Towards more valid scoring criteria for integrated reading-writing and listening-writing summary tasks

Abstract

Despite the increased use of integrated tasks in high-stakes academic writing assessment, research on rating criteria which reflect the unique construct of integrated summary writing skills is comparatively rare. Using a mixed-method approach of expert judgement, text analysis and statistical analysis, the current study examines writing features that discriminate summaries produced by 150 candidates at five levels of proficiency on integrated reading-writing (R-W) and listening-writing (LW) tasks. The expert judgement revealed a wide range of features which discriminated R-W and L-W responses. When responses at five proficiency levels were coded by these features, significant differences were obtained in seven features, including relevance of ideas, paraphrasing skills, accuracy of source information, academic style, language control, coherence and cohesion and task fulfilment across proficiency levels on the R-W task. The same features did not yield significant differences in L-W responses across proficiency levels. The findings have important implications for clarifying the construct of integrated summary writing in different modalities, indicating the possibility of expanding integrated rating categories with some potential for translating the identified criteria into automated rating systems. The results on the L-W indicate the need for developing descriptors which can more effectively discriminate L-W responses.

Keywords: Scoring validity, rating scale, integrated tasks, summary, reading-writing, listening-writing

Introduction

The use of integrated tasks such as those involving summarising ideas from reading passages and listening recordings is well supported in the current literature on writing assessment. Integrated tasks require “complex cognitive, literate, and language abilities for comprehension as well as to produce written compositions that display appropriate and meaningful uses of and orientations to source evidence, both conceptually and textually’’ (Cumming et al., 2005, p.34). As language use is ultimately a meaning-making process, Pollitt and Taylor (2006) made a convincing argument for integrated tasks, as did Weigle (2004), on the grounds that asking candidates to produce a summary in writing/speaking of what they have read/listened to allows access to the otherwise unobservable mental representation of source use.

The recent increased use of integrated tasks, including summary tasks, in academic writing assessment has stimulated much research interest in clarifying the construct of

integrated skills. Earlier studies compared candidates' cognitive processes (e.g., Chan, 2011; Plakans, 2008, 2010) and written products (e.g., Cumming et al., 2005; Kyle & Crossley, 2016) between independent and integrated writing tasks. The findings of these studies were conclusive that the two task types elicit distinct constructs of writing. Knoch and Sitajalabhorn (2013) argue that the processes underpinning an integrated performance need to be adequately translated into the rating descriptors used to evaluate the written product. This is, however, not always the practice. Descriptors are particularly opaque in the context of tests where integrated tasks are machine scored. Before more adequate rating criteria can be developed, there is an urgent need to identify the distinctive features which discriminate integrated performances produced by learners at different levels of proficiency.

Literature review

The construct of integrated writing has received an increasing amount of attention in the language testing literature over the last two decades. Summary writing, among other kinds of integrated tasks, has been of particular interest. The literature review first describes the current understanding about summary writing and identifies gaps in research regarding the construct of this particular kind of integrated writing in relation to its distinguishing features across proficiency levels. The review then establishes the importance of developing a better understanding of rating criteria of relevance to summary writing in tests that are machine-scored. Lastly, an overview of different approaches to scale development is provided to establish the basis of the design of the current study.

Studies on integrated tasks and summary writing

One important dimension of research on integrated tasks is to examine features of source use in integrated writing performances. Plakans and Gebril (2013) investigated features of source use on the TOEFL integrated task (reading-listening-writing) and found that three aspects of source use — importance of ideas, use of ideas from inputs, and verbatim reproduction of text from the source — explained over 50% of the variance in the task

scores. High-scoring writers selected important ideas from the sources and used the listening text as the task prompt instructed. In contrast, low-scoring writers depended heavily on the reading text for content and language. Ohta, Plakans and Gebril (2018) compared holistic and multi-trait scoring methods for scoring EFL university essays and reported that source use was the most reliable feature among all dimensions of the multi-trait rubric. Cho and Choi (2018) examined features which reflect writers' awareness of an audience in integrated essays produced by adult English as a Second Language (ESL) writers under two conditions: with and without contextual factors of audience and purpose in the task directions. Based on an analysis of writing from 205 candidates, they found that features of audience awareness (including context statements, content, and source attribution) differed across performance levels. These studies conclusively pointed to the need of incorporating source use in integrated rating scales to reflect the construct of integrated skills. Indeed, following up the inquiry about the importance of source use in summary writing, Sawaki's (2020) recent study examined two types of summary content scoring methods (content point scores and a holistic rating scale) for university L2 academic writing essays. Results showed that source use based on both summary content scoring methods was distinct from a language quality score, leading the researcher to conclude that employing a content criterion with the language quality rating would enhance the representation of the summary writing construct and support the meaningfulness warrant for the test score interpretation claim.

However, as most previous studies focused on argumentative essays, empirical evidence of how source use is operationalised in summary test tasks is largely absent. Summary tasks require considerably less writing than many of the tasks in research previously surveyed, and therefore warrant attention as a unique integrated task type. Consideration of summary tasks in this way is crucial given the important role of task effects in candidates' writing performance (In'nami and Koizumi, 2016). To date, however, there have been few studies investigating the relationship between task features and integrated writing performance.

Li (2014)'s study, one of the handful of studies on the relationship between task features and summary writing (Hidi & Anderson, 1986; Kirkland & Saunders, 1991; Yu, 2009), found that Chinese university students performed better on expository summary writing than on summaries of a narrative text. In addition, the difficulty estimates of all rating criteria included main idea coverage, integration, source use and language use. Taken together, these results indicate that features of sources influence candidate's performance in integrated tasks. However, there is limited research, if any, on impact of input modality on candidate's source use in summary test tasks.

Rating criteria of integrated writing in machine-scored tests

Another ongoing focus of research has been the rating approach of integrated writing in machine-scored tests. International examination providers have developed or adapted automated essay evaluation (AEE) applications to score integrated writing performances in their tests; for example, Educational Testing Service (ETS)'s e-rater, Pearson's Intelligent Essay Assessor (IEA) and Vantage Learning's IntelliMetric. The scoring of essays by most AEE applications is conducted through a statistics-based approach to identify the most dominant observed variables in essays (Koskey & Shermis, 2013). These dominant variables, such as number of sentences and words, use of verb/noun agreement, spelling, punctuation, and capitalization tend to be mechanics of writing and features that are quantifiable. They might not necessarily reflect the key features of integrated writing, including selection of ideas and effectiveness of source use, and hence risk representing only a small fraction of the construct of integrated writing skills. In addition, although most AEE systems establish high correlations between human scores (often obtained by using holistic scales) and the machine scores, researchers have cautioned that evidence of high reliability for the automated approach should not be the only criterion to establishing scoring validity of a test (Bennett & Bejar, 1997; Bennett & Zhang, 2016; Weir, 2005).

Test providers incorporate AEE in their scoring systems in a number of ways. After a series of validation studies to ensure that machine scores are comparable to scores given by skilled human raters, the integrated writing tasks (i.e., 50- to 70-word reading-writing/listening-writing summaries) in the Pearson Test of English Academic (PTEA) are scored by machine solely in relation to content (i.e. the extent to which relevant aspects are mentioned in the summary), grammar, spelling, vocabulary and form (Pearson PTE, 2019). The TOEFL iBT integrated writing tasks (i.e., 250-word reading-listening-writing essays) are scored by the scoring models of E-rater® using a set of 11 features with nine representing aspects of writing quality and two representing content, in conjunction with a human scorer using a holistic scale (Burstein, Tetreault, & Madnani, 2013). As established earlier in the literature review, selection of ideas and content integration are important features of effective integrated writing. AEE systems, in general, are able to score 'content' by comparing candidates' responses with the large set of training responses and assigning a score based on computed similarities. Nevertheless, the rating scales used for integrated writing tasks in the PTE machine-scored test seem to rely more heavily on mechanic writing features than content. Researchers have called for a fuller representation of the integrated writing construct in integrated scales, especially those used in machine-scored tests (Knoch & Sitajalabhorn, 2013, Yu, 2013). Advancements in technology should not lead us away from the calls for a more empirically-based approach to rating scale development (Fulcher, 1996). The development and use of integrated tasks enables closer definition of the construct of integrated skills, which should then inform the development of appropriate integrated scales so that the features selected in the machine-scored models can be expanded to represent a fuller range of the construct of integrated skills. To establish the theoretical basis to achieve this, we now review studies on the role of expert judgement and data-driven rating scale construction.

Role of expert judgement and data-driven rating scale construction

The research on the role of expert judgement and data-driven integrated rating scale construction has implications for identifying key features of integrated written performance.

Using think-aloud protocols, Cumming et al. (2001) investigated how raters evaluated integrated (reading-listing-writing) and independent writing tasks from TOEFL iBT. Based on the think-aloud protocols of seven raters, they found that raters focused more on rhetoric and content when rating integrated tasks, whereas they focused more on language when rating independent tasks. In the context of a University integrated R-W test, Gebril and Plakans (2014) examined raters' processes of evaluating integrated essays in detail. Raters in their study engaged in procedures including locating source information, checking citation mechanics and evaluating quality of source use. Hence, they argued that source use should be incorporated in integrated rating criteria. In the context of standardised multiple level tests (CEFR A2-C1), Chan, Inoue and Taylor (2015) reported a large scale study of data-driven rating scale construction for Trinity Integrated Skills of English (ISE), a standardised multiple level integrated tests (CEFR A2-C1). As part of the study, they investigated raters' processes of evaluating integrated R-W essays using level-specific analytic scales. The results revealed that some features, including misunderstanding of sources, were more prominent in lower proficiency levels. Other features such as reading comprehension and identifying main ideas discriminated more effectively between lower proficiency levels than higher proficiency levels where most candidates demonstrated effective reading comprehension skills. In contrast, raters noticed more features of paraphrasing and content integration among higher-level responses. These studies demonstrated the role of expert judgement in identifying features of integrated performance within and between proficiency levels, and the value of empirical data-driven rating scale construction. Nevertheless, they tended to focus on one type of integrated writing task – typically, 250-word (argumentative) essays. Less research attention has been placed on other types of integrated test tasks, especially summary tasks which require limited production.

