Approaches to Assessing Summary Content: Implications for Teaching Material Development and Recommendations for Research

Woranon Sitajalabhorn[1]

Abstract

Amongst the different aspects of summary writing, summary content is viewed as one of the primary constructs in measuring the quality of summaries (Alderson, 2000; Hijikata et al., 2015; Putri, 2020; Yamanishi et al., 2019). Although summaries have been defined in a variety of ways by researchers (e.g., Chuenchaichon, 2022; Dewi & Saputra, 2021), authors of books and writing handbooks (e.g., Turabian, 2019; Wette, 2020), as well as online writing resources (e.g., Pressbooks, 2022; Purdue University Online Writing Lab, 2022), the one thing that is held constant across these publications is the idea that summaries must capture all the essential information from the source materials. This mutual agreement only suggests that summary content (or the ideas summary writers include in their summarised texts) must be taken into consideration as far as the assessment of summaries is concerned. Given the importance of summary content in assessing summary quality, this article aims to present how summary content has been evaluated and discuss implications for teaching material development and recommendations for research.

Keywords: Summary; Summarisation; Summary writing; Summary content; Academic writing; Integrated writing; Summary assessment; Integrated writing assessment

[1]Lecturer, Chulalongkorn University Language Institute, Chulalongkorn University, Bangkok, Thailand, 10330

*Corresponding author email: woranon.s@chula.ac.th

Received: 6 January 2023; Revised: 10 February 2023; Accepted: 24 February 2023

Introduction

The ability to summarise well is crucial in an academic context (Chuenchaichon, 2022; Ono, 2021; Putri, 2020; Yamanishi et al., 2019). Students, especially those in higher education, definitely need this skill to complete their degrees (Kirkland & Saunders, 1991) as most class assignments and test tasks at the university level require them to abstract and incorporate essential information from various source texts and materials (Carson, 2001; Marshall, 2017; Ono, 2021; Plakans, 2008). Although summarising enjoys its status as one of the most important writing genres and is used widely as both assigned tasks and assessment tools in educational settings (Hult & Huckin, 2008), it is no mean feat to evaluate this type of writing.

> Evaluating summaries is difficult: Do you give test-takers a certain number of points for targeting the main idea and its supporting ideas? Do you use a full/partial/no-credit point system? Do you give a holistic score?
>
> (Brown & Abeywickrama, 2019, p. 223)

This statement is put by Brown and Abeywickrama (2019) right at the beginning of the discussion about assessing summaries in their book on language assessment. Several other researchers share the same view and further offer different reasons underlying the difficulty in marking summaries, such as the problems with designing and producing effective assessing schemes and scoring criteria (Khvatova & Krutskikh, 2020; Yamanishi et al., 2019; Yu, 2007), achieving rating reliability (Cohen, 1993, 1994; Hijikata et al., 2015; Yu, 2007), and dealing with subjectivity in summary evaluation (Alderson, 1996; Alderson et al., 1995; Weir, 1993). Experts in integrated writing tasks also posit that since summarising requires more than one skill to complete, i.e., listening and writing, reading and writing, or in some cases listening, reading, and writing, its assessment is undoubtedly far more complicated than assessing independent writing or writing-only tasks (Carson, 2001; Hirvela, 2016; Ono, 2021; Putri, 2020; Weigle, 2004; Weigle & Parker, 2012; Yamanishi et al., 2019). In assessing independent compositions, raters may pay attention to certain common marking criteria, such as content clarity, organisation, grammatical accuracy, syntactic and lexical complexity/diversity, cohesion, and coherence. In assessing summaries, however, raters, in addition to the aforementioned criteria, are required to assess additional features exclusive to summary writing, such as conciseness, use of source materials, and paraphrasing (Hijikata et al., 2015; Yamanishi et al., 2019).

As is evident from the discussion above, summary assessment is fraught with challenges, and one of the main challenges originates from the fact that this genre of writing encompasses

various writing constructs that need to be taken into account when it comes to the issue of assessment. Though interesting, it is not within the scope of this article to exhaustively discuss how each aspect of summary writing has been and/or should be assessed. This article will primarily focus on the evaluation of summary content because this aspect is considered by several scholars to be at the heart of summary assessment (Alderson, 2000; Hijikata et al., 2015; Putri, 2020; Yamanishi et al., 2019).

Summary Definitions: Revelation of One of the Main Constructs

Definitions of summaries are ubiquitous. They are offered by various researchers (e.g., Chuenchaichon, 2022; Dewi & Saputra, 2021; Hood, 2008; Kim, 2001; Kirkland & Saunders, 1991; McAnulty, 1981; Ono, 2021; Rinehart & Thomas, 1993; Roig, 2001; Yamanishi et al., 2019), authors of books and writing handbooks (e.g., Davies, 2011; Harris, 2017; Kissner, 2006; Oshima & Hogue, 2006; Swales & Feak, 2004; Turabian, 2019; Wette, 2020), and online writing resources (e.g., Pressbooks, 2022; Purdue University Online Writing Lab, 2022). Each source cited above has its own version of the definition, and this gives rise to some discrepancies. One such discrepancy, for instance, lies in the objective nature of a summary. Whereas Harris (2005) and McAnulty (1981) state that a summary should be written without the exertion of a writer's personal opinions, Swales and Feak (2004) contend that a summary writer can critique the source material or express his/her attitude in a summary. Despite the existence of certain discrepancies, all the aforementioned sources are in consensus that a summary is a condensation of someone else's work which abstracts only the essence of the original.

This mutually agreed definition suggests that the essence of the original text (i.e., main points and key supporting details) is a sine qua non for successful summaries. Thus, whether a summary writer can identify the main points and the key supporting details in the source material and include them in a summary undoubtedly determines how his/her work will be assessed. In other words, summary content or the main points and key supporting details in a source text included in a summary can be considered the main construct as far as the evaluation of summary quality is concerned.

As with other aspects of summary writing, assessing the content of summaries is no less challenging. Rost (1990) argues that assessing summaries is a challenging task not only because summary writers differ in their judgement as to what information should be abstracted but also because they have a wide variety of strategies for presenting their perceived crucial information.

Furthermore, as pointed out by Alderson (2000), raters or language instructors may perceive the importance of each piece of information in a source text differently and may not necessarily agree on which ideas should be included in summaries.

