Tamara Sladoljev Agejev and Višnja Kabalin Borenić
University of Zagreb, Croatia

ANALYTIC ASSESSMENT OF SUMMARIES IN LSP CLASSES: THE CHALLENGES AND BENEFITS INVOLVED

Abstract

Analytic assessment of discipline-specific summaries is a demanding task that may involve several criteria and more raters. Further impediments to the habitual use of analytic assessment of summaries in ESP classrooms are related to the development of reliable scoring rubrics, inter-rater agreement and objectivity of assessment. Wishing to outline possible solutions to assessment-related difficulties, and to gain more knowledge of our students' summary writing abilities, we analysed summaries of a college-level text according to their content (factual accuracy, completeness, relevance and coherence) and text quality (cohesion and text organization). We discuss the assessment process (involving two raters), students' summary writing performance and implications for teaching. Inter-rater reliability was satisfactory when assessing cohesion, text organisation, completeness and coherence, but was lower for relevance and factual accuracy. Students wrote accurate and fairly relevant summaries but they largely underperformed in coherence, cohesion and text organisation aspects of summary writing.

Keywords: writing assessment, summary writing, analytic assessment, ESP

* E-mail addresses: tsladoljev@gmail.com (T. Sladoljev Agejev), vkabalinb@efzg.hr (V. Kabalin Borenić).

1. Introduction

As a result of an unprecedented flow of rapidly changing data and ideas in the digital age, writing is “the primary conduit for effective transmission of the vital information,” and a “crucial skill across almost all career paths” (Foltz, 2016, p.659). The need to develop students’ writing skills in their respective disciplines is therefore essential in university-level languages for specific (LSP) courses (Flowerdew & Costley, 2017). Disciplinary writing skills, however, develop slowly (Christie, 2013; Dressen-Hammouda, 2008). They are difficult to teach and even more difficult to assess owing to the length and complexity of discipline-specific writing. One of the ways to overcome these challenges in university-level LSP courses is to use summary writing tasks to assess students’ writing skills.

This approach appears to be particularly suitable for advanced L2 students, especially in the case of English as a Foreign Language (EFL). Summary writing tasks require the writer to assess the relevance of ideas and facts presented in the text, categorize them, find connections between them, as well as to use other higher cognitive skills. As such, summary writing tasks represent a valuable learning tool (Du, 2014) complementing other useful writing-to-learn activities (Fulwiler & Young, 2000), such as note-taking, mind-mapping, writing reading journals, or reaction papers. Moreover, the result of summary writing tasks is a text of manageable length, and the LSP teacher needs only to assess summaries against the source text, which eliminates the need for the teacher to possess advanced disciplinary knowledge. We would, therefore, argue that the benefits that may arise from summary writing in LSP courses, both for students and teachers, outweigh the related difficulties.

In this paper we consider how analytic assessment of summaries can be used to provide effective formative feedback and ensure writing progress in LSP classes. More specifically, our purpose is to analyse the complexities and advantages of an analytic approach to the assessment of summaries written by junior business students in a university-level ESP course. Our research questions are broadly defined as follows:

1. What challenges are involved in the analytic approach to assessing summaries in an academic LSP class?
2. How can these challenges be efficiently overcome?
3. What information do we obtain in terms of our students’ summary-writing performance?

Having outlined our motivation and purpose in the section above, we proceed by presenting a brief theoretical overview of the processes and skills involved in summary writing in LSP classes. Next, we discuss issues relating to writing assessment, such as the benefits and drawbacks of the analytic rather than holistic approach. Then we describe our experience with the analytic assessment of discipline-specific summaries in an ESP course and give a detailed account of the assessment procedure employed. The outcomes of the assessment process are also presented as they illustrate both the challenges involved in the analytic assessment and the benefits reaped. Finally, based on insights derived from the assessment process, we discuss the implications of our findings for teaching discipline-specific summary writing.

2. Theoretical overview

2.1. The importance of summary writing

One fairly simple way to develop students' writing is to practise summary writing as a skill, which according to the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2001) reflects the ability to collect information from multiple sources, reconstruct arguments and coherently present the content to a third person. In other words, the ability to summarize is pervasively used to exemplify intermediate and advanced competences for processing text in speech and writing (Council of Europe, 2017). At higher levels of language proficiency (C1 and C2 levels), language users "can summarise long, demanding professional or academic texts in well-structured language, inferring attitudes and implicit opinions, and explaining subtle distinctions in the presentation or facts and arguments" (Council of Europe, 2017, p. 106).

Summary writing is particularly well suited to improving students' literacy skills in two interrelated ways. First, summarising is a reading-to-understand process (Sabatini, O'Reilly, & Deane, 2013) in which language competence is used to establish text coherence and its related conceptual and social content (Madnani, Burstein, Sabatini, & O'Reilly, 2013), or discipline-specific content in the context of LSP. From the cognitive perspective, summarising triggers active processing and deep engagement with the text (Perfetti, Landi, & Oakhill, 2005). As a result, a mental model of the text is built. Writers identify the main ideas in the source text and establish proper logical relations between them, which is a coherence-building exercise. Yu (2003) therefore claims that summaries can be used to assess reading comprehension as they include central ideas, distinguish facts from opinions and reflect the structure of the source text. After examining the effectiveness of different types of reading assessment, Oded and Walters (2001) also find that summary writing is a superior comprehension check in comparison with other tasks (e.g., listing concepts from the text). Numerous empirical research studies have thus proved that summarising improves reading comprehension and boosts students' reading skills (e.g., Armbruster, Anderson, & Ostertag, 1987; Bean & Steenwyk, 1984; Friend, 2001; Hill, 1991; Hirvela, 2004; Thiede, Anderson, & Therriault, 2003), which means that summarising can be productively used in LSP, L2 and the discipline classes (Center & Niestepski, 2014**)**.