In short, the extant literature indicates the need for closer analysis of the features of integrated summary writing and the need to expand integrated rating criteria, especially those

used in machine-scored tests to better reflect the integrated writing construct. This study aims to address these gaps by addressing the following research questions:

1. How do expert raters characterise the broad construct of integrated summary writing tasks?
2. What are the distinguishing features of effective summary writing elicited through integrated reading-writing (R-W) and listening-writing(L-W) tasks from the perspective of expert judgement?
3. Are there differences in test-takers' responses to integrated summary writing tasks across levels of proficiency?

Method

Using a mixed methods approach of expert judgement, text analysis and statistical analysis, the study examined features of summary writing elicited by two integrated task formats - reading-writing (R-W) and listening-writing (L-W) - and investigated the extent to which these features discriminated performances between levels of proficiency.

Test performance corpus

The current study examined a subset of the Pearson Test of English Academic (PTEA) writing responses produced by test-takers during the administration of one version of the Summarize Written Text task and the Summarize Spoken Text task in 2014 and 2015 (for an example of tasks, see https://pearsonpte.com/the-test/format/). The corpus of test performances was composed of a stratified random sample that included a balanced number of responses from the two integrated tasks that were scored at mid-points of five levels of performance (CEFR A2 to C2). In total, 150 samples of each task were used.

In the PTEA R-W summary task, candidates are required to read a passage (up to 300 words) and summarise it in one sentence (no more than 75 words) in ten minutes. In the L-W summary task, candidates are required to listen to a recording of a lecture (60-90 seconds) and then write a 50-70 word summary in ten minutes. Candidates are allowed to take notes.

Both tasks are machine scored using a partial credit approach[1] in relation to content, grammar, spelling, vocabulary and form (PTE, 2019). An audience is specified only in the L-W task in the written instruction: "You will hear a short lecture. Write a summary for a fellow student who was not present at the lecture".

Expert judgement of salient features

An expert panel was convened to identify the features of summary writing elicited through the integrated R-W and L-W tasks and to provide insights into the nature of effective summary writing. The expert panel consisted of four panellists: two language assessment experts who were experienced in developing and researching summary tasks and two experienced English for Academic Purposes (EAP) teachers who regularly designed and used summary tasks in their teaching contexts (see Table 1).

Table 1. Panel members' profile

Panel member ref	Gender	Expertise	Years of experience
R1	F	Language testing, academic literacy	> 30 years
R2	F	EAP teaching	> 10 years
R3	M	EFL teaching, language testing	> 30 years
R4	F	EAP teaching, academic literacy	> 15 years

EFL (English as a Foreign Language); EAP (English for Academic Purposes)

Prior to the panel discussion, the four panellists were given the prompt, input material and 10 sample responses to the R-W and L-W tasks. The samples included two responses from candidates at each of the five overall writing levels, however the panellists were not informed of the assigned levels. The panellists were asked to rank the 10 R-W and 10 L-W responses from most effective to least effective, making notes on features of each summary that influenced their decision. They brought their notes to the panel discussion. The researchers assumed the role of facilitators and the panellists were asked to discuss their

[1] Item types are scored as correct, partially correct or incorrect. If responses to these items are correct, the maximum score points available for each item type will be received, but if they are partly correct, some score points will be given, but less than the maximum available for the item type. If responses are incorrect, no score points will be received. For example, for the L-W task, "content" is scored as 2 (Provides a good summary of the text. All relevant aspects are mentioned), 1 (Provides a fair summary of the text, but one or two aspects are missing) or 0 (Omits or misrepresents the main aspects) (Pearson, 2018).

rankings for each task (see Supplemental Appendix 1 for discussion questions). This arrangement enabled panellists to clearly articulate the reason for their rankings, engage with other perspectives and justify their rankings by identifying particular features in the R-W and L-W summaries. A shared understanding of what constitutes an effective response to the two integrated summary tasks emerged from this discussion (RQ1).

In order to document a profile of distinguishing features noted on the 10 R-W and 10 L-W samples, the rankings of each response, written notes provided by the panellists and their verbal comments during the panel discussion were collated. The transcript of the panel discussion was analysed through both deductive and inductive approaches (Yin, 2011). Three overarching themes were identified: features of summary, panellists' perceptions of the nature of summary and the rating process. To answer RQ2, we focused on the 317 segments (out of 453 segments) coded under features of summary. We also drew upon panellists' perceptions of the nature of summary to illustrate the findings; however, segments concerning the rating process are beyond the scope of the current article. The 317 transcript segments of summary features were coded by one of the authors using Nvivo 11. A total of 14 features were identified: 1) *academic style*, 2) *accuracy of information*, 3) *coherence and cohesion*, 4) *comprehensibility of response*, 5) *comprehension of source text*, 6) *comprehensiveness of main ideas*, 7) *defensibility of inferences*, 8) *identification of main points*, 9) *inclusion of irrelevant information*, 10) *keyboard skills*, 11) *language control*, 12) *paraphrasing skills*, 13) *staging of information*, and 14) *understanding of task requirements*. For a description of each feature with examples of panellist's comments, see Table 4 in Results. Twenty percent of the segments (i.e., 64 segments) were double coded by the other author. The agreement rate was 95.1%, with discrepancies noted, discussed and resolved (for a summary of the process, see Figure 1).

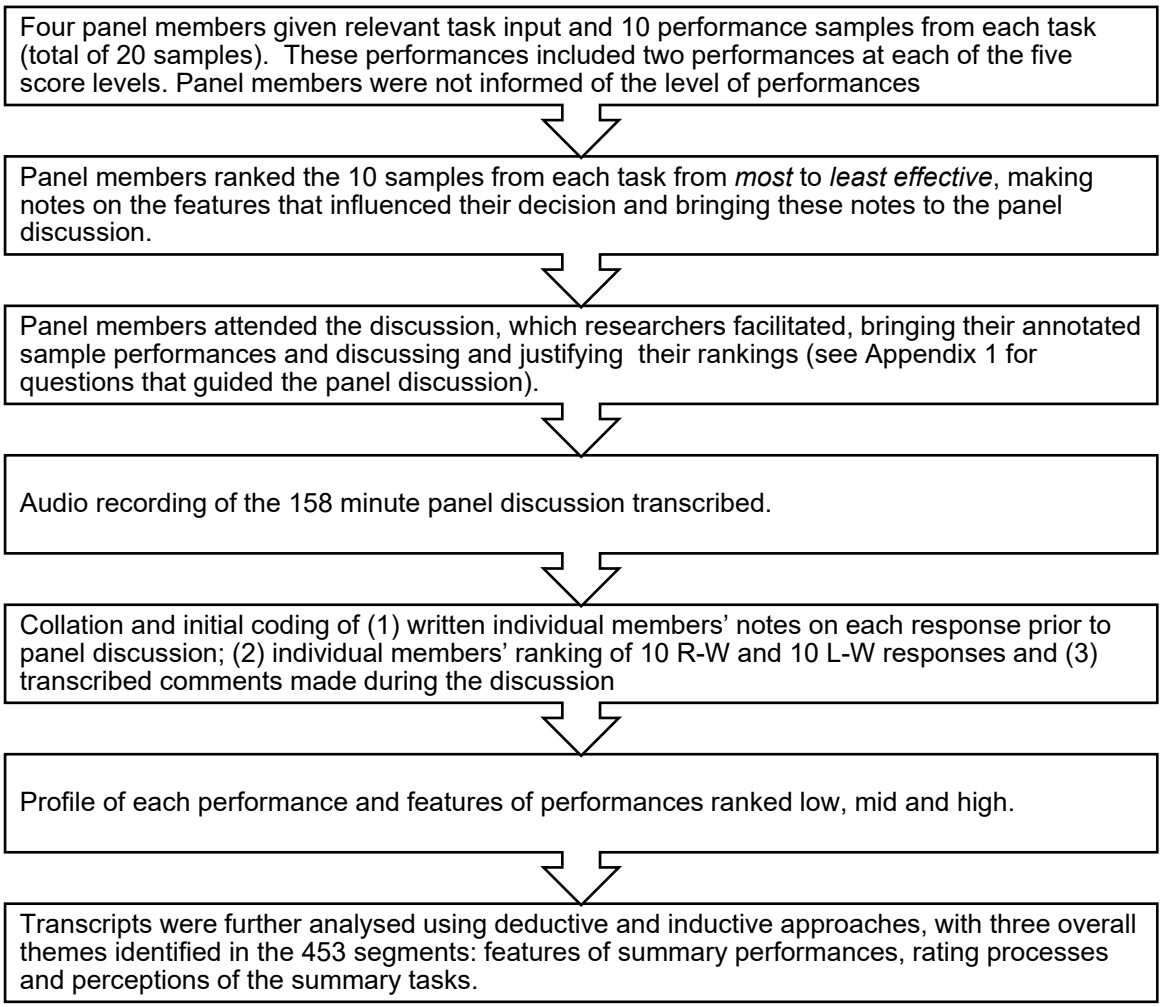

Figure 1. Process for eliciting expert panel discussion and analysis of the panel data

Features of summary writing in candidates' responses across levels

In order to examine features of effective summary writing in the full set of 150 R-W and 150 L-W responses, a revised coding scheme was developed. All categories were informed by the features noted in the panel discussion as well as aforementioned research (e.g., Cho & Choi, 2018; Ohta et al. 2018; Plakans & Gebril, 2013). Several features noted by the panellists were merged together (see Table 2). For example, the panellists made separate comments about the extent to which factual information in the source input was conveyed accurately (*accuracy of information*) and the extent to which they felt the candidate understood the input (*comprehension of source text*). However, as both features reflect a candidate's ability to comprehend the input and then accurately relay the information in the summary, it was not

practical to develop two sets of performance descriptors. These two categories were merged into *accuracy of source information* in the revised coding scheme for RQ3.