To overcome these challenges, three main approaches have been proposed and employed throughout the years to measure the quality of summary content: rules of summarisation, judgement of informational importance, and rating scales. The details of each approach will be provided in the subsequent sections, and implications for teaching material development and recommendations for research will also be discussed towards the end of the article.

Rules of Summarisation: An Initial Approach to Summary Content Assessment

Research on summary writing began to flourish after the emergence of Kintsch and van Dijk's (1978) model of text comprehension and production. In this model (see also van Dijk, 1979), a summary or a recall is generated through an interaction of three processes: 1). comprehension of the meaning of a text as a coherent whole, 2). condensation of the entire meaning into its gist, and 3). production of a new text. Kintsch and van Dijk's (1978) further propose that the process of summarising entails the schemata of a reader, the microstructure (propositions), and the macrostructure (inductive interpretations) of a text, all of which operate interactively according to a set of macro-rules to reduce and organise the information of a text to its essence. The macro-rules are as follows: 1). *deletion* of unimportant and redundant information, 2). *generalisation* of ideas to generate a superordinate proposition, and 3). *Construction* of a topic sentence. Not only has this processing model become a conceptual framework for several subsequent studies on summary protocols (e.g., Brown & Day, 1983; Brown et al., 1983; Johns, 1985; Johns & Mayes, 1990; Taylor, 1984; Winograd, 1984), but it has also shed light on the issue of summary assessment.

Brown and Day (1983) further expanded Kintsch and van Dijk's (1978) macro-rules to include six basic rules for summary writing: 1). deletion of trivial information, 2). deletion of redundant information, 3). substitution of a superordinate term or event for a list of items or actions (i.e., substituting *pets* for *cats, dogs, and parrots*), 4). substitution of a superordinate action for a list of subcomponents of that action (i.e., substituting *John went to London.* for *John left the house., John went to the train station., and John bought a ticket.*), 5). selection of a topic sentence (if there is one available), and 6). invention of a topic sentence (if none exists). In other subsequent studies, these rules are referred to with different terminology, such as

reproductions, combinations, run-on combinations, and inventions (Winograd, 1984), correct replications and distortions (Corbeil, 2000; Johns, 1985; Johns & Mayes, 1990), and reproduction, transformation, and intrusion (Coffman, 1994).

Aiming to investigate the development of the ability to use summarisation rules by learners at different ages and levels of language proficiency, Brown and Day (1983) conducted a series of three experiments by having fifth, seventh, and tenth graders, college students (novices), and graduate students (experts), all of whom are native speakers of English, write summaries of two expository texts specially designed to elicit the use of summarisation rules by the participants. The results show that all subjects, regardless of their age and language proficiency, could use the deletion rules, the most basic rules, effectively. However, when it comes to more cognitively-challenging rules like substitutions, selection, and invention, tenth graders and college students did better than their younger research cohorts whilst graduate students outperformed the other groups of participants in their ability to use these rules.

Based on these findings, Brown and Day (1983) conclude that "[t]hroughout this series of studies a clear developmental pattern was found, with deletion rules emerging first followed by superordination and then selection. Invention, the most difficult rule, was late developing" (p.12). Similar research results have also been obtained by other studies on first-language summary writing (Brown et al., 1983; Johns, 1985; Taylor, 1984; Winograd, 1984). To the best of my knowledge, the work by Johns and Mayes (1990) is the only piece of research that explored the summarisation rules used by non-native speakers of English to produce summaries, and its results remain much the same.

In light of such findings, some researchers suggest that the quality of summaries can be judged by considering the rules applied to produce them. For example, Brown et al. (1981) argue that the most difficult rule of invention which requires a summary writer to create and incorporate a new piece of information into his/her summarised text is essential for good summarisation. Likewise, Hidi and Anderson (1986) assert that information across sentences or paragraphs from the source material needs to be integrated and combined for summaries to be of high quality. In other words, effective summaries rely upon more sophisticated rules rather than simple copy-delete strategies.

However, assessing summaries merely by considering what rules of summarisation are applied to produce them is inadequate. It might create a misleading impression of the quality of summaries as it is possible that they might still contain a great deal of trivial and redundant

information even should their writers frequently use sophisticated rules. Additionally, using summarisation rules to judge the quality of summaries can negatively affect the practicality aspect of the assessment. It can make the entire marking process a long and painstaking task considering that a rater needs to meticulously identify all the rules used in a summary and assign scores based on the level of sophistication of each rule. Last but not least, the meticulosity of this assessment method can compromise the reliability of the evaluation since raters might disagree on the rule being applied to a particular summarised instance. For example, one rater may consider one instance as the deletion of trivial information whilst another rater might regard the same instance as the deletion of redundant information. The use of summarisation rules, therefore, may not be an appropriate method for summary assessment.

Judging the Importance of Information: A Promising Approach to Summary Content Assessment

As discussed earlier, the word 'summary' has been variously defined, and this results in discrepancies in its definition. Even so, one attribute of summary which holds true across all the sources is that it must capture the gist of the original. This fact, then, raises a question of how the essence of a text can or should be identified. A review of past literature on this issue reveals two main methods employed to identify the gist of a source material.

The Use of Units of Analysis

In order to identify the essence of a text and judge the importance of information, several researchers turn to the concept of idea units (sometimes referred to as propositions, content units, content idea units, linguistic subunits, pause acceptability units, or pausal units). Nevertheless, the methods employed to determine idea units differ from study to study. In Kintsch and van Dijk's (1978) work, for example, an idea unit is the unit consisting of one predicate and its argument(s). Similarly, Coffman's (1994) idea units are "propositions that contained a subject and predicate combination plus restrictive clauses. Compound predicates were divided into separate propositions" (p. 26). Kim (2001) also followed this means of idea unit identification.

Winograd (1984) used punctuated sentences in the original text to segment idea units, arguing that this system "made it possible to identify which ideas from the original passage were included in the summary, as well as to record what transformations had been performed on those ideas" (p. 408). This method of coding idea units was later criticised by Johns (1985)

and Johns and Mayes (1990) in that the use of punctuated sentences to code idea units was difficult, especially if the sentences were composed of two or more clauses or reduced clauses. Instead, Johns (1985) adopted Kroll's (1977, as cited in Johns, 1985) concept of an idea unit (Appendix A) as a way to segment these units. Afterwards, Johns and Mayes (1990) modified Kroll's (1977, as cited in Johns, 1985) original work by integrating ideas from Carrels' (1985) study on text structure and used this modified version to code idea units (Appendix B). In the modified version, several minor items were improved with the consideration of a clause with a compound verb added.