Secondly, summarising is also a reading-to-write process. Delaney (2008) provides a brief overview of the intricate interplay of reading and writing processes as addressed from several theoretical perspectives. For example, from the reading perspective, those who read in order to learn (Carver, 1997; Kintsch, 1998) or to integrate information from a text (Grabe & Stoller, 2002) construct text structure and situation models of the reading material, which enables them to use information from the source text in a meaningful way (Kintsch, 1998). This means that readers who write summaries study the text carefully, establish connections between its segments (words, sentences, paragraphs or larger sections of the text) and relate the meaning to their own experience or prior knowledge (situation model), thus strengthening their comprehension. From the constructivist perspective both reading comprehension and text composition involve the construction of meaning (Delaney, 2008; Kucer, 1985; Nelson & Calfee, 1998). According to Spivey (1990, 1997), meaning is constructed through organizing, selecting and connecting the information from the reading, which is exactly how summaries are constructed. Finally, reading-to-write tasks such as summaries can improve students' writing skills beyond their ability to provide a concise factually accurate account of written content authored by others. While summarizing, students distinguish between relevant and irrelevant, practice content organisation, monitor their own coherence and think of the resulting linguistic output, which is a set of activities necessary in real-life and various professional environments.

2.2. Summary writing in LSP

Clearly, since summary writing is a reading-to-integrate task and a valuable learning tool, it is particularly useful for university-level LSP classes. This claim can be justified by reasons that fall into three groups: student characteristics, task characteristics, and teacher characteristics. As far as student characteristics are concerned, the majority of students in academic LSP classes have at least an upper-intermediate level of general L2 proficiency, which is a minimum requirement for summarising complex texts successfully, as suggested by Kirkland and Saunders (1990). Moreover, while it is true that junior students may have insufficient prior knowledge of the concepts that source texts deal with, a careful selection of readings and appropriate pre-reading activities may in fact increase students' motivation to learn and enhance their reading engagement in the field. As regards task characteristics, summary writing tasks based on authentic professional texts require that professional vocabulary be processed and reused and, as such, represent good vocabulary practice for LSP classes. They are limited in scope and allow for more frequent usage in class than longer types of writing such as essays. Most importantly, summarising supports students' learning effort and simultaneously develops two language skills, i.e., reading and writing in the professional context. Summary writing is thus usable across fields on a range of disciplinary texts, which provides grounds for a desirable fusion of academic and disciplinary writing (Du, 2014). Finally, while it is generally demanding for language teachers to develop advanced reading and writing skills in academic LSP classes due to a high requirement in terms of the knowledge of the discipline, with summaries teachers have complete control of the content and can more reliably check the quality of their students' writing content or text structure. Summaries, by definition, include only the most relevant ideas from the source text. Consequently, language professionals are better equipped to deal with summaries than with full-length essays on academic or professional topics.

Summary writing is thus a practical and useful integrative task (Delaney, 2008) that supports reading comprehension, learning, development of writing skills, and development of higher cognitive skills. A good summary reflects the writer's understanding of the source text and (possibly) the wider context, ability to discern the most relevant information, to synthetize and restate it in his or her own words. It should, therefore, not come as a surprise that the value of summary writing has generally been recognized by teachers: summary writing tasks have been used in recruitment tests for the most desirable positions, and not only to establish the candidates' language proficiency. For example, information on the written examination structure published by the recruitment pages of the United Nations' Young Professionals Program (UNYPP) includes the following guidelines on the summary writing task:

"*Please summarize the following 878 word text by reducing it to approximately one third of its original length; the summary should have around 300 words in English, you should use your judgement in deciding what the main ideas are and which points should be stressed while respecting the balance of the original. Clarity and organization will be among the elements taken into account in evaluating your summary... Your summary must be written in your own words and NOT copied directly from the text."* (UNYPP)

The instruction clearly shows that the task is aimed at testing candidates' drafting and critical thinking abilities rather than their language competence in English. This is often the case with writing tasks in corporate recruitment or in business education, for example in the Graduate Management Admission Test. It seems necessary that LSP classes, especially academic ones, should develop reading and writing skills demanded on the job market, and summaries appear to be the core writing skill here.

2.3. Assessing summaries

Providing accurate and informative feedback on writing performance (Bachman & Palmer, 1996), including summaries, is often demanding, particularly in the context of higher education. It requires a substantial amount of time and effort due to the multi-faceted nature of quality writing (Weigle, 2002). For example, Cumming, Kantor, and Powers (2002) report that all raters pay attention to a range of features such as complexity and accuracy of language, fluency, coherence, content development and organization, all of which makes assessment difficult. Attempting to reduce raters' cognitive load, the pilot version of the recently developed DELTA writing rubric employs simple, one-word labels for the scale descriptions of three lexico-grammatical criteria (e.g. exemplary, good, fair, etc.), whereas the criteria relating to content, register, and organization are fully elaborated and exemplified (Lockwood, 2016). Coping with the complex nature of writing is not the only challenge. While simultaneously engaged with the many different aspects of a written text, raters need to maintain a satisfactory level of objectivity and reliability. According to Alderson (2000), subjective rating is a major problem in summary assessment. This is generally due to raters' different professional, academic, linguistic, or cultural background (Song & Caruso, 1996; for more studies, see Guo, Crossley, & McNamara, 2013). The problem of subjectivity can be resolved with rating scales which clearly define rating criteria throughout the assessment procedure. Depending on the approach used, most scales can broadly be classified as either holistic or analytic and both can be used for grading writing tasks.

When holistic assessment is applied, every writing sample receives a single score that is based on the assessor's overall impression. The advantage of such scoring is that the writing sample is read quickly and assessed using a scoring rubric that outlines the criteria for the single grade (Weigle, 2002). The assessor weighs the different aspects of writing intuitively and delivers an overall, synthetic judgement (Council of Europe, 2001), which makes the assessment faster and less expensive. Also, such assessment can be efficiently used in selection procedures (e.g., in corporate recruitment) when assessors are not so much interested in particular qualities of a candidate but only in shortlisting those who meet professional literacy requirements and eliminating those who do not. Holistic scales thus represent a compromise between practicality and precision. Weigle (2002) lists several disadvantages of holistic scoring. Holistic scores are difficult to interpret. They obscure the quality of the different aspects of writing (e.g. syntax, organization, vocabulary) and cannot be used as a diagnostic tool. The mentioned disadvantages are particularly strongly felt in L2 contexts since different aspects of writing ability do not develop at a uniform rate. As a result, individual writers demonstrate unique combinations of strengths and weaknesses that may not be properly assessed holistically (Alderson, 1991).