Table 2. Revised coding scheme based on features noted by expert judgement

Features noted in panel discussion	Revised coding categories for RQ2
1. Comprehensiveness of main ideas from source text	1. Relevance of source ideas
2. Identification of main points, key ideas, topics	
3. Inclusion of irrelevant information, details	
4. Accuracy of information	2. Accuracy of source information
5. Comprehension of source text	
6. Paraphrasing skills	3. Paraphrasing skills
7. Academic style	4. Academic style
8. Comprehensibility of response	5. Language control
9. Language control	
10. Keyboard skills	
11. Coherence and cohesion	6. Coherence and cohesion
12. Staging of information	7. Staging of information
13. Understanding of task requirements (staging for the intended reader)	
14. Understanding of task requirements (length requirement)	8. Form

After refining the coding categories, we developed performance descriptors for each feature using the same 10 R-W and 10 L-W responses given to the panellists. Given the limited length of the responses, descriptors were developed at three levels: effective, limited effectiveness, and ineffective. In order to reflect that some R-W responses were copied from the source text, the R-W descriptors also included a fourth level, which indicated that a feature was 'absent' as there had been a direct copy with no meaningful evidence of the candidate's summary writing. Table 3 presents the final coding scheme for RQ2, with seven categories applied to the R-W task and eight to the L-W task (full descriptors are provided in Supplemental Appendices 2 and 3).

Table 3. Final coding scheme for RQ2

Features	Description	Measurement
Average relevance scores	Extent to which ideas selected from source text are relevant	Rating of relevance of idea at the level of T-units in each response (procedures are explained in detail below)

Features	Description	Measurement
Accuracy of source information	Extent to which factual information from source text is conveyed accurately or misconstrued	Rating of effectiveness at the level of the entire response (for specific descriptors, see Supplemental Appendices 2 and 3)
Paraphrasing skills	Extent to which paraphrasing is precise	
Academic style	Extent to which the response conforms to features of academic writing	
Language control	Extent of control of linguistic aspects	
Coherence and cohesion	Extent to which the response is coherent. Includes flow, elegance, use of conjunctions, logic of organisation at the phrase/clause level	
Staging of information (L-W task only)	Extent to which the response is staged for the intended audience	
Form	Extent to which the response conforms to the task requirement (no. of sentence/words)	

The analysis of *relevance of ideas* involved four steps:

- Identifying and rating idea units in the source text
- segmenting T-units in candidates' responses
- rating relevance of ideas in each T-unit
- calculating average ratings in each response

We followed Kroll's (1977) definition that an idea unit is "a chunk of information which is viewed by the writer cohesively as it is given a surface form" (p. 85). As explained by Kroll, this is more of a psychological classification focusing on the communicative nature rather than grammatical structure of writing. The two authors independently identified the idea units in the input texts and then discussed to finalise the lists. Eighteen idea units were identified in the R-W input and 21 in the L-W input. According to their relevance to the summary task, each idea unit was classified as main idea, supporting information, specific detail or irrelevant information. This is a process adopted from Spivey (1984) and Plakans and Gebril (2013).

The next step was to divide the 150 R-W and 150 L-W responses into T-units i.e., the shortest unit that can stand alone as a sentence (Hunt, 1965). A T-unit can be an independent clause (e.g., *the speaker talked about a ill woman*) or an independent clause plus its dependent clause(s) (e.g., *Marry Mellen immigrated in 1883 to America, working as a cook for the elite, before contracting a Tyfoid disease*). While we followed Hunt's (1965) definition, T-unit marking was challenging at times, especially when segmenting responses by lower proficiency candidates. Their responses tended to include ungrammatical sentence structures. When there was ambiguity due to grammatical errors, as shown in the example below, the structure was counted as one T-unit instead of two or more.

She was very studing just her weight at least did be 107 (L-W, Ref23, A2)

We standardised the analysis of *relevance of ideas* by number of T-units instead of number of sentences. This was considered more appropriate, given that the R-W task required candidates to produce one sentence. In addition, as candidates summarised from spoken input in the L-W task, their responses tended to resemble the spoken structure which often included several embedded structures in a sentence. To establish reliability of T-unit marking, the first author and a research assistant segmented T-units of 30 responses (10% of the full data set). The marking of T-units between the two had an agreement of 0.93%. After this check, the first author segmented the remaining T-units.

After T-units had been segmented, we rated the *relevance of ideas* in each T-unit as 4 (main ideas), 3 (supporting information), 2 (specific details), 1 (irrelevant information) or 0 (not from the source text). An average rating of all T-units in each response was then computed. In some cases, when a T-unit contained more than one idea unit (IU), an average rating was given based on the composition of the idea units included. For example, the T-unit in the example below consisted of main ideas with some specific details, a rating of 4 was given.

> Marry Mellen immigrated in 1883 to America [IU1; Main idea; IU13; Specific details], working as a cook for the elite [IU2; Main idea], before contracting a Tyfoid disease [IU3; Main idea]

Using an average rating helps to take account of the relevance of all idea units included in a T-unit and a response but the limitation is that the average relevance value may not refer to idea units of the same value in the original source text. However, following Plakans and Gebril (2013), this approach gives a reasonable representation of candidates' ability to include relevant ideas in a summary task.

For the remaining seven features - *accuracy of source information*, *paraphrasing skills*, *academic style*, *language control*, *coherence and cohesion*, *staging of information*, and *form* (see Table 3), an effectiveness rating was given at the level of the entire response using a three-point scale for L-W and a four-point scale for R-W (see Supplemental Appendices 2 and 3 for the full schemes). For example, for *paraphrasing skills* a response could be rated as:

3 – Effective: paraphrasing is precise (using appropriate words) and concise;
2 – Limited effectiveness: paraphrasing is generally precise and concise with some ineffective use of alternative wording;
1 – Ineffective: ineffective use of alternative wording; or, for R-W tasks
0 - Absent: direct copy where there is no meaningful evidence of paraphrasing

For R-W, if a response consisted of more than 80%[2] verbatim, it was coded zero for all categories except *form*. This did not apply to the L-W responses because candidates did not have access to the written transcript. Some candidates might have noted down parts of the audio input but this was a process different from verbatim.

[2] When considering the threshold of direct copy, the authors trialed the percentages of 50%, 60% and 80%. As the R-W task was a rather short task, the summary by nature would include many keywords from the input. It was evident that when the threshold was set at 50%, most responses would have been regarded direct copy. It was a judgment call that when the ratio of direct copy is beyond 80%, there was insufficient evidence of own writing. Nevertheless, the threshold set in this study may only apply to this type of short R-W task.

Staging of information (i.e., the extent to which the response indicated a context for the intended audience) was analysed for the L-W task only. In the L-W task, candidates were asked to write to a peer who missed the lecture, whereas the R-W task under investigation did not explicitly provide an intended audience. We did not analyse staging of information for the R-W task as it was beyond the task requirement.

Having finalised the coding schemes, the authors each rated three or four features in the remaining 140 R-W and L-W responses. To establish reliability, the authors coded 20% of the other's coding. The agreement rate was 100%[3] on the R-W task and 90.2% on the L-W task features. All discrepancies were between one point of the scale. The discrepancies were discussed and agreed.

Statistical analyses

For RQ1 (features of effective summary writing from the perspective of expert judgement), frequency counts of the feature categories commented on the R-W and L-W tasks were calculated. For RQ2 (features of summary writing in candidates' responses across levels), descriptive statistics of the different features identified in R-W and L-W responses were calculated. One-way ANOVA was performed with the different features as dependent variables and the five levels of performance as the between subjects independent variables. Most assumptions were met. However, there were extreme outliers (more than 10 out of 150) in two categories including *language use* and *staging of information* for the L-W task. We included these outliers in the final analyses as removing them did not change the conclusions reached. Secondly, the outliers helped us to interpret the potential effectiveness of the feature categories. The data for the L-W task were not normally distributed. We compared the results using a non-parametric test but this did not change the results. As One-way ANOVA can be considered robust to non-normality (Maxwell & Delaney, 2004) when the sample sizes (numbers in each group) are equal, or nearly equal (Lix et al., 1996), we decided to report the

[3] The complete agreement of the R-W rating was most likely due to fact that the authors developed the coding scheme together and had several rounds of initial coding. Also, most R-W responses were only one-sentence long.

results. Nevertheless, the results need to be interpreted with caution. Correlational analysis was conducted to examine the relations between frequency of the features of effective summary writing and the five proficiency levels.

Results and Discussion

RQ1 - The nature of summaries from written and spoken input

RQ1 aimed to identify features of effective summary writing elicited through the R-W and L-W summary tasks from the perspective of expert panellists. To provide context of the findings, we first report panellists' perceptions of the nature of summary writing and the impact of different modalities on their ranking of responses. Next, we provide an overview of distinguishable features across low, mid and top ranked responses, followed by frequencies of these feature categories coded in the panel discussion.

The nature of a summary speaks to the heart of construct definition. The impact of the modality involved in summarisation and features of the source input emerged as key considerations in the expert panel. Panellists noted the ephemeral nature of the spoken input in the L-W task, which are likely to require different processes to make meaning and retain key points, in comparison with the reading input in the R-W task, which was available for candidates to use while engaging with the task. Panellist 4 explained her perception of the essential differences in the nature of summary between the modalities:

> *What they have in common is that in both cases … you are asked to understand the original and in both cases you have to make decisions about what parts of it you need to relay, but the important difference is that in the listening you have to capture as much as you can in real time in order to be able to make those decisions and that's*

> *something very demanding and something that makes an important difference, whereas with the reading within the 10 minutes you can reread that short text you can pick over it quite carefully …check your understanding … so it seems to me … they are two rather different kinds of summarising… in the reading it's got much more to do with making it shorter whereas in the listening it has got much more to do with passing on the essential meaning as completely and as accurately as you can.* [Extract 1, P4]

It is possible that the L-W task orients more towards retelling than summarising. A consequence of these differences is that panellists expected and valued different features in the two tasks, with Panellist 3 explaining:

> *To me, reading, because they have the text in front of them, would allow less flexibility on interpreting or misinterpreting or misunderstanding whereas due to listening being just heard once so I would focus on more general understanding and a bonus would be a good organization of the narrative but … that's only if their language skills allow this.* [Extract 2, P3]

The role of note-taking while listening was another discussion point in relation to the construct of integrated listening-writing tasks. Panellist 4 concurred that *"listening is very different to reading in that you only get one chance to hear so it seems to me that probably a key skill here is note-taking"*. However, other panellists felt that this was unrealistic in terms of the amount of information that was conveyed in such a short, spoken text. Panellists also noted that candidates had little context nor idea of organisation of the input before hearing the audio. Lack of contextualisation would add challenges to note-taking. Panellist 3 commented:

I noticed many students for listening it seems that they took notes and they attempted to insert the years which may not be that essential for a summary really ... and I was wondering with them trying to capture that … is it a good way to approach the task or may not be … since

some of them captured the years but did miss the information about what illness it was how it was transmitted and how many people died because of it and things so I thought that maybe the years weren't that essential. [Extract 3, P3]

Panellist 2 questioned the authenticity of note-taking in this context*:*

Even if you're a native speaker if you spend time noting down the detail you are going to lose the overall thread of what is being said and therefore especially if they're only able to listen once then I think you have to prioritise a more holistic impression of it …I think you would find it very difficult to listen and note down these facts and I am not sure that you could continue… if it was part of a lecture and which was say an hour long we couldn't continue at that rate of knots. [Extract 4, P2]

Panellist 3 felt note-taking might explain the frequent errors in accuracy of information in L-W responses. The same descriptors worked effectively on the R-W task as 'absolute accuracy' was achievable for candidates as they had access to the source text as they wrote the summary. This points to the need for further research into the notion of accuracy of information in a summary task which involves listening.