Another way in which idea units are determined is found in Johnson's (1970) work. The researcher asked 23 university students to read a passage and divide it into pause acceptability units. Prior to this activity, the students were briefed on the functions of pausing, namely catching a breath, emphasising the story, or enhancing meaning. According to Johnson (1970), "[t]he validity of a pausal location was accepted when at least one-half of the 23 raters agreed that a pause was acceptable" (p. 13). This approach was also adopted by Brown et al. (1983), and Brown and Smiley (1977).

Once idea units are determined, their structural importance to the theme of the original text is further rated to distinguish the gist from trivia. To illustrate, Brown et al. (1983), Brown and Smiley (1977), and Johnson (1970) had each idea unit retyped on a separate line and instructed a different group of college students to remove a quarter of the idea units that they considered least important to the theme of the passage. Then, they had to repeat the same procedure twice more until there were only a quarter of the idea units left. These remaining units were deemed the most important to the theme of the text and were assigned Importance Level 4 whereas the first set of units that had been eliminated were considered least important and were assigned Importance Level 1. These rated levels of importance are, then, used to develop a scoring scale for summaries.

One obvious benefit of this method is that the information in a source text is divided into clear units of analysis. This, in turn, ensures that each unit will be rated for its importance. A rating scale developed based on this method should greatly facilitate the assessment of summary content because it will be evident whether and to what extent important ideas from the original text are incorporated into a summary. Nevertheless, Brown et al. (1983), Brown and Smiley (1977), and Johnson (1970) failed to explain why there should be four, and not more or fewer, levels of importance. Moreover, their method for judging the importance of information

by eliminating a quarter of idea units three times means that in the end, each level of importance will consist of the same number of idea units (25% at each importance level). This is probably too rigid and may not reflect reality as the number of main points, key supporting details, and trivial or unimportant pieces of information should largely depend on each individual text. In fact, the number of ideas at each importance level should not be predetermined. Whilst some texts might contain more main points than trivial details, others may be composed predominantly of examples and minor pieces of information.

The Use of Native Speakers' or Experts' Judgement

Another alternative is to have native speakers or experts judge the priority of information or write a model summary of the source text. In Cohen's (1993) study, two Hebrew texts and three English texts were used as source materials. Nine Hebrew native speakers and nine English native speakers, all of whom were experts in reading and writing areas, were asked to read and write summaries of the texts composed in their mother tongue. Then, their summaries were analysed, and the ideas agreed upon by five or more experts were used to generate a scoring key for each text.

In much the same vein, Corbeil (2000) also relied on summaries written by native speaker experts to identify the essence of the texts. The researcher had five native speakers of English read and write summaries of two magazine articles written in English and had five native speakers of French perform the same tasks on two French magazine articles. These experts prepared their summaries under the same time and word limit constraints imposed on the participants. The ideas that were included in at least seven or more experts' summaries were regarded as important points and were further used to evaluate the participants' summary content.

Employing a scientific article written in French as a source text, Rivard (2001) had eight teachers, half of whom taught French and the other half taught science, judge the importance of information in the article. Each teacher considered and assigned a score to each of the sentences in the article using the following criteria: a score of 1 indicating that the sentence was important and should be included in a summary; a score of 2 indicating that the sentence was somewhat important; and a score of 3 indicating that the sentence was unimportant and should be excluded from a summary. The researcher, then, reported that five sentences were considered important whilst four sentences were regarded as unimportant, and these nine sentences were used to rate the content of participants' summaries. It is not clear whether

there were discrepancies in the judgement made by the eight teachers and if there were, how the discrepancies were resolved. Also unclear is what happened to sentences assigned a score of 2 because Rivard (2001) did not discuss whether or not such sentences should be included in a summary. What is more, no information about the first language of the eight teachers was provided, so it is not clear whether they are French native speakers or only highly proficient users of French.

Yu (2007, 2008, 2009, 2010) conducted a series of studies on summary writing in which he used extended English texts (two texts in 2007 and three texts in the other studies) as the source materials. All the texts were summarised by five native speakers of English with a good educational background and professional experience. They were not timed whilst producing their summaries but were instructed to write within the word limit of 300–350. Then, their summaries were coded using a computer program before the first 10 most frequently occurring statements were identified and used to construct the rating templates.

Although this approach has led to some criticisms, such as the fact that there can be disagreements among experts as to which ideas are important and should be kept in a summary (Cohen, 1993) and that native speakers do not always perform well on a language test nor do they always perform better than their non-native speaker counterparts (Bachman, 1990), its convenience and ease of implementation still make it appealing (Yu, 2007).

Rating Scales: A State-of-the-art Approach to Summary Content Assessment

In addition to the summarisation rules and the levels of importance discussed earlier, the quality of summary content has also been measured through the use of rating scales. This approach has grown in popularity especially in the past few decades as more and more researchers and writing instructors have developed their own scoring rubrics for summary assessment (e.g., Brown & Abeywickrama, 2019; Baba, 2009; Chen & Su, 2012; Coffman, 1994; Hijikata et al., 2015; Kim, 2001; Kissner, 2006; Rivard, 2001; Yamanishi et al., 2019; Yu, 2007, 2008, 2009, 2010).

