National and Kapodistrian University of Athens

Department of English Language and Literature

MA Programme "Linguistics: Theory and Applications"

L2 C level Writing Proficiency: KPG Integrated Tasks and Language Interaction in Rated Scripts

Eleni Pappa

217022

Supervisor: Bessie Mitsikopoulou

2019

Declaration

This submission is my own work. Any quotation from, or description of, the work of others is acknowledged herein by reference to the sources, whether published or unpublished.

Other supervisors:

Elly Yfantidou

Anna Xatzidaki

Abstract[1]

Research into L2 writing assessment has largely focused on mapping textual features onto rater-judged candidate performance, exploring issues related to writing quality and rater reliability.

However, due to issues such as the ambiguous wording of the marking scale (Lumley, 2002) or the raters' difficulty to score borderline essays (Gebril & Plakans, 2014), very little has been found with respect to fine distinctions between adjacent levels of language proficiency, C level (C1-C2) in particular.

In this line, the current research aimed to investigate KPG C level rater-judged candidate performance in integrated tasks of two types, an intralingual and an interlingual mediation task. Using a sample of 66 rated scripts (33 candidates), three points were addressed: a) the effect of two different types of texts, an expository blog and a narrative encyclopedic entry with an expository task requirement, on language realization, b) interrater variation and c) cohesion and coherence as a potential candidate performance differentiating language criterion within C level.

Quantitative analysis results indicate that, first and foremost, Coherence & Cohesion and, second, Vocabulary Range, can allow for distinctions within C level. What is more, their accentuated predictive strength when combined with the Appropriacy criterion can yield a more global (in terms of tasks) account of observed

[1] A Greek version of this abstract can be found on the last page of this dissertation.

variance. Furthermore, although in sum interrater differences were deemed negligible, inter-activity variance tied Vocabulary Range with the intralingual mediation task and Grammaticality/ Accuracy with interlingual mediation.

Qualitative analysis targeting Cohesion & Coherence using Coh-Metrix suggested that it is mostly surface level features, and word information indices in particular, that differentiate C1 and C2 situated performance.

These findings can inform current KPG research into task analysis and ongoing rater training programmes. Additionally, they are in support of KPG's practice of applying different gravity status to language criteria depending on the level of proficiency, while recommending that the task process in question be taken into consideration as well.

Acknowledgements

This dissertation would not have been possible but for Associate Professor Bessie Mitsikopoulou’s invaluable and continuous help and support. Thank you for going above and beyond. Thank you for not giving up on me.

I would like to thank Associate Professor and Associate Director of RCeL Kia Karavas for generously giving me access to the KPG Script Database, which allowed me to draw data for this study.

I would also like to extend my sincere thanks to Dr. Vassiliki Oikonomidou for her insight and advice.

Finally, to all the professors in the 2017-2019 ‘Linguistics: Theory and Applications’ postgraduate course, I would like to express my deepest gratitude. I consider myself lucky to have been one of your students. I learnt more than I could have ever hoped for and, more importantly, I learnt how to keep learning. Thank you.

Disclaimer

Although the data were kindly bestowed to me by RCeL, all conclusions and opinions are derived from my subjective understanding of the various notions and, as such, they do not represent the views of RCeL and its affiliates.

Contents

To Georgia Sotiropoulou

Chapter 1

Introduction

Writing assessment in first and second language research is a rich area of study, featuring various research foci and objectives. For example, some longitudinal studies have focused on writing development and explored the degree to which formal instruction can effect change on aspects of writing performance, such as language quality (e.g., Mazgutova & Kormos, 2015), genre awareness (e.g., Yasuda, 2011; Yoon & Polio, 2016) or ideas development, to name just a few. Others have focused on task design and have investigated the cumulative impact of different factors on writing performance in high-stakes language examinations (e.g., Cho et al., 2013; Oikonomidou, 2016). Another strand of researchers has looked into writing quality of rater-judged scripts highlighting language performance (e.g., Aryadoust & Liu, 2015; Casal & Lee, 2019), rater reliability (e.g., Wind, 2019), or both (e.g., Jeong, 2016).