Another consideration is that the features of source input could pose particular challenges for L2 learners, especially in the L-W task, with the difficulty in spelling names and specialised vocabulary identified as an issue by all panellists. Panellist 3 commented:

As I started reading the [L-W] responses here I just made a note 'spelling' so when you are listening obviously you haven't got an opportunity to copy you haven't got an opportunity to view and therefore there is such difference in the spelling of the [proper nouns], so it's just

your overall understanding and it requires on-the-spot processing and then reconstructing that text. [Extract 5, P2]

The panellists in this study tended to be lenient towards spelling mistakes, which partly explains the discrepancy between the panel's ranking of responses and the PTE scores which evaluated spelling for both the L-W and R-W tasks. Another challenge posed by the inclusion of specialist vocabulary was the difficulty of paraphrasing these terms, with Panellist 1 noting of candidates who attempted this:

W*hen we're coming into technical vocabulary or quasi-specialist vocabulary there aren't many options apart from using the original word or phrase. I felt at times there might have been writers who were trying to find a different way of saying something and that impacted negatively on their performance because they should have just used 'heart attack' or whatever the technical term was in the original text*. [Extract 6, P1]

Having reported the panellists' perceptions of the nature of the R-W and L-W tasks, we now present an overview of the distinguishable features across low, mid and top ranked responses.

Distinguishable features across low, mid and top ranked responses

For the R-W summary task, features that panellists commented upon when they ranked a response as low included the verbatim use of all or most of a summary from the source, with one panellist commenting "*it has been lifted entirely, there has been no summarisation*". Problems with language control and comprehensibility were also evident in these responses, as a panellist noted: "*so incomprehensible it was difficult to judge*". In the mid-ranked responses, panellists commented on issues with cohesion and clarity, while noting that some key ideas had been comprehended and communicated. Responses ranked as top were

characterised by "*precision*" and "*good control of the language*", with candidates "*capturing the main points in a way that has similar prominence to the original*".

For the L-W summary task, panellists ranked responses as low when they conveyed "*little evidence of comprehension of the source text*" or were "*incomprehensible*". Responses ranked as mid included those that were "*well-written but got some facts wrong*" and demonstrated "*good content but some distracting [language] errors*". Highly ranked responses were well structured, with a "*good opening*", conveyed "*a good sense of the story and its importance*" and were "*quite sophisticated linguistically*".

At this stage of analysis, it emerged that panellists identified and commented on a range of features which allowed them to rank the ten responses. Nevertheless, they did not always completely agree on the relative ranking of every response as specific features were often valued differently. For example, one rater might prefer a summary which effectively communicated overall gist while another might appreciate the ability to summarise details from a listening input. There were also some discrepancies between the panellists' rankings and the PTEA scores. For example, as mentioned previously, a response with spelling mistakes would have been penalised by the PTEA scoring system whereas the panellists in this study were more lenient towards spelling mistakes if they did not affect meaning. As argued by Gebril and Plakans (2014), while rating integrated responses, raters attend to various aspects of performance quality, for example, evidence of source use and overall communicative effectiveness. The fact that most learners have a jagged summary writing profile points to the need, especially for holistic scoring, to clearly specify criteria features which are most relevant to the task.

RQ2 – distinguishing features of effective summary and frequency

Based on the 317 segments coded as features of the summary from the panel discussion, we now present the results of the 14 features that emerged in more depth (see Table 4). Although we discuss the comments in terms of frequency counts, it is important to

note that while a feature may be less frequently commented by the panellists, it does not necessarily mean that the feature is less important than the others.

Table 4. Features noted by panellists and number of segments coded.

Feature	**Description**	**R-W** No of segments	**L-W** No of segments	**General** No of segments	**Panel ref no**
Academic Style	Extent to which writing could be part of an assignment/task in an academic context	5	1	0	P1, P2, P3, P4
Accuracy of Information	Factual information is conveyed accurately or misconstrued	17	17	1	P1, P2, P3, P4
Coherence and Cohesion	Extent to which the response was coherent. Includes choice of conjunctions, linking and mentions of flow, elegance	14	4	2	P1, P2, P3, P4
Comprehensibility	Extent to which the response was comprehensible to the reader. Includes clarity, ease of reading	11	8	1	P1, P2, P3, P4
Comprehension of the source text	Extent to which the judge felt the source text was comprehended	9	5	2	P1, P2, P3, P4
Comprehensiveness	Comment on the (in)comprehensiveness of the selection of ideas. Includes thoroughness	2	11	0	P1, P2, P3, P4
Defensibility of Inferences	Extent to which the same inference could be made by others	0	5	1	P1,P2,P4
Inclusion of irrelevant information, details	Comment on inclusion of irrelevant information, questionable relevance, focus on relatively unimportant details	5	5	2	P1,P3,P4
Keyboard skills	Control of/lack of keyboarding skills	6	0	0	P1,P2,P4
Language control	Control of/lack of linguistic aspects. Includes lexis, syntax, punctuation, spelling	24	17	2	P1, P2, P3, P4
Identification of main points, key ideas, topics	Extent to which main points/key ideas/topic are identified and communicated	26	26	1	P1, P2, P3, P4
Paraphrasing skills	Comments on paraphrasing skills, includes precision and conciseness	45	15	5	P1, P2, P3, P4

Staging of information	Comments on organisation of summary as a whole, ways that information is structured within the summary	2	9	0	P1, P2, P3, P4
Understands task requirements	Orients to stated audience in the L-W task; responds in1 sentence response in the R-W task	8	3	0	P1, P2, P3, P4
	Total no of segments	174	126	17	

In general, panellists made more comments on the R-W (174 comments) than the L-W responses (126 comments) during the panel discussion. This was partly because there were more variations in how candidates constructed to one-sentence R-W summary. The panellists were more agreeable on their rankings of the L-W responses. The order of the discussion could be another reason why the R-W responses were discussed more extensively as R-W was discussed first, followed by L-W during the lengthy (over two hours) panel discussion.

For the 14 features of effective summary identified, five features, including *identification of main ideas, inclusion of irrelevant information, accuracy of source information, language control* and *comprehensibility of response* were discussed to a similar extent between the two tasks. It is noticeable that concept-related features were commented on most extensively during the discussion, showing the importance of relevant ideas in summary writing with limited production is aligned with the importance of relevant ideas found in other types of integrated writing tasks (Chan et al., 2015; Plakans & Gebril, 2013; Ohta et al., 2018). However, these findings for shorter summary tasks would not necessarily generalise to integrated writing tasks involving longer texts.

Shin and Ewert (2015) argue that these concept-related features reveal the extent to which learners recognize viewpoints conveyed in the input and display the degree of text engagement. Similar to previous research (Chan et al., 2015; Plakans et al., 2019), panellists

in this study also identified language use and comprehensibility of response as important features which discriminate high-, mid- and low-responses on both tasks. However, it is important to differentiate two aspects of language use in integrated tasks: one relates to the extent to which language of the sources was appropriately summarised and the other concerns general language complexity and accuracy, a distinction also advocated by Knoch and Sitajalabhorn (2013).

Certain features were discussed more frequently in either one of the tasks. For example, the panellists found features including *comprehension of input, paraphrasing skills, coherence and cohesion,* and *academic style* more dominant in the R-W task. Comprehension of input has always been regarded to be an important aspect of integrated writing as good comprehension facilitates the process and outcome of integration (Feak & Dobson, 1996; Plakans, 2009; Plakans & Gebril, 2012). However, as commented by the panellists, good comprehension might not always be "directly observable" as effective paraphrasing skills are required to relay the comprehension which took place. On the other hand, features including *comprehensiveness* and *staging of information* were more evident in the L-W responses. As discussed in the previous section on the nature of summaries, different approaches were used to construct the L-W responses. To summarise the lecture for a peer, some candidates attempted to provide a comprehensive "dictated" version of the lecture whereas others tried to provide the overall gist of the lecture. This distinction will be further discussed in conjunction with the findings of RQ2, i.e., the extent to which these features were evident in candidate's responses.

RQ3 - Features of effective summary in candidate's responses by proficiency group

Having reported the features that emerged from the panel discussion, we now report the findings regarding the extent to which these features were observable in the full set of 150 R-W and 150 L-W responses.

Features of R-W summary by proficiency level

Figure 2 illustrates the mean ratings of the seven features on the R-W responses by proficiency group. It was consistently observed that the higher proficiency groups showed more evidence of these features of effective summary writing than the lower proficiency groups. Nevertheless, the degree of increase across levels was not homogenous. This agrees with the literature that learners' progress in different features is not linear and that some features are better at discriminating texts at certain adjacent levels (e.g., Chan, 2011; Green, 2012).