It should be noted that these researchers and writing teachers adopted different methods of rating scale development and that their scales included differing assessment criteria. Nevertheless, the content criterion was made manifest in all the scales constructed by the scholars mentioned above. To illustrate, Rivard (2001) only mentions that the scale used in the study was collaboratively constructed by a language curriculum consultant along with

the other two academics. Yu (2007, 2008, 2009, 2010) states that his holistic scale was adapted from those of Rivard (2001) and the ETS *LanguEdge*^{TM} *Courseware Handbook for Scoring Speaking and Writing* (ETS, 2002). Similarly, Baba (2009) simply mentions that the summaries in her study were holistically scored according to a five-point scale taken from ETS (2002). Chen and Su (2012) modified a holistic scoring rubric developed by Jacobs et al. (1981) and used it to evaluate the content, organisation, vocabulary, and language use exhibited in summaries in their research. Kissner (2006), based on her teaching experience and taking writing genres into consideration, constructed what she called summary checklists for expository and narrative texts. Given the five criteria she established (deleting unnecessary and redundant details, replacing a list of items with one term, including main ideas from the original, paraphrasing the author's words accurately, and reflecting the structure of the source text), her checklists correspond closely to Brown and Day's (1983) summarisation rules and are also based largely on her perception of how an effective summary should be. For a narrative text, a criterion of whether a summary includes key story elements (character names, setting, and conflict) is added. The author, moreover, suggested optional criteria, namely word choice, sentence variety, capitalisation, and punctuation.

To focus more on the assessment of summary content through the use of rating scales, this article will discuss the content criterion of the ETS (2002) holistic rating scale and the content criterion in an analytic scoring rubric developed through a series of research projects by Hijikata, Yamanishi, and Ono (i.e., Hijikata et al., 2015; Hijikata, et al., 2011; Yamanishi & Ono, 2018; Yamanishi et al., 2019).

The ETS (2002) rating scale has been used widely for both research and classroom assessment purposes (Yamanishi et al., 2019). It was one of the pilot rating scales examined during the development of the Internet-based version of the Test of English as a Foreign Language or TOEFL iBT. This rubric is a five-point holistic scale with descriptors detailing the quality of summaries at different levels of writing proficiency. Table 1 below shows the ETS (2002) scale's descriptors of summary content at the five band levels (p. 47).

Table 1

The ETS (2002) Scale's Descriptors of Summary Content at the Five Band Levels

Band	Descriptors
5	principal ideas presented accurately with ample and accurately connected key supporting points/elaboration as required to fulfill the task effectively
4	principal ideas presented accurately as required by the task, though one or two key supporting points/details/elaboration may be omitted, misrepresented, or somewhat unclear, inexplicit, or inexplicitly connected
3	principal ideas inconsistently presented: some are discussed accurately with key supporting points/elaboration; other support/elaboration may be absent, incorrect or unclear/obscured by weaknesses in language
2	significantly incomplete, inaccurate, or unclear presentation of principal ideas and key supporting points
1	little or no comprehensible presentation of principal ideas and key supporting points required by the task

As shown in Table 1, summary content is evaluated based on how accurately and completely summary writers can present principal ideas and key supporting points in their summaries. At a glance, the scale demonstrated in Table 1 may seem easy to be utilised as a tool for evaluating summary content due to its simplicity and focus. Note, however, that the actual scale is more complicated than it appears here because each band level also consists of other assessment criteria, such as sentence formation, organisation, word choice, and paraphrasing ability. This means that a rater is required to pay attention to several written features simultaneously. A study by Hijikata et al. (2011) found unsatisfactory reliability (Cronbach's alpha reliability coefficient = .51) amongst the three Japanese raters who used this ETS (2002) holistic rating scale to assess summaries written in English by 51 Japanese university students. The raters reported that it was difficult to assign a score to a summary when two or more written aspects were at different band levels. For example, a summary may exhibit content at Band 4 whereas its sentence formation and word choice may belong to Band 2. Moreover, the raters commented that the assessment results obtained from the use of this scale were only overall scores of the students' performance, and such scores could not be used to provide detailed diagnostic information for the students. Whilst holistic scoring rubrics

usually receive compliments on their ease of use, practicality, and cost-effectiveness (Bacha, 2001; Hamp-Lyons, 1995; Hyland, 2003; Weigle, 2002), Hijikata et al. (2011) did not report whether or not the raters in their study mentioned these advantages.

Acknowledging the lack of effective summary rating scales, Hijikata, Yamanishi, and Ono conducted four research projects (i.e., Hijikata et al., 2015; Hijikata et al., 2011; Yamanishi & Ono, 2018; Yamanishi et al., 2019) to address this issue. Their first research study (Hijikata et al., 2011), as discussed earlier, indicated that the raters could not use the ETS (2002) holistic scale to evaluate summaries reliably and that the rating outcomes were not useful for pedagogical purposes. Based on these findings, they set out to develop a summary scoring rubric in their subsequent studies. In Hijikata et al.'s (2015) work, the researchers designed a provisional analytic scoring rubric with four criteria, including content, quantity of paraphrase, quality of paraphrase, and language use. Yamanishi and Ono (2018) further refined this provisional scale based on the comments and opinions provided by three experts in language testing. At this stage, their rating scale became 'hybrid' as they followed the experts' suggestions and added the overall quality criterion, which requires a rater to assign a holistic score to a piece of summary, to the existing four analytic criteria of content, quantity of paraphrase, quality of paraphrase, and language use. In the final research study conducted by Yamanishi et al. (2019), the hybrid rating scale developed in the previous project was tested both quantitatively and qualitatively for its applicability. The content criterion of this hybrid scoring rubric is displayed in Table 2 below.

Table 2

Yamanishi et al.'s (2019) Scale of the Summary Content Criterion

Dimension	Level		Criteria
CONTENT	4	very good	Can grasp all of the main ideas. Can develop the main point substantially by occasionally using secondary information.
	3	good	Can grasp most of the main ideas. Includes somewhat incorrect information or information beyond the original text, but it does not substantially deviate from the main point.
	2	fair	Can grasp only limited main ideas. Cannot demonstrate an adequate development of the main

Dimension	Level	Criteria
		point. Noticeably includes incorrect information or information beyond the original text.
	1 poor	Cannot identify main ideas. Cannot grasp main ideas correctly.

Yamanishi et al.'s (2019) scoring rubric requires a rater to assess summary content analytically. A summary writer who includes all the main ideas from the original text in his/her summary will receive a score of 4 whereas a summary writer who fails to identify the main ideas will receive a score of 1. Since Yamanishi et al. (2019) did not run separate statistical analyses for the content criterion, the quantitative results regarding the applicability of this criterion, such as inter-rater reliability, were not reported. In terms of qualitative findings, even though the scale does not explain the difference between 'the main ideas' and 'the main point' in the descriptors, the six raters (three native speakers of English and three native speakers of Japanese) who used this scale to mark 16 summaries praised the scale for its ease of use, commenting that the descriptors were clear and distinctive. The researchers may have pointed out the differences between these two terms during the rater training session; hence, this potential confusion could have been eliminated prior to the marking process. Interestingly, one of the raters expressed concern over the fact that this hybrid rating scale seemed to give more scoring weights to paraphrasing ability than to content because summaries were evaluated based on two paraphrasing criteria, i.e., quantity of paraphrase and quality of paraphrase. In this rater's view, content should be the main construct in assessing summary writing.