This dissertation has opted for the third approach, zeroing into the assessors' quality judgements in an L2 high-stakes examination context. Using data from the intergraded C level (C1-C2) KPG (i.e. Kratiko Pistopiitiko Glossomathias) language examination suite, an attempt was made, first, to examine the salience of the assessed language criteria, namely Spelling & Punctuation, Language Range, Grammaticality/Accuracy, Appropriacy and Cohesion & Coherence, and, second, to retrieve potential instances of rater severity.

It should be added that the research design of this study derives from the rare opportunity of having been given access to rated scripts. What is more, having two trained raters for the same script(s) is significant for L2 writing research, as has been repeatedly stated in related literature (see for instance Crossley & McNamara, 2012).

The unique design of the KPG writing module targets both writing production and written (interlingual) mediation at C level by incorporating a read-to-write task with an English source text and an also integrated mediation task with a Greek source text. With these in mind, the questions that were formulated are as follows:

RQ1: How does task completion interact with the overall language performance criteria (Spelling & Punctuation, Vocabulary Range, Grammaticality/ Accuracy, Appropriacy and Cohesion & Coherence)?

RQ2: To what extent do rater judgements overlap/vary and where can this be attributed to?

RQ3: Are there any indicators regarding cohesion and coherence which can potentially allow for distinctions between C1 and C2 level candidate performance?

To address these questions, the following sections are organized accordingly. In Chapter 2, Literature Review, previous research is presented linking writing quality with raters' judgements, task types and genre effects. Chapter 3, Methodology, includes a description, along with the statistical affordances and limitations, of the corpora in question. Furthermore, a detailed account of the C level writing module of the KPG language examination is given.

The Results and Analysis section (Chapter 4) is divided into three parts. The first concerns quantitative analysis of the scripts with a focus on textual features. The second employs quantitative analysis with a view to exploring overlap and variation in the raters' judgements. Both of the aforementioned sections include predictive models. The third one stems from the results of the previous sections and singles out Coherence & Cohesion through Coh-Metrix-driven qualitative analysis. In order to accommodate the reader, each of these chapters is followed by analyses, translating numbers into linguistic terms.

In Chapter 5, Discussion and Conclusion, each of the research questions are dealt with separately. They are followed by overall comments, the limitations of this dissertation and further research suggestions.

.

.

Chapter 2

Literature Review

2.1. Introduction

This chapter reports on research relevant to key elements of this study, namely integrated tasks, writing quality indicators and issues related to rater reliability.

2.2. Defining Integrated Tasks

With integrated tasks commonly used in academic contexts, the need for learners to be able to cope in such environments has prompted the use of integrated tasks in language proficiency assessment (cited in Plakans & Gebril, 2017). Plakans (2012: 249) defines integrated tasks as "tasks that require more than one skill for completion," emphasizing the contribution of source text processing reading skills to writing production (Plakans; 2009; Plakans & Gebril, 2012, 2013; Plakans et al., 2019). Knoch and & Sitajalabhorn (2013: 306), after reviewing previous definitions (Ascención Delaney, 2008; Cumming et al., 2005; Plakans, 2009, 2012), and the various conventions employed in integrated tasks in language examinations (reading, listening and graphs), proposed a more comprehensive definition specifically for read-to-write tasks:

> Integrated writing tasks are tasks in which test takers are presented with one or more language- rich source texts and are required to produce written compositions that require (1) mining the source texts for ideas, (2) selecting

> ideas, (3) synthesising ideas from one or more source texts, (4) transforming the language used in the input, (5) organizing ideas and (6) using stylistic conventions such as connecting ideas and acknowledging sources. The rating scale used to grade such compositions needs to take account of these features specific to integrated writing tasks.