Analytic scoring uses analytic scales and provides separate scores for different aspects of writing ability. Weigle (2002) gives three examples of extensively piloted and revised analytic scales. The ESL Composition Profile criteria, widely known analytic scoring scales designed by Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey (1981) aimed at evaluating different types of expository texts, cover five aspects of writing: content, organization, vocabulary, language use, and mechanics. The set of scales developed by Weir (1988) for the Test in English for Educational Purposes (TEEP) assess relevance and adequacy of content, compositional organisation, cohesion, adequacy of vocabulary for purpose, grammar and mechanical accuracy (punctuation and spelling). The Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1990, 1991) includes scales for ideas and arguments, rhetorical features and language control.

Clear and detailed scoring rubrics are particularly welcome when using such criteria, as has been established by many researchers (e.g. Alderson, 1991; Bachman & Palmer, 1996; Lockwood, 2016). As a result of the analytic approach and explicit criteria, analytic scoring provides valuable diagnostic information to teachers and test takers alike (Bernhardt, 2010; Lockwood, 2016; Weigle, 2002). Using the results of analytic scales, teachers can see if they should put more stress on receptive or productive skills, as well as which aspects of the reading-to-write ability should be given more attention (e.g. whether to focus on local or global text understanding, or whether to encourage or discourage students from using cohesive devices). According to Weigle (2002), analytic scoring is also more reliable than holistic scoring. This is not to say that analytic scoring for diagnostic purposes has no disadvantages: dealing with several aspects of writing is extremely time-consuming and mentally exhausting.

3. Method

In this study, we decided to explore the process and effects of analytic assessment of our students' summaries. Our aim was threefold. First, we wanted to gain a detailed insight into our students' reading and writing competences, inform our teaching, and facilitate meaningful feedback in our English for specific (ESP) classes dedicated to developing writing skills. Next, we aimed to determine the challenges involved in assessing summaries written by junior business undergraduates in an ESP course. The study was part of a broader research investigating reading comprehension and writing skills in ESP. Finally, our aim was to develop an assessment procedure that would reliably provide insight into our students' reading and writing abilities. In order to do that, we implemented a strictly defined protocol that we describe below. Wishing to reduce subjectivity, we opted for the analytic approach and detailed scoring rubrics. Since summarising in an academic LSP context reveals much about students' academic abilities, we focused on the content of summaries and their text

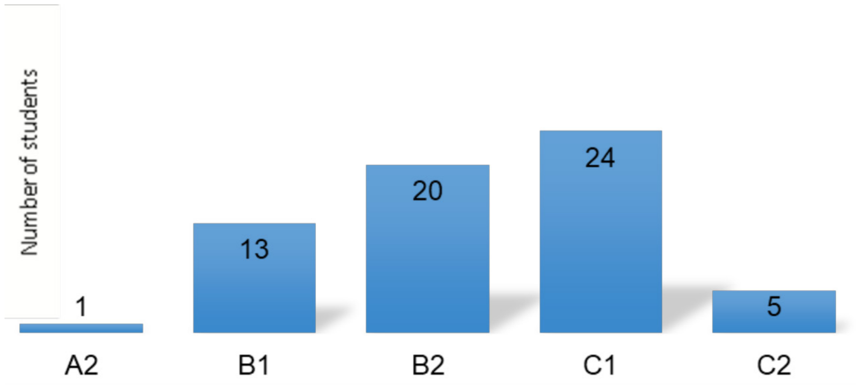

Figure 1. Participants' L2 competence in English based on the CEFR.

quality (e.g. coherence and cohesion) and ignored the aspects of writing (e.g. grammar and lexical knowledge) that are tested in other, simpler ways.

3.1. Participants

Our research included 63 first-year students (28 males and 35 females) attending an English for Business and Economics course. The students belonged to a convenience sample from a total of approximately 1,000 first-year students who are divided into ten alphabetically ordered teaching groups of roughly equal size (approximately 95 students). The participants were students from one such group who volunteered for the task in exchange for a slightly reduced number of written assignments required for the course. The students' competence in English as L2 was determined using the Quick Placement Test (Allan, 2004). As could be expected, the majority of students (N=49) were at B2 level or higher (C1 and C2) according to the Common European Framework of Reference (Figure 1).-

3.2. The summarizing task

The students were asked to summarize a text taken from *The Economist*, a business magazine, entitled "Montessori Management" (2013, 7 September). The text was 902 words long, and the students were required to summarize it by reducing it to approximately a third of its original length (+/- 10%). The students had previously received no instruction on summarizing and the exact wording of the task was as follows: 'Write a summary of the text of about 300 w. (+/- 10%).' Since the text was fairly difficult for L2 students (Flesch-Kincaid Grade Level, 10.79), and we were interested in the content and text quality of our students' summaries in a task resembling a real-life activity, the students were allowed to keep the text throughout their work. The students summarised the text in a controlled classroom environment. There was no explicit time-limit for task completion, but the majority of the students completed the task in about 80 minutes (which also included the reading time).

3.3. Assessment criteria

In our study we used a set of assessor-oriented scales (Alderson, 1991) compiled as part of the previously mentioned research on reading comprehension. All the criteria were derived from literature (e.g., University of Cambridge ESOL Examinations, 2008; TOEFL; Jacobs et al., 1981) and adapted where necessary. The scales were designed for the specific purpose of analytic assessment of academic summaries. The set reflected the integrated nature of the summary task (reading-to-understand and reading-to-write). In other words, it integrated the need to assess a piece of student writing demonstrating reading comprehension on the one hand and writing quality on the other. Consequently, our set of scales comprised two groups of criteria: content criteria (factual accuracy, completeness, relevance and coherence) and text quality criteria (organization and cohesion). The criteria were defined as precisely as possible and accompanied by explicit and detailed descriptors for grades ranging from 0 to 3. Since the scales were appropriate for diagnostic purposes, the grades were to be reported separately (analytically).