It is noteworthy that even should the ETS (2002) rating scale and Yamanishi et al.'s (2019) scoring rubric be of different types (holistic for the former and analytic for the latter), their descriptors for summary content are similar in that they target a summary writer's ability to identify the essence of the source material and incorporate it into his/her summary. One concern can be raised about these two summary content evaluation schemes, however. As asserted by Alderson (2000), it is possible for raters to perceive the significance of each idea in an original material differently, and this can lead to disagreements amongst the raters as to which pieces of information should be placed in summaries. For this reason, in addition to training raters how to use a rating scale, the main points and the key supporting details in

a source text should be clearly identified and agreed upon amongst the raters to ensure the validity and reliability of their markings. This process of essence identification can be carried out through the use of units of analysis or the use of native speakers' or experts' judgement discussed in the previous section.

Implications for Teaching Material Development and Recommendations for Research

The discussion above reveals that summary content has been evaluated based on the following three main approaches: rules of summarisation, levels of informational importance (judged by the use of different types of units of analysis and the use of native speakers' or experts' judgement), and rating scales. Whilst the first approach offers implications for instructional material development, the other two approaches suggest research opportunities.

Implications for Teaching Material Development

As discussed previously, developed by Kintsch and van Dijk (1978) and Brown and Day (1983), the six rules of summarisation (i.e., deletion of trivial information, deletion of redundant information, substitution of a superordinate term or event for a list of items or actions, substitution of a superordinate action for a list of subcomponents of that action, selection of a topic sentence, and invention of a topic sentence) were initially used to evaluate the quality of summaries. However, three major problems are likely to arise from the use of these six rules as an assessment tool: score interpretation, practicality, and reliability. For these reasons, assessing summaries based on these rules is deemed inappropriate.

Despite their ineffectiveness as an assessment tool, these six rules of summarisation can be extremely useful if used for instructional material development purposes. It is generally accepted that summarising skills are of paramount importance in education, particularly for students at the tertiary level (Kim, 2001; Yang & Shi, 2003). For university students, the ability to summarise information plays an indispensable role in their academic success (Kirkland & Saunders, 1991) because most of their assignments and assessments require them to abstract and integrate important information from various texts and other materials (Carson, 2001; Marshall, 2017; Ono, 2021; Plakans, 2008). Given such considerable importance, rules or strategies concerning successful summary production should be explicitly taught to students. This statement is well supported by Yaminishi et al.'s (2019) calls for clear instructions for summarising skills, particularly in English as a foreign language (EFL) classrooms.

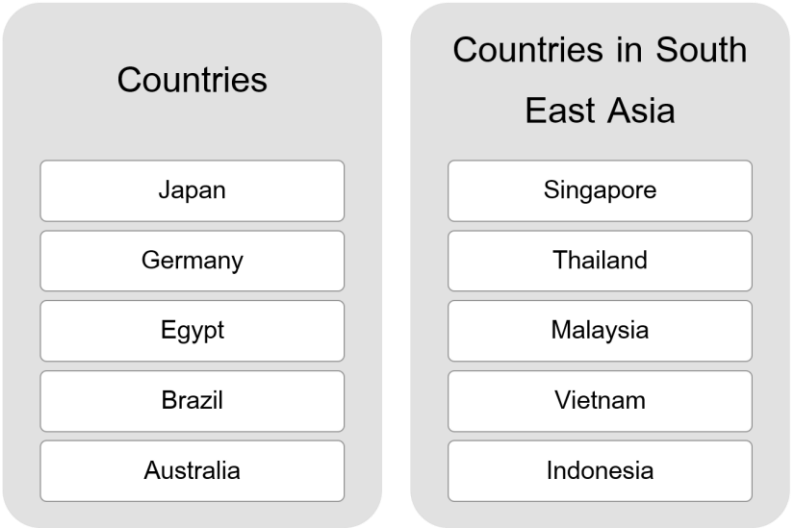
Additionally, Yaminishi et al. (2019) express concern about the restricted number of practical and appropriate instructional materials for summary writing and advocate for the development of teaching guidelines that can facilitate the instruction of summarisation. In most wring handbooks and online writing resources, the crucial topic of summary writing is not discussed as overtly as it should be. For instance, in a chapter or section concerning summary writing, most resources only discuss the significance of this skill, define what a summary is, list characteristics of a good summary, and provide a few examples of original texts versus their summarised versions before ending the chapter or section with summarising exercises as if the learners would master the skill just by reading this information. Hardly are the strategies and processes of summary writing explicitly discussed in these resources. This situation creates a gap as the learners using these learning tools will know what a summary is but are not informed of how to produce it.

As they can be regarded as summarising strategies, the six rules of summarisation can help bridge the aforementioned gap. All the rules along with clear explanations and examples of how to apply them to delete or substitute information and select or create a topic sentence should be overtly and comprehensively discussed in textbooks, writing handbooks, and online writing resources. This way, the learners can learn from these resources both what a summary is and how to successfully write one on their own. Furthermore, it will be helpful for teachers who have to teach summary writing because they can also rely on these resources when planning a lesson (Yaminishi et al., 2019). Instead of just discussing the importance, definition, characteristics, and examples of summaries and then asking their students to complete summary writing tasks, teachers should gradually explain each of the rules in detail to ensure that the students can grasp and hone this essential academic skill.

For example, the section in a textbook that teaches how to substitute a superordinate term for a list of items or actions should begin with a discussion of what this substitution rule is (definition), how this rule can help produce summaries (processes), and why students should use the rule (benefits). Then, some simple examples (see Figure 1 below) can be provided to strengthen the students' understanding of the rules.

Figure 1.

