2012). Developing and validating a mark scheme for writing. *Cambridge English Research Notes*, *49*, 6–10.
Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103–122.
Linacre, J. M. (2005). *A user's guide to FACETS. Computer software manual*. Chicago: Winsteps. com.

Linacre, J. M. (2011). *A user's guide to FACETS. Computer software manual*. Chicago: Winsteps. com.
LoCastro, V., & Masuko, M. (1997). *Plagiarism and academic writing of NNS learners*. Paper presented at the TESOL 1997 Meeting, Orlando, FL.
Lopez, W. A. (1996). The resolution of ambiguity: An example from reading instruction (Doctoral dissertation, University of Chicago, 1996). *Dissertation Abstracts International, 57*(07), 2986A.
Mandin, S., Lemaire, B., & Dessus, P. (2007). Modeling summarization assessment strategies with LSA. In F. Wild, M. Kalz, J. Bruggen, & R. Koper (Eds.), *Mini-proccedings of the 1st European workshop on latent semantic analysis in technology – enhanced learning*, (pp. 20–21). NL: Heerlen.
Matalene, C. (1985). Contrastive rhetoric: an American writing teacher in China. *College English*, *47*(8), 789–808. https://doi.org/10.2307/376613.
McAnulty, S. J. (1981). Paraphrase, summary, précis: advantages, definitions, models. *Teaching English in the Two-Year College*, *8*, 47–51.
McNamara, T. (1996). *Measuring second language performance*. New York: Addison Wesley Longman Limited.
McNamara, T. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, *22*, 221–242. https://doi.org/10.1017/S0267190502000120.
Meyer, B.J.F. (1975). *The organization of prose and its effects on memory*. Amsterdam: North-Holland.
Mills, C.B., Diehl, V.A., Birkmire, D.P., & Mou, L.C. (1993). Procedural text: Predictions of importance ratings and recall by models of reading comprehension. *Discourse Processes*, 16, 279–315.
Myers, S. (1998). *Questioning author(ity): ESL/EFL, science, and teaching about plagiarism*. TESL-EJ, 3. Retrieved August 15, 2000, from http://www-writing.berkeley.edu/TESL-EJ/ej10/a2.html
Papageorgiou, S., Wu, S., Hsieh, C-N., Tannenbaum, R.J., & Cheng, M. (2019). *Mapping theTOEFL iBT®test scores to China's standards of English language ability: implications for score interpretation and use. TOEFL ® research report TOEFL–RR-89 ETS research report no. RR–19-44*.
Pecorari, D. (2001). Plagiarism and international students: how the English-speaking university responds. In D. Belcher, & A. Hirvela (Eds.), *Linking literacies: perspectives on L2 reading–writing connections*, (pp. 229–245). Ann Arbor: The University of Michigan Press.
Pennycook, A. (1996). Borrowing others' words: text, ownership, memory, and plagiarism. *TESOL Quarterly*, *30*, 210–230.
Plakans, L., Liao, J.-T., & Wang, F. (2019). "I should summarize this whole paragraph": shared processes of reading and writing in iterative integrated assessment tasks. *Assessing Writing*, *40*, 14–26. https://doi.org/10.1016/j.asw.2019.03.003.
Rivard, L. P. (2001). Summary writing: a multi-grade study of French-immersion and francophone secondary students. *Language, Culture and Curriculum*, *14*(2), 171–186. https://doi.org/10.1080/07908310108666620.
Sawaki, Y. (2003). *A comparison of summarization and free recall as reading comprehension tasks in web-based assessment of Japanese as a foreign language*. Unpublished PhD thesis, University of California, Los Angeles (UCLA), Los Angeles.
Schnotz, W. (1983). On the influence of text organization on learning outcomes. In G. Rickheit, & M. Bock (Eds), *Psycholinguistic studies in language processing*, (pp.152–81). Berlin and New York: de Gruyter.
Shaw, S. D., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*. Cambridge: University Press.
Turner, C. E. (2000). Listening to the voices of rating scale developers: Identifying salient features for second language performance assessment. *The Canadian Modern Language Review*, *56*(4), 555–584. https://doi.org/10.3138/cmlr.56.4.555.
Urquhart, A.H., & Weir, C.J. (1998). *Reading in a second language: Process, product and practice*. London: Longman.
van Dijk, T.A., & Kintsch, W. (1977). Cognitive psychology and discourse: Recalling and summarizing stories. In W.U. Dressler (Ed.), *Current trends in textlinguistics*, (pp. 61–80). New York: de Gruyter.
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.
Weir, C. J., Vidakovic, I., & Galaczi, D. E. (2013). *Measure constructs: a history of cambridge english examinations. 1913–2012. Studies in language testing, 37*. Cambridge: Cambridge University Press.
Wiseman, C. S. (2008). *Investigating selected facets in measuring second language writing ability using holistic and analytic scoring methods*. Unpublished PhD thesis, Columbia University.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.
Xi, X. (2008). Methods of test validation. In E. Shohamy, & N. H. Hornberger (Eds.), *Encyclopedia of language and education: vol. 7. Language testing and assessment*, (2nd ed., pp. 177–196). New York: Springer Science+Business Media LLC.
Yamada, K. (2003). What prevents ESL/EFL writers from avoiding plagiarism? Analyses of 10 north-American college websites. *System*, *31*(2), 247–258. https://doi.org/10.1016/S0346-251X(03)00023-X.
Yang, H. (2009). *Exploring the complexity of second language writers' strategy use and performance on an integrated writing test through structural equation modeling and qualitative approaches*. Unpublished PhD thesis, University of Texas at Austin.
Yang, H. (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.
Yang, H., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated Reading-listening-writing task. *TESOL Quarterly*, *46*(1), 49–70.
Yu, G. (2007). Students' voices in the evaluation of their written summaries: empowerment and democracy for test takers? *Language Testing*, *24*(4), 539–572.
Yu, G. (2008). Reading to summarize in English and Chinese: a tale of two languages? *Language Testing*, *25*(4), 521–551.
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.
Zhang, X. (2007). *A construct validation research of the writing task in NMET (GD) — text-based english writing for Chinese EFL Learners*. Unpublished PhD thesis, Guangdong University of Foreign Studies, Guangzhou, China.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.