We applied the criterion of factual *accuracy* to all the facts reported in the summaries of our participants. The criterion of *completeness* was used to determine how many of the main ideas (previously defined by independent experts) were included in the summaries. *Relevance* reflected an absence of unnecessary details and unjustified copy-pasting. The criterion of *coherence* was used to assess the logic of the text at the macro and micro levels, i.e. the logical flow of ideas within sentences, paragraphs and the text as a whole. The text quality criterion of *cohesion* was used to assess whether the text ran smoothly, linguistically speaking. *Text organization* implied the use of topic sentences, good paragraphing, and well-developed internal paragraph structure.

For each criterion a scale of four-level descriptors was obtained by asking general questions about summaries (e.g. Is all the information in the summary accurate? Are all the main ideas covered? Is there a logical flow of ideas, are there coherence breaks?). The descriptors for the grades from 0 to 3 quantified the number of errors (e.g. coherence breaks) allowed at each of the four levels of all the six scales. For example, coherence was assessed on the following scale: 3 = entirely coherent (logical flow of ideas); 2 = few coherence breaks (1 or 2); 1 = several coherence breaks (3 or 4); 0 = numerous coherence breaks (> 4). As a rule, across the six scales, grade 3 signified that the summary fully complied with the criterion, grade 2 reflected one or two lapses, grade 1 three or four lapses, while grade 0 signified that there were more than 4 errors of a specific kind in the summary.

3.4. Assessment protocol

We followed an assessment protocol which consisted of several stages. Since we knew the students personally, we first coded the summaries to obscure the identity of the authors and enhance our objectivity. In order to avoid inconsistencies in scoring, both in the ratings of different summaries by a single scorer and between different scorers, we observed the procedures outlined by White (1984). Our protocol was divided into pre-assessment, preliminary assessment, and final assessment.

In the pre-assessment stage, we engaged in self-training according to the procedure presented by Weigle (2002). The purpose of self-training was to achieve higher inter-rater reliability, i.e. the tendency of different raters to give the same grades to the same summaries. After carefully studying the scoring rubric, we applied it to the sample (non-participant) summaries to discuss the wording and meaning of the descriptors for different criteria and levels, or to resolve specific scoring problems. During the self-training process, several issues were raised and discussed, which we briefly refer to in a later section of this paper.

We then moved on to the stage of preliminary assessment in which we individually assessed five summaries only to discuss them afterwards, checking for possible inconsistencies or ambiguities. This was followed by minor revisions in the assessment criteria where necessary. For example, the descriptor ‘completely incoherent’ changed to ‘incoherent on numerous occasions (>4)’. When such revisions happened, it was necessary to reassess the summaries scored before.

The final stage of the assessment was carried out in a controlled reading session (White, 1984). Both raters obtained copies of all the 63 summaries, divided into six batches of about 10. After reading and scoring the first 10 summaries independently, the raters met to compare the scores. The process was repeated with all the batches. In other words, we checked on the reading in progress after every 10 summaries to make sure that the standards we had previously agreed upon were maintained. Since we were reading and assessing relatively long texts (about 300 words), we developed a specific procedure to deal with the presumed lapses in the raters’ concentration. If we discovered that our grades differed by more than 1 score point out of a total of 4 (grades 0-3), we returned to the source text to analytically establish which grade from the rubric was supported by evidence from the text. Based on the discussion, we would reach an agreement, and the grade would be adjusted accordingly.

The described procedure was applied to all the summaries. All the grades were first individually assigned, then compared, discussed, and adjusted in case of differences. It should be noted that although we agreed on all the grades together, we still recorded the initially awarded ones in order to calculate inter-rater reliability. This means that we used our independently assigned grades to investigate the reliability of our scoring criteria in a demanding assessment procedure involving relatively long summaries based on complex expository reading. Alternatively, we could have opted for a procedure in which two different scores would eventually have been averaged and recorded as a mean value. If this had been the case, however, we would not have had the opportunity to discuss our occasional underperformance as raters. Furthermore, we would not have been fully aware of the complexity of the assessment. We therefore believe that our discussions contributed significantly to the quality of the assessment.

3.5. Statistical analysis

After completing the assessment procedure, descriptive statistics (mean and standard deviation) were calculated for all of the data relating to the six assessment criteria (accuracy,

Table 1.
Inter-rater reliability in the assessment of summaries.

Inter-rater reliability measure	Accuracy	Completeness	Relevance	Coherence	Text organization	Cohesion
Intraclass Correlation Coefficient (ICC)	0.711	0.860	0.719	0.825	0.869	0.906

completeness, relevance, coherence, cohesion, and text organisation). Intraclass Correlation Coefficient (ICC) was used to calculate inter-rater reliability. SPSS statistical package was applied for all the statistical analyses.

4. Results

4.1. Inter-rater reliability

During the assessment procedure, after the previously mentioned three rounds of ten summaries, we used the awarded grades to calculate inter-rater reliability for 30 summaries. In this way we checked if our assessment method was able to provide us with reliable information on the quality of our students' summaries. The results are presented in Table 1.

Inter-rater reliability for the criteria of completeness, coherence, text organization, and cohesion was satisfactory (>0.8), while our grades for accuracy and relevance showed somewhat less agreement (>0.7). Interestingly, as raters we performed worst at what our students apparently were best at: accuracy and relevance (see below). It appears that we were often not quite certain what to make of the students' good choice of copy-pasted, yet relevant sentences, which occasionally failed to fit into a coherent whole with the rest of the summary. In other words, as assessors using analytic assessment scales we frequently found ourselves in bewildering situations, discussing differences between accuracy and coherence, or coherence and cohesion. We had to establish whether something could be accurate and incoherent, or inaccurate and coherent at the same time. We found that both were possible. We defined accuracy as a fact-based criterion that can only be applied at the local micro-level of clauses and sentences as suggested by Sabatini, O'Reilly, and Deane (2013). Coherence, however, belongs to a more global level of the text and is used to assess whether a summary accurately reflects the meaning of the text by correctly presenting the logical flow of ideas, not details. Consequently, according to the descriptors used in our scoring rubric, factual accuracy at the micro-level could exist alongside incoherence at the macro-level of the text.

Also, the question arose if a text segment can be coherent but not cohesive, or incoherent but cohesive. This dilemma illustrates the difference between content quality and text quality. Coherence is a content issue, the result of text-level comprehension, while cohesion is a writing issue, the result of well-connected text segments. Consequently, a text segment may be well-connected and, at the same time, completely incoherent.