In contrast to independent writing tasks which require that the candidate rely mostly on memory to answer the task question (Weigle, 2002), integrated tasks are more taxing, in the sense that more cognitive processes are needed in order to appropriately relay the relevant information and position themselves against them. Cumming et al. (2005) report on a number of differentiating factors between independent and integrated tasks: lexical complexity, syntactic complexity, rhetoric and pragmatics. In fact, their research showed that, across proficiency levels, grammatical accuracy and lexical complexity indices were the most informative.

The cognitive and discoursal differences, to say the least, between integrated and independent writing tasks entail several pedagogical implications. Test designers ought to accommodate candidate needs, all the while adhering to test requirements, teachers should be aware of fine differences between task types and learners need to be properly prepared to tackle such tasks. Actually, Cheong et al. (2019) argue that learner training in reading and employing multiple texts in their L1 can greatly benefit text integration in the L2.

2.3. KPG Integrated Tasks

The data for this study relate to two integrated tasks from the C level writing module of the KPG language examination suite.

The first task (Activity 1) is referred to in literature as a read-to-write task (Oikonomidou, 2016) in which candidates are given a written source text in the target language and they are asked to interact with it in some way in order to produce their own text. It can be claimed that the proposed task features by Knoch and Sitajalabhorn (2013) are in line with those of the KPG. More specifically, they argue (ibid: 304) that the source text needs to display the following features:

> (a) the input material needs to include a significant proportion of language and, directly following from this,
> (b) the task needs to require that the language in the source material is used and transformed to complete the writing task.

Although it is not strictly 'required' that candidates use information from the source text, they do need to understand it and use it in tandem with their own ideas to the extent they feel is appropriate.

What is more, it can be claimed that it is a form of mediation, though 'intralingual', as derived from the Council of Europe, CEFR Companion Volume with New Descriptors (2018: 33):

> Treatment of mediation in the CEFR is not limited to cross-linguistic mediation (passing on information in another language) as can be seen from the following extracts:
>
> ► Section 2.1.3: Make communication possible between persons who are unable, ***for whatever reason***, to communicate with each other directly.
>
> ► Section 4.4: Act as an intermediary between interlocutors who are unable to understand each other directly, normally (***but not exclusively***) speakers of different languages.
>
> ► Section 4.6.6: ***Both*** input and output texts may be spoken or written and ***in L1 or L2***. (*Note: This does not say that one is in L1 and one is in L2; it states they could both be in L1*).

The second task (Activity 2) is an interlingual mediation task, specific only to the KPG examination suite, which instructs candidates to use the Greek source text appropriately to produce another one in English. Abiding by the CEFR notion of a mediation task, it is considered to be a distinct translanguaging practice in an EFL context by Stathopoulou (2015: 37-38), who "sees written [interlingual] mediation as a reading-to-write construct since it involves both reading and understanding a text and writing another" (Stathopoulou, 2019: 418). She moves on to point out that the processes of reading and writing "blend and co-occur" (ibid: 419) and argues that interlingual mediation tasks

> require a reading-to-write ability, which does not involve just reading in order to comprehend, but it implies more than this; mediators' reading is directed towards writing or in other words, they read with a writing goal, i.e. to select and relay information from a Greek source in order to produce another in English. (ibid.: 419).
>
> In this line, to mediate is
>
> to reformulate, to transcode, to alter linguistically and/or semiotically by rephrasing in the same language, by alternating languages, by switching from oral to written expression or vice versa, by changing genres, by combining text and other modes of representation, or by relying on the resources – both human and technical – present in the immediate environment, …" (Coste & Cavali, 2015: 62-63, as cited in Stathopoulou, 2019: 416)

Actually, it appears that the features suggested by Knoch and Sitajalabhorn (2013) are closer to the requirements of the interlingual mediation task, as candidates here are explicitly asked to make use of the source text.

2.4. Lexical, Syntactic, and Cohesive Features and L2 writing quality

Although writing quality is seen as closely connected to content development, audience awareness, as well as discoursal, linguistic and textual features, in the field of writing assessment, it is aligned with "the fit of test-takers' essays to their assessment context, which is usually reflected by scoring rubrics" (Kim & Crossley, 2018: 40).