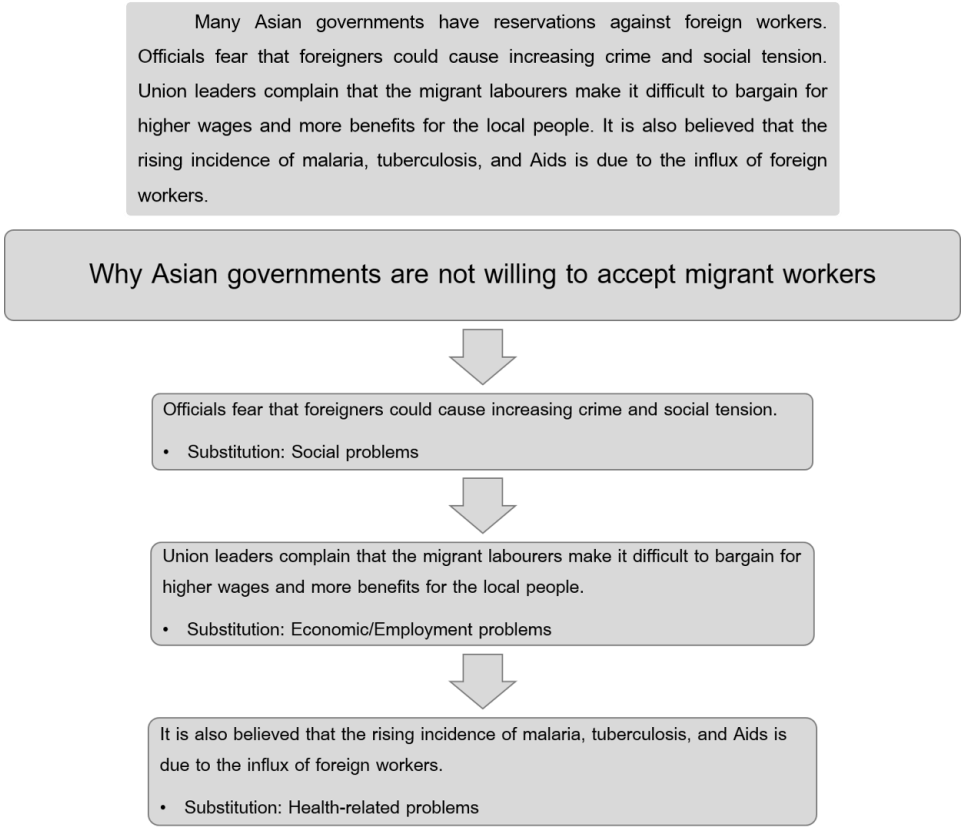
Examples of substitution of a superordinate term for a list of items or actions



Here, the textbook and the teachers using this textbook should explain that whilst 'countries' can substitute both word lists, 'countries in South East Asia' is a more effective substitution for the list on the right because it represents those five countries more accurately. Similar lists of words can be provided further for students to practise using the substitution technique. Next, the textbook can present students with a short paragraph as shown below in Figure 2.

Figure 2

Examples of a short paragraph for students to practise substitution of a superordinate term for a list of items or actions



At this point, the teachers may ask their students to read this short paragraph and discuss the content of the paragraph with them. After the main topic of the paragraph is established (i.e., why Asian governments are not willing to accept migrant workers), the students should be asked to try to replace the content in Sentences 2, 3, and 4 with appropriate noun phrases. Depending on the students' level of proficiency, the teachers might need to adjust the extent to which they facilitate their students. Once the three sentences are appropriately substituted, the teachers may have their students write a complete sentence that summarises this paragraph. At the end, the teachers can provide a sample summary of this paragraph, such as '*Asian governments are unwilling to accept migrant workers due to social, economic, and health-related reasons*.'

Similar materials should be developed to cover the other five summarisation rules. In addition, Ono (2011) found that a lot of EFL students participating in the study learnt how to

summarise on their own, and the researcher also observed that this unfortunate situation was common in the EFL context. Therefore, based on this finding and observation, the samples and exercises provided for each summarisation rule should include both easy and more challenging texts in order to make self-study possible. This will also benefit teachers as they can select the samples and exercises that best suit their students' abilities.

Recommendations for Research

The earlier discussion about how the gist of the original text has been identified shows that there are two common methods used by researchers to achieve this goal: the use of units of analysis (e.g., Brown et al., 1983; Brown & Smiley, 1977; Johnson, 1970) and the use of native speakers' or experts' judgement (e.g., Cohen, 1993; Corbeil, 2000; Rivard, 2001; Yu, 2007, 2008, 2009, 2010). These two methods come with their own pros and cons. Whilst dissecting a source text into smaller units allows for detailed analyses of the level of informational importance of each idea, this method is too rigid because each level of importance will contain the same number (25%) of ideas. This might not always be true in reality, though, as each text can be different in terms of the number of ideas at different importance levels. As for the use of native speakers' or experts' judgement, the method has been praised for being easy and convenient to implement (Yu, 2007), yet native speakers or experts may at times disagree with each other and judge the importance of certain ideas differently (Alderson, 2000; Cohen, 1993).

The fact that these two methods have both advantages and disadvantages makes it interesting to examine which one of them is superior as an assessment tool for summary content. To the best of my knowledge, no research exists that compares the efficacy of these two methods. The assessment of summary content will be made advanced if we know which one of them can better distinguish summaries at different levels of writing proficiency.

Alternatively, the benefits of these two methods can be combined to develop a new approach to identifying the essence of the source material. To illustrate, the information from a source text can be first divided into idea units before being categorised into different levels of importance by native speakers or experts without predetermining the number of ideas in each of the importance levels. Cases of disagreement amongst native speakers or experts, if any, can be resolved through negotiations. This combined approach should then be tested to reveal whether and how effectively it can discriminate between good and poor summaries.