Table 2.
Descriptive statistics for content and text quality of the students' summaries.

	M	SD
Accuracy	2.46	0.64
Relevance	1.86	0.71
Completeness	1.62	0.78
Organization	1.48	0.92
Cohesion	1.44	0.81
Coherence	1.25	0.84

4.2. Descriptive statistics for grades assigned to student summaries

Having established the reliability of our assessment procedure, we then proceeded with a simple statistical analysis of our students' performance when writing summaries. As expected, the results provided us with useful insight into our students' strengths and weaknesses. Descriptive statistics (Table 2) show that on a scale from 0 to 3 our students received the highest grades for *factual accuracy* (mean [M] = 2.46, standar deviation [SD] = 0.64). In other words, the facts presented in their summaries were for the most part accurate, which at first may sound as evidence of good performance. A closer inspection of the summaries, however, revealed that high levels of factual accuracy were in many cases achieved by a fair amount of copying. Namely, students had not been explicitly warned against copying from the source text as we were interested in their reading-to-write abilities in L2 in a pre-teaching stage. It is therefore likely that many of them, especially those with weaker English writing skills, resorted to this practice (Kirkland & Saunders, 1991). Moreover, the average grade for the *relevance* of facts and ideas reported in the summaries was somewhat lower (M = 1.86, SD = 0.71), reflecting that some of the accurately copied information was too elaborative or too superfluous to be considered relevant (in 21 summaries graded 0 or 1). The average grade for *completeness* (M = 1.62, SD = 0.78) signalled that a fair number of the summaries failed to present a number of the main ideas from the source text (30 summaries graded 0 or 1).

The rest of the results point to areas which, in our opinion, require stronger teaching effort in academic LSP classes attended by junior college students. The average grades for *text organization* (M = 1.48, SD = 0.92) and *cohesion* (M = 1.44, SD = 0.81) revealed that the quality of writing was rather poor. Text organization (e.g. topic sentences, clear paragraphing) and cohesive devices were not adequately used (grades 0 and 1 were assigned for text organisation in 36 summaries and for cohesion in 40 summaries). Furthermore, descriptive statistics for *coherence* (M = 1.25, SD = 0.84) demonstrated that performing well in this criterion was the most difficult aspect of the summary writing task for our students. Since coherence is a content criterion related to the process of reading-to-understand (Sabatini et al., 2013), it follows that the summaries reveal serious problems with comprehension. It appears that the students struggled to establish and present logical relations among the main ideas of a text relevant to their professional studies. Overall, the average grades from the bottom of Table 2 point to the need to improve student performance in both reading (establishing coherence of the source text) and writing (achieving cohesion in one's own text).

Table 3.
Number of students with satisfactory performance in one or more assessed criteria (grades 2 or 3).

	1 criterion **N=13**	**2 criteria** **N=15**	**3 criteria** **N=10**	**4 criteria** **N=5**	**5 criteria** **N=5**	**6 criteria** **N=15**
Satisfactory						
Accuracy **(Total 60)**	Acc (detail): 13	Acc (detail): 13 +1 criterion Rel: 7 Comp: 3 Org: 2 Chs: 1	Acc (detail): 10 +2 criteria Comp+Rel: 7 Chr+Rel: 1 Org+Rel: 1 Org+Chr: 1	Acc (detail): 5 +3 criteria Comp, Rel, Chs: 1 Comp, Rel, Org: 1 Comp, Org, Chs: 1 Relev, Chr, Org: 1 Relev, Org, Chs: 1	Acc (detail): 4 +4 criteria Comp, Rel, Chr, Org:1 Comp, Rel, Org, Chs: 2 Rel, Chr, Org, Chs: 1	Acc (detail): 15 +5 criteria All criteria: 15
Low Accuracy **(Total 3)**		Rel+Comp: 1 Rel + Chs: 1			Comp+Rel+ Chr+Org+Chs: 1	

N=number of student, 1= satisfactory in 1 criterion only, 2= satisfactory in 2 criteria, 3= satisfactory in 3 criteria, 4=satisfactory in 4 criteria, 5=satisfactory in 5 criteria, 6=satisfactory in all criteria; Acc =Accuracy, Comp=Completeness, Rel =Relevance, Chr=Coherence, Chs=Cohesion, Org=Organisation

4.3. Student achievement – the number and combinations of grades achieved

Wishing to obtain a deeper understanding of our students' achievement, we decided to reduce the scale to only two grades: pass ('positive' grades 2 and 3) and fail ('negative' grades 0 and 1). The grades were used to determine how many of our students received satisfactory (positive) grades for specific aspects of their summaries (e.g. accuracy, relevance, coherence) and which combinations of the positive grades prevailed. The paragraphs that follow present the participants' achievements in terms of particular combinations of strengths and weaknesses as demonstrated by their writing and captured by the analytic approach to assessment.

As mentioned before, a good level of accuracy was achieved by as many as 60 out of 63 participants, possibly due to the fact that many of them correctly copied from the source text. Consequently, the criterion of accuracy as defined in our study (factual accuracy) could not be meaningfully applied. It can, therefore, be concluded that instructions to students should explicitly prohibit copying.

Table 4.
No. of students with satisfactory performance (grades 2 or 3) in individual criteria (N=63).

Accuracy	Completeness	Relevance	Coherence	Cohesion	Organisation
60	33	42	21	23	27

Since the ability to correctly copy from the text was nearly universal among our participants, we divided them into groups according to the number of positive grades they received for other aspects of their writing. As a result, six groups were formed, ranging from the group that satisfied only the criterion of accuracy to the group that received positive grades for all the six aspects of writing covered by the scoring rubric. The groups are presented in Table 3 together with the number of students with satisfactory performance in one of the criteria (Table 4).

Overall, the results show that over a third of the students were able to do little more than copy factual details correctly from the source text. On the other hand, relevance was the second most frequent criterion in which the students achieved a high grade (after accuracy). It characterised over half of the summaries that met only two criteria and as many as two thirds of all the summaries, which indicates some knowledge construction. Also, relevance was often followed by completeness, which may be defined as the ability to select *all* the relevant parts of the source text. Still, the grades assigned in the criteria measuring text integration in terms of relations between ideas, i.e. coherence as a RU measure and cohesion/organisation as RW measures, showed that only a third of the students connected the ideas meaningfully (mostly those that performed well in all the other criteria).