In this line, this section presents research on the effect of linguistic choices on L2 writing quality. Specifically, it discusses a number of studies employing both integrated and independent task types which have investigated language realization in various genres, among which expository and narrative.

A large amount of research involves cohesion and coherence. The indices of the automated tool Coh-Metrix have been used to investigate global, local and text cohesion, with its connections to lexical diversity, fluency and readability, to mention only a few. Recently, Crossley et al. (2016) looked at such connections and juxtaposed them to Coh-Metrix-calculated and rater-derived writing quality. The results indicated that that human raters capture less variance in candidate performance.

In another seminal work, Crossley and McNamara (2012) argued for the production of more linguistically sophisticated essays by advanced learners instead of more cohesive. More specifically they found that the more advanced the learners the “less frequent, less familiar and less meaningful” (ibid: 17) the words they used. At the same time temporal cohesion and word overlap were not preferred.

Kim and Crossley (2018: 51) synthesize and contribute to relevant research. First, they report on previous studies on lexical, syntactic and cohesive features focusing on a single writing task (Crossley & McNamara, 2012, 2014; Crossley et al., 2016; Lu, 2010) and on others looking at different tasks but with separate statistical analyses (Guo et al., 2013; Kyle & Crossley, 2016). Then they cover the gap in literature by adding an investigation of two different writing tasks on the same criteria. The results again argue for the predominance of lexical sophistication over syntactic complexity and cohesion. Moreover, their findings corroborate previous research, arguing for grammatical and textual competence as the basis for effective L2 writing production across tasks.

Jeong (2016) explored the possibility of an expository and/or narrative genre effect on students’ writing scores. Although no such interaction was found, results differed when looking at the genres across proficiency levels. In fact, expository essays was what advanced students excelled at while beginner level students achieved higher marks in narrative tasks. The expository finding was attributed to wider exposure of the advanced students to the genre and better reading comprehension skills, which were aided by these students ability to interact with complex vocabulary and syntactic complexity abundantly found in expository texts.

In contrast, Kormos (2011, as cited in Oikonomidou, 2016) argued that candidate performance in terms of linguistic and syntactic structures heavily relies on narrative generic requirements

With respect to cohesion and coherence in expository texts, Khahil (1989, as cited in Blani, 2008) noted the consistent use of lexical cohesion by EFL learners, and and the marked lack of use of linking devices by ESL students. To an extent, this is in line with Crossley and McNamara (2012) as mentioned before.

In a comparison between narrative and argumentative writing tasks, Rashid and Rafik-Galea (2007, as cited in Oikonomidou, 2016) found that performance in the narrative tasks exceeded that in the argumentative. To some extent Rezazadeh et al.'s (2011, as cited in Oikonomidou, 2016) research yielded similar findings, with argumentative tasks lacking in fluency and accuracy. Moreover, Yoon (2017), who researched argumentative essays, found that complex ideas connected to phrasal density. Moreover, although he did find lexical and morphological complexity across genres, he was unable to clearly differentiate between adjacent proficiency levels (ibid.: 138)

2.5. Rater Reliability in Assessing Writing Quality

With automated scoring gaining more and more ground in writing assessment (see Weigle, 2013), the call for interrater reliability as an indicator of test validity (see Bachman and Palmer, 1996) becomes all the more essential.

Whether rater agreement or rater severity (see Wind et al., 2017) is the focal point, intra- and inter-rater reliability (test internal validity) are important. In this dissertation only interrater reliability was catered for and, thus, the remaining section revolves around the extent to which rater judgement variation can affect overall scoring outcomes.

Perhaps, what is most contentious in human rating evaluation is the assessment method. Hyland (2003: 226) notes three types of criterion-referenced procedures: holistic, analytic and trait based. Trait based evaluation depends on the requirements of the particular task and the scale is modified accordingly. Nevertheless, it would require the availability of a relevant data bank. Holistic evaluation involves the rater awarding a cumulative mark to a script based on general level requirements. It is supposed to "integrate inherent qualities of writing" (ibid: 227). However, this practice has been attacked for being too impressionistic (see Xie, 2015: 23), while a major disadvantage is that it does not afford precise and adequate feedback to learners (Hyland, 2003: 227).