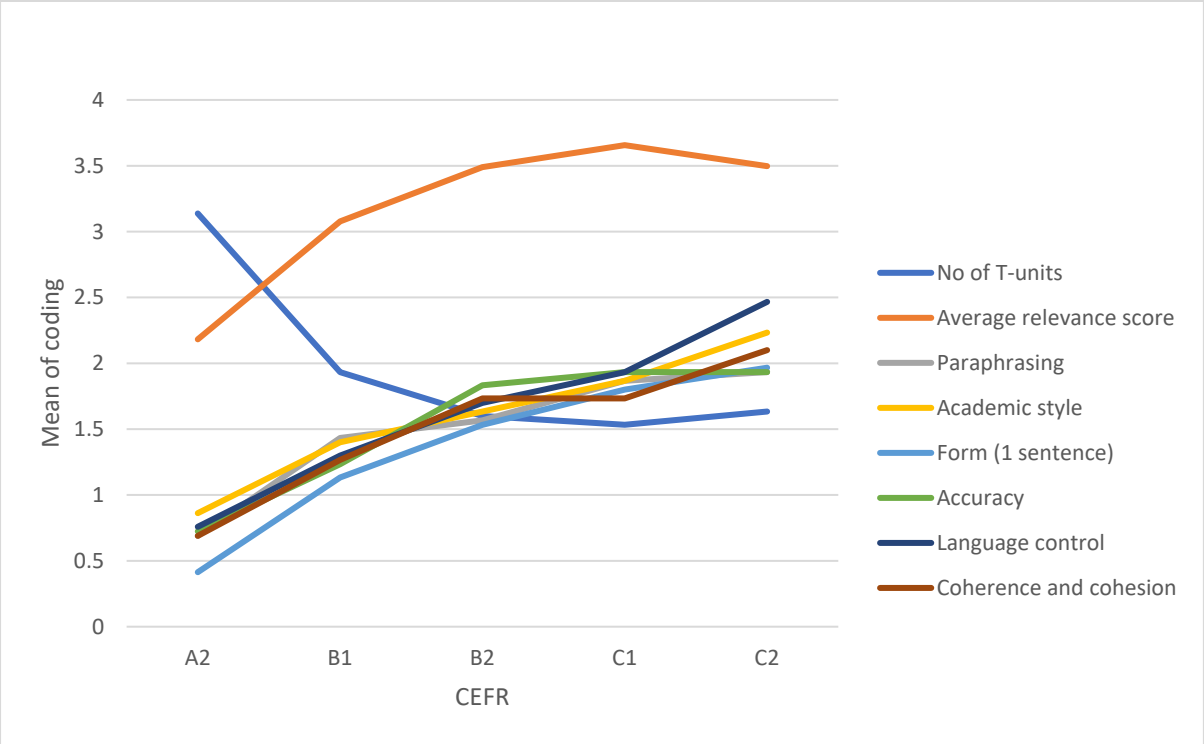

Figure 2. Feature scores by proficiency group on the R-W summary task

Table 5 shows the descriptive statistics for mean ratings of the seven features on the R-W responses by proficiency group. It shows that, overall, the high proficiency groups obtained higher scores on average than did the low proficiency groups on the R-W task. Using a Bonferroni-adjusted alpha of .006, one-way ANOVA detected significant differences by proficiency group for all seven features.

Table 5. Descriptive statistics for feature scores by proficiency group on R-W summary and one-way ANOVA results

		Median	Mean	SD	X^2	F	p
No of T-units	A2	3.00	3.14	1.12	13.20	15.46	0.00
	B1	2.00	1.93	0.83			
	B2	1.00	1.60	0.97			
	C1	1.00	1.53	0.82			
	C2	0.85	1.63	0.85			
Average relevance score	A2	2.33	2.11	1.37	11.85	13.13	0.00
	B1	3.42	3.08	1.00			
	B2	4.00	3.49	0.78			
	C1	4.00	3.66	0.62			
	C2	4.00	3.50	0.81			
Paraphrasing	A2	1.00	0.69	.712	7.24	9.63	0.00
	B1	2.00	1.43	.971			
	B2	2.00	1.57	.679			
	C1	2.00	1.87	1.042			
	C2	2.00	1.93	.868			
Accuracy	A2	0.00	0.72	0.92	8.50	9.20	0.00
	B1	1.00	1.23	1.10			
	B2	2.00	1.83	0.83			
	C1	2.00	1.93	1.05			
	C2	2.00	1.93	0.87			
Academic style	A2	0.00	0.86	0.99	7.80	9.02	0.00
	B1	2.00	1.40	1.10			
	B2	2.00	1.63	0.77			
	C1	2.00	1.87	0.97			
	C2	2.00	2.23	0.77			
Language control	A2	1.00	.73	.785	12.82	16.34	0.00
	B1	1.50	1.30	.952			
	B2	2.00	1.70	.794			
	C1	2.00	1.93	1.048			
	C2	3.00	2.47	.819			
Coherence and Cohesion	A2	1.00	0.69	0.76	8.68	10.92	0.00
	B1	1.00	1.27	0.94			
	B2	2.00	1.73	0.87			
	C1	2.00	1.73	0.94			
	C2	2.00	2.10	0.92			
Form	A2	0.00	0.41	0.78	11.30	10.45	0.00
	B1	1.00	1.13	1.22			
	B2	1.00	1.53	1.07			
	C1	2.00	1.80	1.10			
	C2	2.00	1.97	0.96			

To remind the reader, the R-W task in this study requires candidates to summarize a passage in one sentence of no more than 30 words in ten minutes. As reflected in the rating of form (i.e., task fulfilment of producing the summary in one sentence), lower proficiency candidates were less able to conform to the requirement and received significantly lower ratings than higher proficiency candidates. On the other hand, to meet the task requirement, higher proficiency candidates were more able to produce a one-sentence summary with complex embedded clauses.

As established in the literature, an important part of summary writing involves writers selecting ideas which are relevant to the purpose of writing. As shown in Table 5*,* the higher proficiency groups had significantly higher average *relevance* scores than lower proficiency groups on the R-W. This implies that the ideas selected by the higher proficiency candidates were more relevant than those selected by the lower proficiency candidates. The *paraphrasing* category concerns how well candidates paraphrased these ideas which were selected from the source. The higher proficiency candidates demonstrated significantly more evidence of paraphrasing than the lower proficiency candidates on the R-W task. Echoing Plakans and Gebril's (2013) findings, lower proficiency candidates in this study also depended heavily on the reading input for content and language. In the following examples, the text in bold was lifted from the reading input. The A2 response (Example 1) was largely verbatim whereas the C2 response showed evidence of paraphrasing.

Example 1

> In the researc it saws that **who take regular 30 minitus naps** will help in **heart diseas**, it also benificial for **working men**; also **napping was more likely than diat or physical activities to lower the incidence of heart attacks** and **sleep at any time of the day acts like a valve to release stress of everyday** (R-W, Ref28, A2)

Example 2

> New research has shown that taking **naps** of more than **30 minutes** during the **day** on a **regular** basis has a positive effect on health, possibly **more** so **than diet or** exercise, and can help protect against **heart disease** (R-W, Ref124, C2)

The *accuracy of information* category examines the extent to which factual information from the input was conveyed accurately. Again, the higher proficiency candidates (at B2-C2) had

significantly higher accuracy ratings than the lower proficiency candidates (at A2 and B1). In Example 3, the A2 candidate misconstrued a causal relationship between napping and the incidence of heart attack. According to the input, napping helps to reduce heart attacks instead of causing them. This was accurately conveyed in Example 4 by a C1 candidate.

Example 3

> too much stess is harmful for the people's health, and the cause for the heart attacks is napping (R-W, Ref1, A2)

Example 4

> According to scientific study the chances of heart attack and other heart diseases can be decreas by napping. (R-W, Ref32, C1)

It is also consistent that the higher proficiency candidates had significantly higher ratings on the remaining three categories, including *academic style*, *language control* and *coherence and cohesion*, than the lower proficiency candidates on the R-W. These features relate to general writing skills as they are not exclusive to summary tasks.

Table 6 shows the results of post-hoc analyses of the significance of pairwise comparisons using the Turkey HSD test. The table presents only the comparisons that indicated statistically significant differences. Feature scores in A2 responses in all categories were significantly lower than those at other individual levels. This means that A2 candidates performed significantly worse than learners at other levels in relation to these features on the R-W summary task, indicating that the skill of producing a one-sentence summary of a reading passage is too demanding for A2 learners. Other significant differences were obtained between B1 and C1 and between B1 and C2 in five categories including *accuracy of source information*, *academic style*, *language control*, *coherence and cohesion*, and *form*. This

suggests that B1 might be the level of proficiency where learners develop the ability to convey information acquired from reading.

Table 6. Post hoc pairwise comparisons of features on R-W summary

Dependent Variable	(I) Level	(J) Level	Mean Difference (I-J)	Std. Error	Sig.
No of T-units	A2	B1	1.205*	0.24	0.00
	A2	B2	1.538*	0.24	0.00
	A2	C1	1.605*	0.24	0.00
	A2	C2	1.505*	0.24	0.00
Average relevance	A2	B1	-.96867*	0.25	0.00
	A2	B2	-1.38067*	0.25	0.00
	A2	C1	-1.54733*	0.25	0.00
	A2	C2	-1.38800*	0.25	0.00
Paraphrasing	A2	B1	-.744*	0.23	0.01
	A2	B2	-.877*	0.23	0.00
	A2	C1	-1.177*	0.23	0.00
	A2	C2	-1.244*	0.23	0.00
Accuracy	A2	B2	-1.109*	0.25	0.00
	A2	C1	-1.209*	0.25	0.00
	A2	C2	-1.209*	0.25	0.00
	B1	C1	-.700*	0.25	0.04
	B1	C2	-.700*	0.25	0.04
Academic style	A2	B2	-.771*	0.24	0.02
	A2	C1	-1.005*	0.24	0.00
	A2	C2	-1.371*	0.24	0.00
	B1	C2	-.833*	0.24	0.01
Language control	A2	B2	-.967*	0.23	0.00
	A2	C1	-1.200*	0.23	0.00
	A2	C2	-1.733*	0.23	0.00
	B1	C1	-.633*	0.23	0.05
	B1	C2	-1.167*	0.23	0.00
	B2	C2	-.767*	0.23	0.01
Coherence and Cohesion	A2	B2	-1.044*	0.23	0.00
	A2	C1	-1.044*	0.23	0.00
	A2	C2	-1.410*	0.23	0.00
	B1	C2	-.833*	0.23	0.00
Form	A2	B2	-1.120*	0.27	0.00
	A2	C1	-1.386*	0.27	0.00
	A2	C2	-1.553*	0.27	0.00
	B1	C2	-.833*	0.27	0.02

*The mean difference is significant at the 0.05 level.