More specifically, the group that demonstrated only the ability to correctly copy from the source text included 13 students. Not only was this group rather large, but the group which was able to satisfy only one more criterion besides accuracy also numbered 13 students. As regards the summaries in which accuracy coincided with only one other aspect of writing, seven reflected that their authors were at least able to identify relevant information in the source text, three presented most or all of the main ideas from the source text, two were well-organized, and only one included cohesive devices. It should be noted that not one of these summaries had a satisfactory level of coherence. The next group included ten students who received positive grades for accuracy and two other criteria. Seven of these summaries were accurate, complete and relevant, but only two summaries in this group demonstrated a satisfactory level of coherence. Out of 36 summaries presented so far, only four have demonstrated a satisfactory level of text organization, only two have conveyed the logic of the whole or parts of the source text (coherence), and only one has employed cohesive devices.

Achieving coherence and cohesion posed the most difficult problems even for students who had good results overall. In the group of five students who received four positive grades out of a total of six, two students achieved a satisfactory level of text coherence while another two students achieved an acceptable level of text cohesion. In the group of four students who received five positive grades, text coherence was achieved by only two students. Finally, there were 15 summaries, almost a quarter of the total sample, which received positive grades for all the aspects of writing in the scoring rubric. Given the analysis presented above, despite our awareness that a number of students struggled with the summary writing task, it seems quite remarkable that a quarter of our participants wrote accomplished summaries of a difficult professional text in L2 prior to instruction.

5. Conclusions

Summary writing is useful in academic LSP classes since it integrates both reading and writing skills, composition and disciplinary knowledge. It also provides students with an opportunity to process and reuse professional vocabulary. Assessment of summaries, however, is difficult and costly in terms of time. This is especially true when the analytic approach is used as assessment of individual aspects of writing requires multiple readings and much concentration if longer texts are involved. Nevertheless, analytic assessment is too valuable not to be used in academic LSP classes. As demonstrated by our study, it is a diagnostic tool that provides valuable information on students' abilities and areas where improvements are needed (e.g. avoiding meaningless plagiarism and enhancing coherence), especially if mixed-ability groups are in question.

Our analytic procedure proved to be a reliable tool for scoring university-level summaries in advanced ESP classes, especially when assessing cohesion, completeness, coherence and text organisation. While our moderate inter-rater reliability scores for the factual accuracy and relevance of our students' summaries may not look too impressive to an inexperienced eye, we believe our results are rather satisfactory considering exceptional cognitive demands involved in the assessment which included (a) complex writing, (b) analytic scoring, and (c) more categories (Council of Europe, 2001). Moreover, as a result of a number of problems we encountered during the assessment, we have managed to develop a fairly precise structural framework for resolving our dilemmas. In other words, the complexity of our analytical approach induced us to develop a more precise interpretation of our students' abilities when reading and writing in the ESP context. For example, analysing summaries as a form that allows for the control of writing content turned out to be a highly useful tool for distinguishing between errors of coherence on one side and cohesion or factual accuracy on the other. The benefits of our assessment exercise can therefore not be stressed enough for ESP specialists interested in helping students to prepare for the ever-increasing EFL demands in the professional global community.

As regards our students' performance in the summary writing task presented in this study, the benefits of analytic scales appear even if only simple statistics are used. At this exploratory stage, the overall results of our study demonstrate that almost all of our students were able to correctly copy the facts, figures, and part sentences from the source text, but they were less skilled at selecting the relevant information for their summaries as they frequently focused on irrelevant details. Furthermore, despite the fact that the majority of the students tended to engage in undiscerning, albeit accurate, copying, they frequently left out one or more of the main ideas and thus produced partial accounts of the source text. The sequencing and arrangement of these (relevant and irrelevant) ideas in the summaries often failed to convey the logic of the text, or any logic for that matter. Finally, regardless of the quality of the content presented in the summaries, many were difficult to read as text organisation features and cohesive devices were frequently lacking.

This is not to state that all of our students experienced the problems listed above. Our analytic assessment revealed further useful information about the mixed-ability sample used in the research. For example, almost a quarter of the students in our sample were able to produce very good summaries with no previous instruction, and they achieved satisfactory results for all the aspects of writing included in our scoring rubric. On the other side, there was a group of similar size that performed rather poorly on all the criteria but accuracy, and another equally large group of the students who were only able to correctly copy relevant information or to write summaries that were complete but irrelevant, incoherent, and badly organized.

With regard to the aspects of writing that posed the greatest challenge for the participants in our study (text organisation, cohesion and coherence), our approach enabled us to note several interesting findings. Good text organization typically coincided with text cohesion, and both were achieved by over a third of the total number of the students in our sample. Moreover, a satisfactory level of text coherence was not achieved by as many as two thirds of research participants, which may indicate a major comprehension problem, possibly due to a lower level of EFL and a lack of familiarity with more complex expository reading at an early stage of tertiary education.

Regardless of the assessment-related difficulties illustrated by our own research experience, we believe that the multitude of precise data resulting from our analytic approach justify the amount of time and significant effort invested by the assessors. We also believe that future research will overcome the limitations of our study by including participants from various college-level programmes and different countries and by conducting deeper statistical analyses, possibly comparing holistic and analytic assessment. Moreover, we hope that this account of our assessment experience may dispel doubts about the process and encourage LSP teachers to apply analytic scoring to summary writing tasks regardless of the practical difficulties. The following section on the implications for teaching therefore includes further ideas for classroom practice.

6. Implications for teaching

6.1. Dealing with specific aspects of summary writing in LSP classrooms

Our analytic assessment of academic LSP summaries written by first-year students of business and economics prior to instruction resulted in a series of practical suggestions for an effective application of summary writing tasks in LSP classrooms. Firstly, when teaching summary writing, it is necessary to deal with the widespread tendency of students to copy from the source text. Several strategies could easily be used for that purpose. It is necessary to practice paraphrasing and to insist that summaries must convey the main ideas in students' own words. Furthermore, although text-present summarising has a valuable positive effect on reading comprehension, text-absent summarising tasks need to be introduced too in order to encourage rephrasing. Therefore, a combination of text-present and text-absent summarising tasks may be the best approach to developing reading and writing skills.