Analytic evaluation is based on different scales of writing quality indicators. In this practice, raters are equipped with more information about the weight of the rating criteria, allowing for discriminations to be made even between weaker texts (ibid: 229). Humphry and Heldinger (2019: 3) report on the provision of more diagnostic information which "better enables practitioners to tailor instruction more closely to the needs of their students and to provide feedback to the students." This latter scale (i.e. analytic) is also employed in the KPG examination battery, which this study focuses on.

Critics of analytic evaluation methods express concerns regarding the 'halo effect', in which raters are influenced by the scoring of one scale when marking another (Hyland, 2003: 229). This is one of the many factors that have been found to affect rater writing quality judgements. Goodwin (2016) notes contrast effects, in which raters are influenced by what they have already rated. Wind (2019) reports on inconsistent use of the rating scale and the scripts' textual characteristics (Wind et al., 2017). Humphry and Heldinger (2019) underline the possibility for construct-irrelevant variance if certain assessing criteria are judged as unrelated to the task. Xie (2015) goes as far as to look at the test-takers' perspective and the strategies they employ to manage rater impressions.

In terms of subjectivity, the raters' multifaceted background has drawn a lot of attention in related research (e.g., see Gebril & Plakans, 2014 and Goodwin, 2016). Specifically, in combination with rater expectations and interpretations, raters' world knowledge is said to affect the effectiveness of discourse features (Banerjee et al., 2011; McNamara, Crossley & McCarthy, 2010). Johnson and VanBrackle (2011) have even identified linguistic discrimination against African American Errors (AAE) and in favour of English as a Second Language (ESL) errors. Finally, the already strenuous problem-solving activity that is rating (Deremer, 1998), is noted to be even more complex and demanding when scoring integrated tasks. Characteristically, Gebril and Plakans (2014), report on more judgement than interpretation strategies being implemented by raters in such tasks.

When all is said and done, what else is there? Although improving the rating scale could be a valid suggestion (e.g., Banerjee et al., 2015), most researchers would agree that another worthwhile solution is rater training. Especially about integrated tasks, with which raters may find themselves to be "out of their comfort zone" (Gebril & Plakans, 2014: 66), it is noted that rater severity and rater inconsistency can be dectreased with rater training (ibid: 58). What is more, it has been proposed that raters can be more easily trained to rate analytic instead of holistic scales (Dunsmuir et al., 2005, as cited in Humphry & Heldinger, 2019).

In the KPG examination suite, where, as mentioned before, analytic marking grids are used, many layers of measures are taken to ensure inter-rater reliability. A marked example, with respect to the writing module, is the script rater training programme. It is a 5-stage[2] system that comes into operation right after the exam is administered and before the scripts are made available to the raters. It continues throughout the marking period and ceases with feedback evaluations forms from both coordinators and script raters. Here it should be added that raters do not signal their choices in the scripts and, therefore, cannot influence the other raters' decisions (Hartzoulakis, 2010: 236).

[2] See https://rcel2.enl.uoa.gr/kpg/gr_script_train.htm for more information.

Apart from carefully selecting and training expert raters, and performing classical item analysis, the Rasch model is also employed to investigate test paper validity (Dendrinos & Karavas, 2013: 113). The use of this tool with the view to testing scores for various facets, including rater severity, is commonly found in relevant literature (e.g., Aryadoust & Liu, 2015; Goodwin, 2016; see also Wind et al., 2017).

2.6. Conclusion

In sum, this section has been presented typical characteristics of integrated tasks and the particular way they map on the KPG C level writing tasks. Along with the very brief overview of characteristic findings in L2 writing quality research and rater reliability it is hoped that the reader has been prepared for the subsequent focal points of this dissertation.