Table 7 shows results of the correlations between features scores on R-W and proficiency group. All seven features had a significant ($p<.05$) moderate correlation with proficiency level, ranging from $r=.402$ to $r=.562$. It is worth pointing out that high correlations (r=.8 or above) were observed between three pairs of features including *academic style* and

language control, *academic style* and *coherence and cohesion*, and *language control* and *coherence and cohesion*. This implies that these categories measure in a similar pattern, raising the question of whether they should be combined in operational rating scales. Nevertheless, while candidates in this study might have performed in a similar way on these features, having these features as separate categories would provide richer diagnostic feedback than one combined rating category.

Table 7 Correlations between features and levels (R-W summary task)

		CEFR	No. of T-units	Average relevance score	Paraphrasing	Accuracy	Academic Style	Language Control	Coherence Cohesion	Form
CEFR	*r*	1.000	-.443**	.418**	.423**	.402**	.441**	.562**	.460**	.460**
	p	.	.000	.000	.000	.000	.000	.000	.000	.000
No of T-unit	*r*		1.000	-.631**	-.239**	-.243**	-.186*	-.301**	-.283**	-.468**
	p		.	.000	.003	.003	.024	.000	.000	.000
Average relevance score	*r*			1.000	.189*	.255**	.124	.208*	.173*	.265**
	p			.	.021	.002	.133	.011	.035	.001
Paraphrasing	*r*				1.000	.699**	.727**	.761**	.730**	.674**
	p				.	.000	.000	.000	.000	.000
Accuracy	*r*					1.000	.681**	.683**	.621**	.570**
	p					.	.000	.000	.000	.000
Academic style	*r*						1.000	.833**	.806**	.630**
	p						.	.000	.000	.000
Language Control	*r*							1.000	.859**	.760**
	p							.	.000	.000
Coherence Cohesion	*r*								1.000	.753**
	p								.	.000
Form	*r*									1.000
	p									.

**Correlation is significant at the 0.01 level (2-tailed)
*Correlation is significant at the 0.05 level (2-tailed)

Features of L-W summary in candidate's responses by proficiency level

We now report the results concerning eight features on the L-W responses by proficiency group. To remind the reader, the L-W task requires candidates to listen to a lecture excerpt (60-90 seconds) and produce a 50–70-word summary of the main ideas in ten minutes. As shown in Figure 3, there was not much variation in the rating of the features across proficiency levels. Table 8 shows the descriptive statistics of mean feature scores on the L-W responses by proficiency group and the one-way ANOVA results. Differences in all categories across levels were non-significant. Regarding the correlational analysis, the correlations between feature scores on L-W performance and proficiency group were low. Language control had a significant ($p<0.05$) negative correlation with proficiency level at r=-.166.

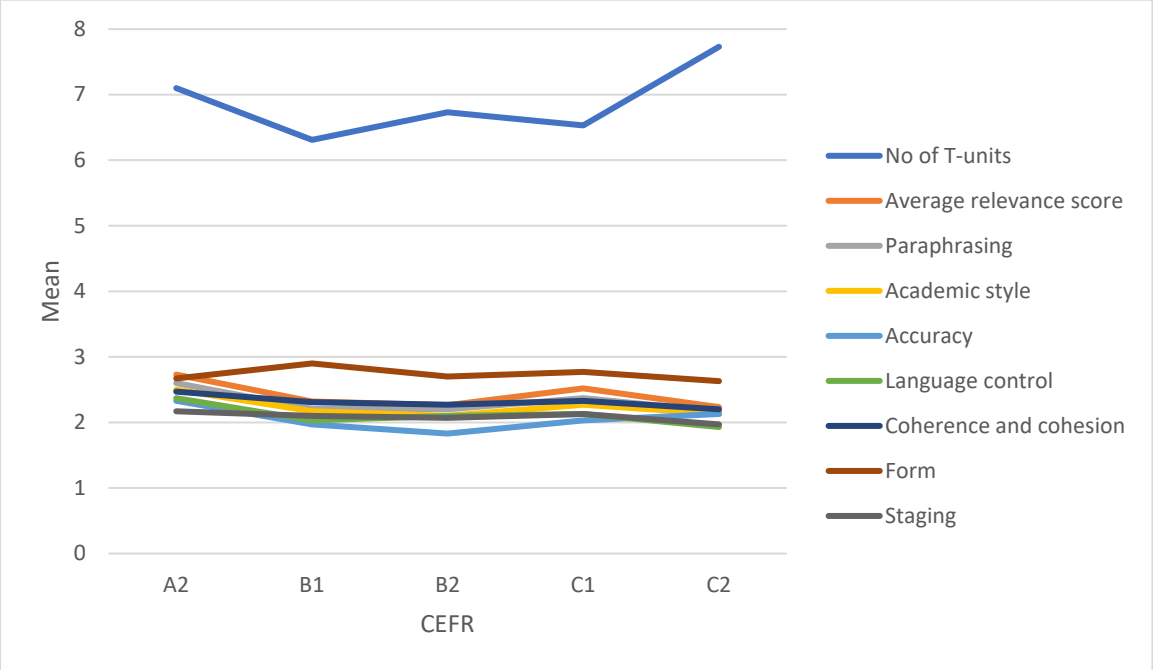

Figure 3. Feature scores by proficiency group on the L-W summary task

Table 8. Descriptive statistics for feature scores by proficiency group on L-W summary and one-way ANOVA results.

		Median	Mean	SD	X2	F	p
No of T-unit	A2	7.00	7.10	1.689	9.238	3.009	.020
	B1	6.00	6.31	1.466			
	B2	6.00	6.73	1.780			
	C1	6.50	6.53	1.570			
	C2	7.00	7.73	2.164			
Average Relevance Score	A2	2.85	2.7298	.70979	1.312	1.425	.229
	B1	2.46	2.3200	1.19474			
	B2	2.50	2.2627	1.03821			
	C1	2.57	2.5206	.89710			
	C2	2.35	2.2339	.89884			
Paraphrasing	A2	3.00	2.60	.563	.977	1.930	.113
	B1	2.00	2.21	.819			
	B2	2.00	2.20	.714			
	C1	2.00	2.37	.556			
	C2	2.00	2.17	.874			
Academic Style	A2	3.00	2.50	.630	.777	1.561	.188
	B1	2.00	2.17	.805			
	B2	2.00	2.10	.662			
	C1	2.00	2.27	.691			
	C2	2.00	2.13	.730			
Accuracy	A2	2.00	2.33	.606	1.056	2.296	.062
	B1	2.00	1.97	.731			
	B2	2.00	1.83	.592			
	C1	2.00	2.03	.669			
	C2	2.00	2.13	.776			
Language Control	A2	2.00	2.37	.615	.774	1.798	.132
	B1	2.00	2.03	.778			
	B2	2.00	2.10	.607			
	C1	2.00	2.13	.571			
	C2	2.00	1.93	.691			
Coherence and Cohesion	A2	2.50	2.47	.571	.292	.689	.601
	B1	2.00	2.31	.761			
	B2	2.00	2.27	.640			
	C1	2.00	2.33	.606			
	C2	2.00	2.20	.664			
Staging	A2	2.00	2.17	.461	.177	.621	.649
	B1	2.00	2.10	.673			
	B2	2.00	2.07	.583			
	C1	2.00	2.13	.434			
	C2	2.00	1.97	.490			
Form	A2	3.00	2.67	.547	.318	1.347	.255
	B1	3.00	2.90	.310			
	B2	3.00	2.70	.535			
	C1	3.00	2.77	.430			
	C2	3.00	2.63	.556			

The reasons why these feature categories did not discriminate L-W writing responses across proficiency groups as they did on the R-W task are complicated. As explained in the Results section of the panel discussion (RQ1), while panellists commented on a range of features which helped them to rank L-W performances from weakest to strongest, they noted that the differences between stronger and weaker L-W responses were more subtle as compared to the R-W responses. Panellist 3 commented "*I found the listening [responses] harder to rank. I think that though is because in my perception they [the responses] were more similar… there were more you know finer distinctions*". For example, in terms of form, most candidates were able to keep to the word limit (50-70 words), with some coming in at exactly 70 words. This fulfilled the criteria for 'effective'. In contrast, the length requirement (one-sentence summary) on the R-W task tended to be met by the higher proficiency candidates, with lower proficiency candidates responding in two or more sentences. This may reflect that the language demands of producing a complex sentence were beyond these candidates.

It is also possible that some descriptors were either not granular enough or did not fully reflect the nature of the L-W construct. For example, for *accuracy of source information*, it was rare that all information' relayed in a L-W response was fully accurate as nearly every response contained some factual errors due to the challenging cognitive demand of the task. This meant that a vast majority of the L-W responses were rated as 'limited' for *accuracy of source information*. Another reason why weaker candidates scored comparatively better on this feature is because candidates with weak listening comprehension or limited proficiency tended to omit information from the input. Candidates with medium listening comprehension, on the other hand, tended to pick up more ideas and details from the listening input but this also means that their responses were more likely to contain content-related mistakes and consequently received low ratings. This raises some important questions for test developers. For example, what does *accuracy of source information* tell us about a candidate's L-W

summary ability? What level of accuracy should be considered achievable, especially when the candidates only listen to the input once?