Regarding the use of rating scales to evaluate the quality of summary content, research opportunities abound. To begin with, since rating scales can be constructed based on several

approaches, such as rater intuition (Brindley, 1991; Fulcher, 2003; Hijikata et al., 2015; North, 1995; Yamanishi & Ono, 2018), theories (Knoch, 2011; McNamara, 2002; North, 2003), and empirical evidence (Knoch, 2009), research studies can be done to examine the effectiveness of summary content rating scales developed through these different approaches. The validation of rating scales is another possible area of research. After being developed, a rating scale should be validated to ensure its validity. The validation results can lead to the improvement or the redesign of the scale. Through the validation process, a series of research studies can be conducted. For instance, validating a rating scale based on Bachman and Palmer's (1996) concept of test usefulness requires a researcher to gather evidence for construct validity, reliability, authenticity, interactiveness, impact, and practicality of the rating scale. Similarly, if a scale designer chooses to validate his/her scale using Chapelle et al. (2008) and Bachman and Palmer's (2010) argument-based approach to validation, he/she needs to undertake research to find backings and warrants in order to support his/her entire argument structure before being able to claim the validity of the rating scale. For those who desire to go beyond summary content assessment, research opportunities are also present during the selection of other marking criteria. Aside from content, summaries, like any other type of writing, entail other aspects that need to be assessed: grammatical accuracy, conciseness, structural complexity, lexical sophistication, and paraphrasing ability, to name but a few. It is important to carry out research to investigate which of these aspects of writing should be utilised as summary assessment criteria (e.g., Hijikata et al., 2015). Based on research findings, the aspects that can efficiently differentiate between summaries at different writing proficiency levels should be incorporated into a rating scale.

The Author

Woranon Sitajalabhorn is currently a lecturer at Chulalongkorn University Language Institute. His research interests lie in the areas of writing assessment, assessing languages for specific purposes, and writing rating scale design and validation.

References

Alderson, J. C. (1996). The testing of reading. In C. Nuttall (Ed.), *Teaching reading skills in a foreign language*. London: Heinemann.

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge University Press.

Baba, K. (2009). Aspects of lexical proficiency in writing summaries in a foreign language. *Journal of Second Language Writing, 18*, 191–208.

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

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.

Brindley, G. (1991). Defining language ability: The criteria for criteria. In S. Anivan (Ed.), *Current developments in language testing*. SEAMEO Regional Language Centre.

Brown, D. H., & Abeywickrama, P. (2019). *Language assessment: Principles and classroom practices.* (3rd ed.). Pearson Education ESL.

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher, 10*, 14–21.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarising texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior, 22*, 1–14.

Brown, A. L., Day, J. D., & Jones, R. S. (1983). The development of plans for summarizing texts. *Child Development, 54*, 968–979.

Brown, A. L., & Smiley, S. S. (1977). Rating the importance of structural units of prose passages: A problem of metacognitive development. *Child Development, 48*, 1–8.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. *TESOL Quarterly, 19*(4), 727–752.

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). The University of Michigan Press.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language*. Routledge.

Chen, Y.-S., & Su, S.-W. (2012). A genre-based approach to teaching EFL summary writing. *ELT Journal, 66*(2), 184–192.

Chuenchaichon, Y. (2022). The problems of summary writing encountered by Thai EFL students: A case study of the fourth year English major students at Naresuan University. *English Language Teaching, 15*(6), 15–31. https://doi.org/10.5539/elt.v15n6p15

Coffman, G. A. (1994). The influence of question and story variations on sixth graders' summarisation behaviours. *Reading Research and Instruction, 34*, 19–38.

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). Alexandria, VA: Teachers of English to Speakers of Other Languages.

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). Longman.

Corbeil, G. (2000). Exploring the effects of first- and second-language proficiency on summarizing in French as a second language. *Canadian Journal of Applied Linguistics, 3*(1–2), 35–62.

Davies, M. (2011). *Study skills for international postgraduates*. Palgrave Macmillan.

Dewi, A. K., & Saputra, N. (2021). Problems faced by students in writing English academic summary. *Middle Eastern Journal of Research in Education and Social Sciences, 2*(2), 126–135. https://doi.org/10.47631/mejress.v2i2.257

Educational Testing Service. (2002). *LanguEdge™ courseware handbook for scoring speaking and writing*. Educational Testing Service.

Fulcher, G. (2003). *Testing second language speaking*. Pearson Longman.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*, 759–765. https://doi.org/10.2307/3588173.

Harris, R. A. (2017). *Using sources effectively: Strengthening your writing and avoiding plagiarism* (5th ed.). Routledge.

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

Hijikata, Y., Ono, M., & Yamanishi, H. (2015). Evaluation by native and non-native English teacher-raters of Japanese students' summaries. *English Language Teaching, 8*(7), 1–12. https://doi.org/10.5539/elt.v8n7p1.

Hijikata, Y., Yamanishi, H., & Ono, M. (2011). *The evaluation of L2 summary writing: Reliability of a holistic rubric*. Paper presented at the 10th Symposium on Second Language Writing in 2011. Howard International House.

Hirvela, A. R. (2016). *Connecting reading and writing in second language writing instruction.* (2nd ed.). University of Michigan Press.

Hood, S. (2008). Summary writing in academic contexts: Implicating meaning in processes of change. *Linguistics and Education, 19*, 351–365.

Hult, C. A., & Huckin, T. N. (2008). *The new century handbook*. Pearson Longman.

Hyland, K. (2003). *Second language writing*. Cambridge: Cambridge University Press.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Johns, A.M. (1985). Summary protocols of "underprepared" and "adept" university students: Replications and distortions of the original. *Language Learning, 35*, 495–517.

Johns, A. M., & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics, 11*, 253–271.

Johnson, R. E. (1970). Recall of prose as a function of the structural importance of the linguistic units. *Journal of Verbal Learning and Verbal Behavior, 9*, 12–20.

Khvatova, E., & Krutskikh, E. (2020). Summary writing as a form of integrated skills assessment in tertiary settings. In S. Hidri (Ed.), *Changing language assessment*. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-42269-1_6

Kim, S. (2001). Characteristics of EFL readers' summary writing: A study with Korean university students. *Foreign Language Annals, 34*, 569–581.

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394.

Kirkland, M., & Saunders, M. (1991). Maximising student performance in summary writing: Managing cognitive load. *TESOL Quarterly, 25*(1), 105–121.

Kissner, E. (2006). *Summarizing, paraphrasing, and retelling: Skills for better reading, writing, and test taking*. Heinemann.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating*

scale. 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*, 81–96.

Marshall, S. (2017). *Advance in academic writing: Integrating research, critical thinking, academic reading and writing*. Pearson.

McAnulty, S. J. (1981). Paraphrase, summary, precis: Advantages, definitions, models. *Teaching English in the Two-year College, 8*(1), 47–51.