Chapter 3

Methodology

This dissertation combines statistical analysis and corpus-driven techniques in order to conduct quantitative and qualitative research of English KPG candidates' scripts at C level (C1 and C2) of language proficiency The overall aims of the study have been to identify and highlight the extent to which

a. different tasks affect candidate performance and
b. raters mark the same tasks differently.

The data for this study have been provided by RCeL, the Research Centre for Language Teaching, Testing and Assessment, of the English Department of the National and Kapodistrian University of Athens, which is responsible for the design of the of the KPG English test.

The KPG English test, and the tests in the other languages as well, adheres to a functional theory of language, which according to Dendrinos & Karavas (2013: 16-17):

> is understood as social practice, and sets out to assess how candidates use the target language to create socially purposeful meanings rather than whether they have a wide range of vocabulary and a firm knowledge of the formal properties of the language in question. Exams aim at measuring candidates' ability to comprehend and produce oral or written discourse and, more specifically, the extent to which candidates can:
>
> - understand messages in different types of oral and written texts
> - make language choices that indicate language awareness and one's ability to negotiate socially situated meanings
> - produce context-appropriate speech and writing

- act as mediators and, from B1 level onwards, extract information from a Greek text so as to relay it in the target language either orally or in writing.

From the above it becomes clear that the KPG writing module places emphasis on genres and on candidates' ability to function as mediators. In fact, the introduction of mediation activities in the two production modules (writing and speaking) constitute one of KPG's innovation features (Karavas & Mitsikopoulou, 2018). Another innovative feature of the KPG is that it assesses candidates' abilities through intergraded exams for A, B or C levels. This practically means that candidates sit for a single C level exam which assesses both C1 and C2 level competences and candidates can be awarded with a C1 or C2 certificate, based on their performance. The intergraded C level writing examination, specifically, consists of two activities, Activity 1 and Activity 2, targeting writing production and intrerlingual mediation respectively.

In the KPG Activity 1, candidates are given a written source text in the target language (English in our case), whose understanding is presupposed in order to produce a generically different one in L2 (Dendrinos & Karavas, 2013: 102). Activity 2 is an interlingual mediation task in which candidates are asked to relay information from a Greek source text given to them into English.

Both writing activities at C level in the KPG exam can be considered integrated in the sense that they both employ a source text and require candidates to interact with it. One difference between the two activities is that Activity 1, the read-to-write task, allows/ expects the use of candidates' personal ideas, as well as relevant information they can draw from the source text. In contrast, in Activity 2, candidates are explicitly instructed to use the information in the source text and are expected to relay all the necessary information in an appropriate way. This last requirement also distinguishes the C level KPG mediation task from the respective B level.

Unlike the KPG B level writing examination where the writing activities are divided by language level (2 activities for B1 and 2 activities for B2), the C level writing module with its two activities is internally graded for language level performance (KPG Script Rater Guide, 2017: 6). This means that it is the questions within the tasks that correspond to different levels and not the individual tasks

In terms of length, the two source texts can range from 400 to 550 words (Dendrinos & Karavas, 2013: 103). However, there is 15% of variation, not necessarily for text word count, allowed from one examination period to the next to protect the batteries from excessive standardization (ibid: 54).

3.1. Data

The data used in this study include candidates' scripts and their associated marks for the C level intergraded KPG examination of English language proficiency (C1-C2) administered in December 2017. A random sample of thirty-three candidates' scripts were selected, which were rated by the same two raters, the profile of whom is not known. Overall, sixty-six scripts, thirty-three for each activity[3],

[3] In this dissertation the terms 'activity' and 'task' are used interchangeably.

containing answers to the two corresponding activities, a read-to-write and a mediation task (Table 1.). They were all transcripted into Word Document format without altering the initial texts in any way. The only thing that could not be controlled for was the space between words.