The findings of *relevance of ideas* in the L-W responses, again, indicate the need to rethink the construct of L-W. Unlike the R-W task where higher proficiency candidates showed better ability to select relevant ideas from the reading input (which was available during the R-W task), the L-W task required candidates to listen to the input and make the selection concurrently either by taking notes or holding these ideas in their mind. The listening input was not available to the candidates as they composed the L-W summary. This proved to be very demanding for candidates across all proficiency levels. The lack of schema and contextualisation for what was about to be heard also added to the challenge of selecting relevant ideas during listening. Candidates, as a result, tended to non-selectively include any ideas which they were able to capture during listening. This was found in L-W responses across lower and higher proficiency candidates (for example, see Examples 5 and 6), where T-units are coded as 4 (identifying main ideas), 3 (supporting information), 2 (specific details), 1 (irrelevant information) or 0 (not in the source text).

Example 5

> *This lecture is about a Lady call Mary. (4) She was born in 1869 (2) and was famous for her excellent cooking in NewYork. (4) She lived with friends sometimes. (3) She's generally happy both in work and life. (1) She's ill with a diease with symptoms of fever, headache. (4) This diease can be transmitted by food, water. (4) Strong evidence suggests this fever is serious (1), one of the people died of it. (4)* (L-W, Ref74, A2)

Example 6

> *Mary Allen was an Irish woman who migrated to America. (4) She worked as a cook for elite American families. (4) We don't know much about her personal happiness (1) but she had great pride in her work.(1) In the 1900's she unknowingly contracted Typhoid.*

> *(4) Typhoid causes fever and headaches. (2) It is said that between 1900 and 1907 she infected twenty two people out of which one person died. (4)* (L-W, Ref133, C2)

In addition, to compensate for the gaps in listening comprehension, candidates often improvised their own ideas (which were scored as zero *irrelevant*) on the L-W tasks. Higher-level candidates were more likely to improvise or elaborate on ideas based on what they had understood, but then received low ratings on relevance in general (see Example 7).

Example 7

> *… she completed her studies in the year of 1883. (0) She lived in the newyork city working as cook. (4) Most of the time she spend with her friends.(0) In her personel life she suffered with problems (0) …* (L-W, Ref50, C1)

In most situations, being able to spontaneously elaborate and interpret information is a sign of advanced writing ability but this ability might not be appropriate for L-W tasks, especially when the intended construct of L-W is to produce a brief summary without additional personal interpretation. Candidates are instructed that “your response will be judged on the quality of your writing and on how well your response presents the key points presented in the lecture”. Thus, the expectation of the summary task needs to be communicated very clearly to candidates.

Future studies should also investigate how candidates determine relevance of ideas on L-W tasks because task variables such as that speed and intonation of speech on the recording might influence what information is perceived as importance or relevant. There is limited discussion on appropriate ways of evaluating candidates’ ability to select relevant ideas in summary tasks. A clear construct definition of L-W tasks would be essential to determine how relevance of ideas among others should be awarded. However, more research is needed to

(re)define the intended construct of L-W and to determine best practice approaches to operationalising this construct through L-W tasks, with recent research showing that the various independent and integrated listening tasks all measure the same construct of listening (Wei & Zhang, 2017).

This leads us back to the fundamental question of the nature of summary writing and thus construct definition. Key features of summary writing such as accuracy of information, source use and paraphrasing skills are often predicated on the assumption that candidates have understood enough of the reading and/or listening to be able to use their writing skills to summarise what they have comprehended. However, as indicated by the findings, this assumption should not be taken for granted in test tasks, particularly L-W summaries. The findings suggest that the modality involved in summary writing, features of the source input and the manner in which they are rated impact on the construct of summary writing that is being operationalised in each task.

Conclusions

It is important to note several limitations of the study. As features of candidates' summary responses are largely influenced by specific task features, the findings of this study might not be generalizable for performances elicited by other R-W and L-W summary tasks which require more substantial writing. Modality of integration was only one variable contributing to the differences between the R-W and L-W features discussed. Indeed, other task features such as the purpose and intended audience of the summary play an essential role in determining the nature of the summary elicited. The expert panel consisted of only four members. Although care was taken to include panellists each with different relevant professional backgrounds, e.g., testing, teaching and research, their opinions and analyses might have limited generalizability. Frequency counts of coded panellist comments do not mean that a feature that was mentioned more frequently is necessarily more important. For

the coding scheme used to analyse R-W and L-W responses, we aimed to develop the coding features to be mutually exclusive. Nevertheless, in reality, a weak summary performance often contained irrelevant, inaccurate information which was poorly paraphrased.

The contribution of this study is, nevertheless, unique in that the short length of the written responses pose not only a challenge to candidates, but also to rating summary responses of one sentence or 50-70 words. The findings thus provide insights into rating short performances in relation to skill integration as well as broader implications for integrated task design.

As argued by Ohta et al. (2018), the need for closer analysis of the features of integrated summary writing and for the development of analytic rating criteria that reflect the integrated construct is evident. This study demonstrates the important contribution of expert judgement, provided through the panel discussion, in enabling a deeper understanding of essential features of the integrated writing performances. Through taking into account the rankings of each performance, it was possible to collate a rich profile of discriminative features associated with the R-W and L-W tasks. These features can then inform development of analytic categories and descriptors. This supports the importance of incorporating expert judgment perspectives into rating systems to identify distinguishing features of integrated writing skills involving different modalities.

The features that emerged from expert judgment in this study informed the development of a coding checklist (seven categories on R-W and eight on L-W). Significant differences in all eight features by proficiency group were found on the R-W summary task. It shows that higher proficiency groups performed significantly better than did the lower proficiency groups in relation to these features including *relevance of ideas, paraphrasing skills, accuracy of source information, academic style, language control, coherence*

andcohesion and *form*. All seven features had a significant ($p<0.05$) moderate correlation, ranging from $r=.402$ *to* $r=.562$, with proficiency levels on the R-W task. The findings provide further support that integrated writing performances should be assessed both conceptually and textually (Cumming et al., 2005; Knoch & Sitajalabhorn, 2013). The over-reliance on linguistic aspects in some integrated rating scales may be a threat to construct validity, as it could lead to under-representation of the complex construct of integrated summary writing. Although content-related features are prompt/version dependent, test developers should consider best practice to incorporate relevance of ideas, accuracy of information and paraphrasing skills in the integrated rating scales.

Based on the sample investigated in this study, despite panellists noticing discriminating features while ranking ten L-W responses from strongest to weakest, the coding categories did not statistically discriminate L-W performance across five proficiency levels. This challenges the practice of using the same (or very similar) rating descriptors for evaluating R-W and L-W summary tasks. As revealed by the findings, modality among other task variables has an impact of the nature of summary writing. The assumption that L2 learners (or indeed most L1 writers) can achieve the same level of content accuracy and relevance in their L-W and R-W summaries is problematic. This points to the need to develop specific descriptors for L-W summary responses.

Regarding implications for integrated task design, it is important to clearly specify the genre and textual features of the source input. Task designers should clearly specify the nature of vocabulary used in the source input, as panellists felt that the specialised vocabulary contained in the listening posed a threat to both construct validity and authenticity. Thirdly, there is a need to clearly contextualise the excerpt that is provided as source input, for example, providing a background statement and/or visuals might help activate candidates' schema. The lack of context would lead to confusion of what could/should be inferred from the

text. As argued by Panellist 1 “there are multiple ways of doing summarisation”; thus, there is a need for a more elaborated construct definition for item writers, raters and candidates. We suggest this should incorporate information about the target reader and context for the summary, the medium and nature of the input, an understanding of the impact of key task features on the response (e.g., one sentence for the R-W and single play of input for the L-W) and a shared understanding of what constitutes effective summarisation for a specific task.

Future research should continue to explore the impact of modality on summary writing, for example, by investigating candidates’ performance in summarising the same text between R-W and L-W. In terms of summarising from listening, at present, there is insufficient understanding of how candidates comprehend and select ideas from audio input. Use of retrospective verbal reports to understand the processes that candidates at different levels engaged in L-W tasks would be one promising direction for future research. A unique aspect of this study was the length of the written responses; it would be useful to explore the extent to which summary features identified in this study apply to longer responses, and how length requirement might influence summary performance. To build a more comprehensive construct definition of integrate skills, it would be important to apply the same research approach taken in this paper to spoken summary tasks (i.e., reading-speaking and listening-speaking). Finally, future research should investigate how advances in technology can lead us to develop rating scales which more closely reflect integrated writing skills. It would be important to investigate the extent to which feature categories identified in this study such as relevance of ideas, paraphrasing skills and accuracy of source information could be applied in integrated scoring systems. We also recommend future research to investigate the potential value of the application of these features for providing diagnostic feedback to candidates.

Acknowledgements:

The authors acknowledge and thank Pearson for their funding and support of this project.

References

Bennett, R., & Bejar, I. (1997). *Validity and automated scoring: it's not only the scoring*. (ETS RR-97-13). Princeton, NJ: Educational Testing Service.

Bennett, R., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.). *Technology and testing: improving educational and psychological measurement* (pp. 142–173). New York: Routledge.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 55–67). Routledge/Taylor & Francis Group.

Chan, S. H. C. (2011). Demonstrating cognitive validity and face validity of PTE Academic Writing Items Summarize Written Text and Write Essay. *Pearson PTE Research Notes*, 2011, 1–16.

Chan, S. H. C., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into-writing skill: A case study. *Assessing Writing*, *26*, 20-37. https://doi:10.1016/j.asw.2015.07.004

Cho, Y., & Choi, I. (2018). Writing from sources: Does audience matter? *Assessing Writing*, *37* (1), 25–38. https://doi.org/10.1016/j.asw.2018.03.004

Cumming, A. H., 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 no. MS-22). Princeton, NJ: Educational Testing Service.

Cumming, A., Kantor, R., Eedosy, U., Eouanzoui, K., James, M., & Erdosy, U. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing*, *10*, 5–43.

Feak, C., & Dobson, B. (1996). Building on the impromptu: A source-based writing assessment. *College ESL*, *6*(1), 73–84.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing,* 13(2), 208–238. https://doi.org/10.1177/026553229601300205

Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73. https://doi.org/10.1016/j.asw.2014.03.002

Green, A. (2012). *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range,* English Profile Series 03, Cambridge: Cambridge University Press.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, *56*(4), 473-493.