Secondly, LSP teachers should devote some time to raising their students' awareness of the characteristics of well-written expository texts: clear structure, usefulness, and functions of paragraphs and topic sentences. Numerous researchers warn that teachers in academic LSP classes should not assume that their students are efficient readers, let alone skilful writers (e.g., Center & Niestepski, 2014; Du, 2014). In order to scaffold both their students' reading comprehension and writing, students could be presented with structured models of different types of writings, summaries included.

Thirdly, it is necessary to raise students' awareness of the effects achieved by an appropriate use of cohesive devices. While some students will only need to be reminded of the importance of cohesive devices for the unity of text in the English language, others will only improve the quality of their texts through systematic practice.

Finally, students need to be encouraged to produce coherent texts. The lack of coherence in students' summaries derived from the integrated nature of the summary task as summarising involves both reading-to-understand and reading-to-write processes. Consequently, a lack of coherence in students' summaries may reflect both a lack of understanding and a poor ability to write clearly. Students should be made aware of both their own lack of understanding and the need to make their writing content clear to others. Introducing students to text organization principles and models of text structure will both boost students' reading comprehension and provide them with the tools necessary to effectively convey meaning in writing.

Work on the mentioned aspects of summary writing requires appropriate source texts, and teachers sometimes worry that texts are too difficult in terms of vocabulary or that they have a structure which is not clear enough. While this is often true, we should not forget that summarising is in fact a meaning construction task, the purpose of which is to deeply engage with the text in a way that is sometimes beyond students' comfort zone. Tasks of the kind do not provide quick, easy-to-use solutions, but require deep concentration and dedication as prerequisites for reading comprehension or for writing production. If such prerequisites are met, most students will find a way to resolve the gaps in their understanding/writing of the text, thus developing a vital professional skill that might be described as 'mental resourcefulness'. This may be a painful process, so the role of the teacher as a motivator, facilitator and provider of relevant feedback becomes most prominent here.

6.2. Time-saving approaches to analytic summary assessment in LSP classes

When carefully tailored to the specific needs of academic LSP students, summary tasks can clearly be described as excellent practice. But these tasks open the problem of time-consuming feedback. A natural solution might be to use shorter source texts resulting in shorter summaries and, consequently, shorter assessment time. This is a legitimate teaching approach in which the size of the text can gradually increase if deemed appropriate time-wise. Using longer texts (e.g. over 500 words), however, offers more opportunities to assess the key aspects of summary writing, such as the ability to establish discourse-level coherence (which requires more inferencing when linking more distant parts of the text). As a result, we argue for the use of longer texts in class as group work or as take-home exercises despite the practical difficulties described above. The question then is how to bypass the obvious problem of the substantial amount of time used to read and correct such assignments.

A few suggestions might be considered here. First, while it is true that individual feedback works best, a frustrating workload for the teacher must be avoided. We therefore recommend providing individual feedback on summaries only twice during the term if large classes do not allow for more: once in the middle of the term (after a brief training session and practice in class), and once towards the end of the term so that students can assess their own progress and possibly improve before the completion of the course. We also strongly recommend keeping to well-designed analytic scoring scales in order to ensure objectivity to the extent possible.

Secondly, despite all the difficulties for the teacher, feedback is necessary throughout the term. Teacher-guided self-assessment and peer review might be considered instead of individual assessment. Teachers could first make students aware of the criteria involved in summary writing, use a few examples to illustrate the scoring procedure, provide some collective feedback, and then delegate assessment to students themselves (either as self-assessment or peer assessment). While the idea of self-assessment or peer assessment may sound contradictory considering the amount of effort the authors of this paper took to complete the assessment themselves, our teaching experience tells us that students tend to take peer assessment rather seriously, especially if they are asked to justify their scoring. As an awareness-raising exercise, this can prove highly useful for students and their overall attitude to reading and writing critically. Having been trained to read and write coherently and relevantly, they will be more able to competently evaluate the work of others and of themselves.

In conclusion, the aim of this paper was to present the importance of writing summaries as a skill worth developing in LSP classes as well as to present difficulties related to summary writing assessment. We pointed out that assessment may represent an obstacle to practicing such assignments, as it is highly demanding for assessors, yet often unreliable if inadequate practices are in place. We therefore presented our own assessment experience to illustrate the advantages of the analytic approach to scoring and strict assessment protocols. Our experience with analytic assessment of summaries in university-level ESP classes and the resulting insight into our students' reading and writing competences testify to the rich benefits of such scoring practices for formative assessment or diagnostic purposes. As demonstrated, the effectiveness, reliability, and objectivity of the assessment process depend on the quality and precision of the analytic scales employed, as well as on raters' readiness to adhere to a clearly defined assessment procedure. Our analytic approach to the assessment of summaries has shown that discipline-specific summary writing tasks can be usefully and efficiently employed to direct LSP practitioners towards introducing more coherence-building reading and writing tasks. This, in turn, will help students become competent critical readers and writers.

Acknowledgments

We would like to thank Professor Svjetlana Kolić-Vehovec for her most helpful advice on the construction of the analytic scales used in this study and to Professor Anamaria Ekert-Kabalin for helping us with statistics. Our thanks go to the anonymous reviewers whose comments significantly contributed to the quality of this paper. We are also grateful to Marin Andrijašević for technical support.

References

Allan, D. (2004). *Oxford Placement Test*. Oxford: Oxford University Press.

Alderson, J. C. (1991). Bands and scores. In J. C. Alderson & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71-86). London: Modern English Publications/British Council/Macmillan.

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

Armbruster, B. B., Anderson, T. H., & Ostertag, J. (1987). Does text structure/summarization instruction facilitate learning from expository text? *Reading Research Quarterly*, *22* (3), 331-346.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford: Oxford University Press.

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Reading Behavior, 16*(4), 297-306.

Bernhardt, E. B. (2010). *Understanding advanced second-language reading*. New York: Routledge.