McNamara, T. (2002). Discourse and assessment. *Annual Review of Applied Linguistics, 22*, 221–242.

North, B. (1995). The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. *System, 23*(4), 445–465.

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

Ono, M. (2011). Japanese and Taiwanese university students' summaries: A comparison of perceptions of summary writing. *Journal of Academic Writing, 1*, 191–205. https://doi.org/10.18552/joaw.v1i1.14

Ono, M. (2021). Japanese university students' integrated writing skills in listening-to-write tasks. *Keio Associated Repository of Academic Resources, 142*, 89–110. Retrieved January 2023, from https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/download.php/AN00062752-00000142-0089.pdf?file_id=159058

Oshima, A., & Hogue, A. (2006). *Writing academic English* (4th ed.). Pearson Longman.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111–129. https://doi.org/10.1016/j.asw.2008.07.001

Pressbooks. (2022). *The roughwriter's guide: Integrating sources*. Retrieved December 2022, from https://pressbooks.pub/roughwritersguide/chapter/quotes-paraphrases-and-summaries/

Purdue University Online Writing Lab. (2022). *Quoting, paraphrasing, and summarizing*. Retrieved December 2022, from https://owl.purdue.edu/owl/research_and_citation/using_research/quoting_paraphrasing_and_summarizing/index.html

Putri, M. N. (2020). Lecturer's perception of using an analytical rubric for assessing summary writing. *Proceedings of the Twelfth Conference on Applied Linguistics (CONAPLIN 2019)*.

https://doi.org/10.2991/assehr.k.200406.025

Rinehart, S. D., & Thomas, K. E. (1993). Summarization ability and text recall by novice studiers. *Reading Research and Instruction, 32*(4), 24–32.

Rivard, L. P. (2001). Summary writing: A multi-grade study of French-immersion and Francophone secondary students. *Language, Culture and Curriculum,14*, 171–186.

Roig, M. (2001). Plagiarism and paraphrasing criteria of college and university professors. *Ethics and Behavior, 11*(3), 307–323.

Rost, M. (1990). *Listening in language learning*. Longman.

Swales, J. M. & Feak, C. B. (2004). *Academic writing for graduate students: Essential tasks and skills* (2nd ed.). University of Michigan Press.

Taylor, K. K. (1984). The different summary skills of inexperienced and professional writers. *Journal of Reading, 27*, 691–699.

Turabian, K. L. (2019). *Students' guide to writing college papers* (5th ed.). The University of Chicago Press.

van Dijk, T. A. (1979). Relevance assignment in discourse comprehension. *Discourse Processes, 2*, 113–126.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing, 9*(1), 27–55. https://doi.org/10.1016/j.asw.2004.01.002

Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing, 21*(2), 118–133. https://doi.org/10.1016/j.jslw.2012.03.004

Weir, C.J. (1993). *Understanding and developing language tests*. Prentice Hall International.

Wette, R. (2020). *Writing using sources for academic purposes: Theory, research and practice*. Routledge.

Winograd, P. N. (1984). Strategic difficulties in summarizing texts. *Reading Research Quarterly, 19*, 404–425.

Yamanishi, H., & Ono, M. (2018). Refining a provisional analytic rubric for L2 summary writing using expert judgment. *Language Education & Technology, 55*, 23–48. https://iss.ndl.go.jp/books/R000000004-I029417776-00.

Yamanishi, H., Ono, M., & Hijikata, Y. (2019). Developing a scoring rubric for L2 summary writing: A hybrid approach combining analytic and holistic assessment. *Language Testing in Asia,*

9(13), 1–22. https://doi.org/10.1186/s40468-019-0087-6

Yang, L., & Shi, L. (2003). Exploring six MBA students' summary writing by introspection. *Journal of English for Academic Purposes, 2*, 165–192.

Yu, G. (2007). Students' voices in the evaluation of their written summaries: Empowerment and democracy for test takers? *Language Testing, 24*, 539–572.

Yu, G. (2008). Reading to summarize in English and Chinese: A tale of two languages? *Language Testing, 25*, 521–551.

Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing, 14*, 116–137.

Yu, G. (2010). Effects of presentation mode and computer familiarity in summarization of extended texts. *Language Assessment Quarterly, 7*(2), 119–136.

Appendix A. Kroll's (1977) concept of an idea unit

1. A noun phrase and verb phrase are counted as one idea unit including (when present) a direct object, a prepositional phrase, adverbial element and a mark of subordination.
2. Full relative clauses are counted as one idea unit when the relative pronoun is present.
3. Phrases which occur in sentence initial position followed by a comma (e.g., participial phrases) or phrases which are set off from the sentence with commas are counted as separate idea units.
4. Reduced clauses in which a subordinator is followed by a non-finite verb (e.g., "as if to destroy the government") are one idea unit.
5. Post-nominal -ing phrases used as modifiers are counted as one idea unit (e.g., Lincoln was left to his thoughts, *worrying).*
6. Other types of elements counted as individual idea units are:
    a. Absolutes: e.g., *His plans thwarted.* Lincoln was discouraged.
    b. Appositives: Lincoln, *the Republican leader,* was able to contact the people.

Source: Kroll, 1977, p. 90, as cited in Johns, 1985, p. 500

Appendix B. Johns and Mayes's (1990) modified version of Kroll's (1977, as cited in Johns, 1985) concept of an idea unit.

1. A main clause is counted as one idea unit including (when present) a direct object, an adverbial element and a mark of subordination.
2. Full relative and adverbial clauses are counted as one idea unit.
3. Phrases, excluding 'transitional' ones, which occur in sentence initial position followed by a comma or phrases which are set off from the sentence with commas are counted as separate idea units.
4. Reduced clauses of various types, including most gerundives and infinitival constructives, are separate idea units.
5. Post-nominal *-ing* phrases used as modifiers are counted as one idea unit (for example, So animals just remain in the water, *dying).*
6. In a clause with a compound verb, the second verb phrase is counted as a separate idea unit. Multiple subjects and multiple direct objects also indicate separate idea units.

7. Other types of elements counted as individual idea units are:

   a. Absolutes: for example, *Its concern heightened,* the government will urge industries to improve.

   b. Appositives: A major type of pollution, *thermal pollution,* is discussed in this article.

Source: Johns & Mayes, 1990, p. 258