Hunt, K. (1965). Grammatical structure written at three grade levels. *Research Monograph No. 3*. National Council of Teachers of English.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, *33*(3), 341–366.

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

Knoch, U., & Sitajalabhorn, W. (2013). A closer look at integrated writing tasks: Towards a more focussed definition for assessment purposes. *Assessing Writing*, *18*(4), 300–308. https://doi.org/10.1016/j.asw.2013.09.003

Koskey, K. L. K., & Shermis, M. D. (2013). Scaling and norming for automated essay scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 199–220). New York, NY: Routledge.

Kroll, B. (1977). Combining ideas in written and spoken English: A look at subordination and coordination. In E. Ochs & T. Bennett (Eds.), *Discourse across time and space* (Southern California Occasional Papers in Linguistics, 5). University of Southern California.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Li, J. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing*, 22, 75–90. https://doi.org/10.1016/j.asw.2014.08.003

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*, 579-619.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective (2nd ed.).* New York, NY: Psychology Press.

Ohta, R., Plakans, L., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing*, *38*(2), 1-36.

Pearson PTE. (2019). *Pearson PTE academic score guide (Version 11).* https://www.pearsonpte.com/scoring.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, 13(2), 111–129.

Plakans, L. (2009). Discourse synthesis in integrated second language writing assessment. *Language Testing*, *26*, 561–585. https://doi.org/10.1177/0265532209340192

Plakans, L. (2010). Independent vs . Integrated Writing Tasks: A Comparison of Task Representation. *TESOL Quarterly*, *44*(1), 185–194. https://www.jstor.org/stable/27785076

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing*, *17*(1), 18–34. https://doi.org/10.1016/j.asw.2011.09.002

Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, *22*(3), 217–230. https://doi.org/10.1016/j.jslw.2013.02.003

Plakans, L., & Gebril, A. (2017). Exploring the relationship of organization and connection with scores in integrated writing assessment. *Assessing Writing*, *31*, 98-112. https://doi.org/10.1016/j.asw.2016.08.005

Plakans, L., Gebril, A., & Bilki, Z. (2019). Shaping a score: Complexity, accuracy, and fluency in integrated writing performances. *Language Testing, 36* (2), 161-179. https://doi.org/10.1177/0265532216669537

Pollitt, A., & Taylor, L. (2006). Cognitive psychology and reading assessment. In M. Sainsbury, C. Harrison, & A. Watts (Eds.), *Assessing reading: from theories to classrooms* (pp. 38–49). Cambridge: National Foundation for Educational Assessment.

Sawaki, Y. (2020). Developing Summary Content Scoring Criteria for University L2 Writing Instruction in Japan. In: Ockey, G.J., Green, B.A. (eds) *Another Generation of Fundamental Considerations in Language Assessment.* Springer. https://doi.org/10.1007/978-981-15-8952-2_10

Shin, S.-Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing*, *32*(2), 259–281. https://doi.org/10.1177/0265532214560257

Spivey, N. (1984). *Discourse synthesis: Constructing texts in reading and writing. Outstanding Dissertation Monograph*. International Reading Association.

Wei, W. & Zheng, Y. (2017) An investigation of integrative and independent listening test tasks in a computerised academic English test, *Computer Assisted Language Learning*, *30*(8), 864-883. https://doi.org/10.1080/09588221.2017.1373131

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, *9*(9), 27–55.

Weir, C. J. (2005). *Language Testing and Validation: An Evidence-Based Approach*. Basingstoke: Palgrave Macmillan.
Yin, R.K. (2011).Qualitative Research from Start to Finish. Guildford Press.
Yu, G. (2009). The shifting sands in the effects of sources text summarizability on summary writing. *Assessing Writing*, *14*(2), 116-137.
Yu, G. (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

Acknowledgments

The authors acknowledge and thank Pearson PLC for their funding and support of this project.

Funding

The research reported in the article was supported by Pearson PLC Research Grants.

Appendix 1 Panel discussion questions

Could you please share your overall impression of the features of an effective summary, in terms of the scripts you have ranked?

Could you please share your overall impression of the features of an ineffective summary, in terms of the scripts you have ranked?

Overall, were there noticeable differences in the responses to the 'summarise written text' and 'summarise spoken text' task responses?

Could you please each take us through your ranking of the 10 'summarise written text' responses, sharing the features of the responses that influenced your ranking?

[after each participant has shared their rankings] Is there anything that you would like to add, or has your ranking changed as a result of this discussion?

Could you please each take us through your ranking of the 10 'summarise spoken text' responses, sharing the features of the responses that influenced your ranking?

[after each participant has shared their rankings] Is there anything that you would like to add, or has your ranking changed as a result of this discussion?

At the beginning of the discussion, you identified [summarise the features] as features of effective and ineffective summary writing. Are there any other features that you would like to add, based on the discussion?

Appendix 2 – Coding scheme of features of R-W summary response

Number of T-units: The number of T-units in each candidate's performance.

Average relevance score: each T-unit in students' writing is rated as 0 (not from the source text), 1 (not so important), 2 (specific details), 3 (supporting information) or 4 (main ideas). Average relevance score is the sum of all ratings divided by the number of T-units.

Paraphrasing skills

- Extent to which paraphrasing is effective
  - effective – paraphrasing is precise (using appropriate words) and concise
  - limited effectiveness – paraphrasing is generally precise and concise with some ineffective use of alternative wording
  - ineffective- ineffective use of alternative wording
  - absent- direct copy where there is no meaningful evidence of paraphrasing

Accuracy of source information

- Extent to which factual information is conveyed accurately or misconstrued
  - effective- all information accurately reflects that in the source text
  - limited effectiveness- most information accurately reflects that in the source text, with minor inaccuracies
  - ineffective- major inaccuracies in conveying information from the source text
  - absent- either completely inaccurate/irrelevant or direct copy with no meaningful evidence of comprehension of information

Academic style

- Extent to which the writing conforms to features of genre of academic writing, e.g. nominalization, use of reported speech, hedging, appropriate source attribution etc.
  - effective – the writing shows many academic writing features and attributes to the sources appropriately
  - limited effectiveness – the writing shows some academic writing features but might not attribute to the source
  - ineffective- the writing is mostly absent of academic writing features
  - absent- direct copy with no meaningful evidence of own academic writing features

Language Control

- Control of/lack of control of linguistic aspects, including lexis, syntax, punctuation, spelling
  - Effective - highly accurate, with few or no errors. If there are any minor errors, they do not impact on conveying meaning
  - limited effectiveness - mostly accurate and coveys meaning; however, minor errors might distract the reader
  - ineffective - noticeable major inaccuracies that impact on conveying meaning and distract the reader
  - absent - direct copy where there is no meaningful evidence of own language

Coherence and Cohesion

- Flow, elegance, use of conjunctions, logic of organization at the phrase/clause level
  - effective- a range of conjunctions used accurately, consistently logical flow of ideas at the phrase/clause level
  - limited effectiveness- a limited range of conjunctions used with varying degree of accuracy and/or minor issues with logical flow of ideas at the phrase/clause level

  - ineffective-very limited use of conjunctions that may not be accurate/appropriate and/or major issues with the logical flow of ideas at the phrase/clause level
  - absent – direct copy with no meaningful evidence of (re)organizing the text

Form

- Extent to which the writing conforms to the task requirement
  - effective – 1 sentence with an appropriate and effective sentence structure
  - limited effectiveness – 1 sentence but not fully effective
  - ineffective – 1 sentence but the sentence structure is incorrect or unnatural
  - absent- more than 1 sentence or only a sentence fragment

Appendix 3 - Coding scheme of features of L-W summary response

Number of T-units: The number of T-units in each candidate's performance

Average relevance score: each T-unit= in students' writing is rated as 0 (not from the source text), 1 (not so important), 2 (specific details), 3 (supporting information) or 4 (main ideas). Average relevance score is the sum of all ratings divided by the number of T-unit.

Paraphrasing skills

- Extent to which paraphrasing is effective
  - effective – paraphrasing is precise (using appropriate words) and concise
  - limited effectiveness – paraphrasing is generally precise and concise with some ineffective use of alternative wording
  - ineffective- ineffective use of alternative wording

Accuracy of source information

- Extent to which factual information is conveyed accurately or misconstrued
  - effective- all information accurately reflects that in the source text
  - limited effectiveness- most information accurately reflects that in the source text
  - ineffective- major inaccuracies in conveying information from the source text

Academic style

- Extent to which the writing conforms to features of genre of academic writing, e.g. nominalization, use of reported speech, hedging, appropriate source attribution etc.
  - effective – the writing shows many academic writing features and attributes to the sources appropriately
  - limited effectiveness – the writing shows some academic writing features but might not attribute to the source
  - ineffective- the writing is mostly absent of academic writing features

Language Control

- Control of/lack of control of linguistic aspects, including lexis, syntax, punctuation, spelling
  - effective- highly accurate, with few or no errors. If there are any minor errors, they do not impact on conveying meaning
  - limited effectiveness- mostly accurate and coveys meaning; however, minor errors might distract the reader
  - ineffective- noticeable major inaccuracies that impact on conveying meaning and distract the reader

Coherence and Cohesion

- Flow, elegance, use of conjunctions, logic of organization at the phrase/clause level
  - effective- a range of conjunctions used accurately, consistently logical flow of ideas at the phrase/clause level
  - limited effectiveness- a limited range of conjunctions used with varying degree of accuracy and/or minor issues with logical flow of ideas at the phrase/clause level
  - ineffective-very limited use of conjunctions that may not be accurate/appropriate and/or major issues with the logical flow of ideas at the phrase/clause level

Staging of information (L-W only)

- logic of organisation at the whole response level, inclusion of context for intended audience
  - effective: a logically organised response that orients to the intended audience

    - limited effectiveness: a mostly logically organised response that may or may not orient to the intended audience
    - ineffective: a response that is not logically organised and does not orient to the intended audience

Form

- Extent to which the writing conforms to the task word requirement
    - effective: contains 50- 70 words
    - limited effectiveness: contains 40-49 words or 71-100 words
    - ineffective: contains less than 40 words or more than 100 words.