Carver, R. P. (1997). Reading for one second, one minute, or one year from the perspective of reading theory. *Scientific Studies of Reading, 1*(1), 3-43.

Center, C., & Niestepski, M. (2014). "Hey, did you get that?": L2 student reading across the curriculum. In M. Cox & T. M. Zawacki (Eds.), *WAC and second language writers: Research towards linguistically and culturally inclusive programs and practices* (pp. 93-111). Anderson: Parlor Press and the WAC Clearinghouse.

Christie, F. (2013). *Language education throughout the school years: A functional perspective.* New York: John Wiley & Sons Inc.

Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Available at http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf (Accessed 11 April, 2017)

Council of Europe (2017). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Available at https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/168074a4e2 (Accessed 10 September, 2017)

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*(1), 67-96.

Delaney, Y. A. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7* (3), 140-150.

Dressen-Hammouda, D. (2008). From novice to disciplinary expert: Disciplinary identity and genre mastery. *English for Specific Purposes, 27*, 233-252.

Du, Q. (2014). Bridging the gap between ESL composition programs and disciplinary writing: The teaching and learning of summarization skill. In M. Cox & T. M. Zawacki (Eds.), *WAC and second language writers: Research towards linguistically and culturally inclusive programs and practices* (pp. 113-128). Parlor Press and the WAC Clearinghouse.

Flowerdew, J., & Costley, T. (2017). *Discipline-specific writing: Theory into practice.* Abingdon, Oxon: Routledge.

Foltz, P. W. (2016). Advances in Automated Scoring of Writing for Performance Assessment. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of Research on Technology Tools for Real-World Skill Development*, (pp. 659-678). ICI Global.

Friend, R. (2001). Effects of strategy instruction on summary writing of college students. *Contemporary Educational Psychology, 26* (1), 3-24.

Fulwiler, T., & Young, A. (2000). *Language connections: Writing and reading across the curriculum.* WAC Clearinghouse Landmark Publications in Writing Studies: https://wac.colostate.edu/books/language_connections/ (Accessed 3 February, 2016)

Grabe, W., & Stoller, F. (2002). *Teaching and researching reading.* London: Longman.

GMAT. The official website of the GMAT exam. Available at https://www.mba.com/global/the-gmat-exam/gmat-exam-format-timing/analytical-writing-assessment/analysis-of-argument-question.aspx (Accessed 15 September, 2017)

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. *Assessing Writing, 18* (3), 218-238.

Hill, M. (1991). Writing Summaries Promotes Thinking and Learning across the Curriculum: But Why Are They so Difficult to Write? *Journal of Reading, 34* (7), 536-539.

Hamp-Lyons, L. (1990). Second language writing. Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). New York: Cambridge University Press.

Hamp-Lyons, L. (1991). Pre-text: Task related influences on the writer. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 97-107). Norwood, NJ: Ablex.

Hirvela, A. (2004). Connecting reading and writing in second language writing instruction. Ann Arbor, MI: University of Michigan Press.

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

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.

Kirkland, M., & Saunders, M. (1990). Maximizing student performance in summary writing: managing cognitive load. *TESOL Quarterly, 25,* 105-121.

Kucer, S. (1985). The making of meaning: reading and writing as parallel processes. *Written Communication, 2,* 317-336

Lockwood, J. (2016). Towards a specific writing language assessment in Hong Kong. In J. Flowerdew & T. Costley (Eds.) *Discipline-specific writing: Theory into practice* (p.p. 196-215). Abingdon, Oxon: Routledge.

Madnani, N., Burstein, J., Sabatini, J., & O'Reilly, T. (2013). Automated scoring of a summary writing task designed to measure reading comprehension. *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–168). Stroudsburg, PA: Association for Computational Linguistics.

Montessori Management (7 September, 2013). Available at https://www.economist.com/news/business/21584947-backlash-against-running-firms-progressive-schools-has-begun-montessori-management (Accessed 11 April, 2017).

Nelson, N., & Calfee, R. (1998). The reading-writing connection viewed historically. In N. Nelson, & R. Calfee (Eds.), *The reading-writing connection. Ninety-seven yearbook of the National Society for the Study of Education* (pp. 1-52). Chicago: National Society for the Study of Education.

Oded, B., & Walters, J. (2001). Deeper processing for better EFL reading comprehension. *System, 29* (3), 357-370.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The Acquisition of reading comprehension Skill. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227-247). Oxford: Blackwell.

Sabatini, J., O'Reilly, T., & Deane, P. (2013). *Preliminary reading literacy assessment framework: Foundation and rationale for assessment and system design*. ETS Research Report Series. Princeton, New Jersey: Educational Testing Service.

Song, B., & Caruso, I. (1996). Do English and ESL faculty differ in evaluating the essays of native English-speaking, and ESL students? *Journal of Second Language Writing, 5,* 163-182.

Spivey, N. (1990). Transforming texts: Constructive processes in reading and writing. *Written Communication, 7*, 256-287.

Thiede, K. W., Anderson, M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95* (1), 66.

TOEFL (n.d.). iBT/Next Generation TOEFL Test. Integrated Writing Rubrics (Scoring Standards). Educational Testing Service. Available at https://www.ets.org/Media/Tests/TOEFL/pdf/Writing_Rubrics.pdf (Accessed on 25 July, 2017).

University of Cambridge ESOL Examinations (2008). *Certificate of Proficiency in English, Handbook for teachers.* Cambridge: UCLES.

UNYPP (United Nations Young Professionals Program). (n.d.) Written Examination Structure. Available at: https://careers.un.org/lbw/home.aspx?viewtype=NCES (Accessed on 1 February, 2017).

Weigle, S. C. (2002). *Assessing Writing*. Stuttgart: Ernst Klett Sprachen.

Weir, C.J. (1988). Construct validity. In A. Highes, D. Porter & C.J. Weir (Eds.), *ELTS validation project report (ELTS Research reports I (ii)*. London: The British Council/UCLES.

White, E. M. (1984). Holisticism. *College Composition and Communication, 35* (4), 400-409.

Yu, G. (2003). Reading for Summarization as Reading Comprehension Test Method: Promises and Problems. *Language Testing Update, 32,* 44-47.