Table 1. The corpus

Writing Module	*Number of texts*	*Number of words*
Activity 1	33	11.420
Activity 2	33	9.928
Total	66	22.348

Although the final grade for the writing module or the overall test performance of these candidates is not known, the 'scoring' (see 3.2.2 for elaboration on the choice of term) of the individual writing paper scripts is available. In particular, assessment is based on two general criteria, task completion and overall language performance. Candidates are awarded a score for each one of the two activities for Task Completion separately. They are also awarded 5 different marks for their overall language performance in both activities (Activity 1 and Activity 2) in Spelling & Punctuation, Vocabulary Range, Grammaticality/Accuracy, Appropriacy and Cohesion & Coherence. Taking into account that each one of the scripts was rated by two different raters, the researcher had available 7 marks for each candidate from each rater (overall 14 different marks from the two raters). Following is an example of the available marks for one candidate (Table 2.).

Table 2: Marks for candidate 116

	Rater 1	*Rater 2*
	Task Completion	
Activity 1	3	3
Activity 2	4	3
	Overall Language Performance	
Spelling & Punctuation	3	2
Vocabulary Range	4	4
Grammaticality/ Accuracy	4	3
Appropriacy	3	3
Cohesion & Coherence	4	3

3.1.1 Writing Activities

In this corpus, in Activity 1, candidates were given an English poem to produce an expository personal blog entry and, in Activity 2, an online magazine article to transform into an encyclopedic entry with both narrative and expository features. Following are the task instructions (retrieved from https://www.minedu.gov.gr/themata-kpg). See Appendices A.1. and A.2 for the complete writing exam (task instructions and source texts).

Activity 1:

Read an extract of a poem written by an actor –not a poet. Charlie Chaplin, famous from the Silent Movie era, wrote it on his 70th birthday, 16 April 1959. You like the poem and decide to write **an entry** (about 350 words) in your **personal blog**, **explaining**:

- what this poem means for you

- what it could mean to parents and teachers (who have to help young people become well-adjusted adults)

Activity 2:

Using information from the text below, write an entry (about 300 words) for an electronic encyclopaedia:

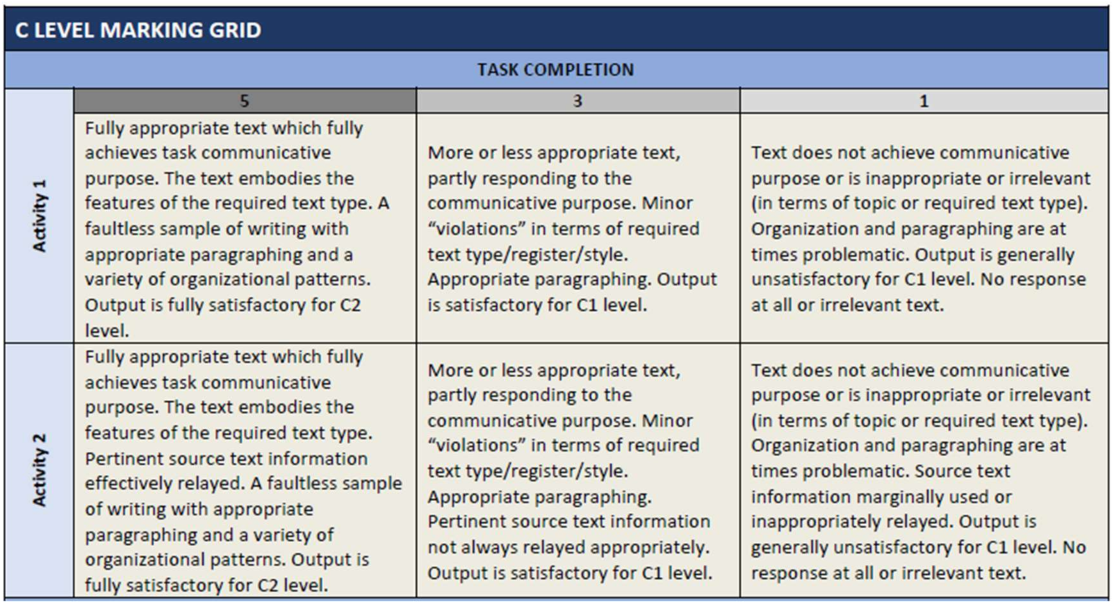
- **Provide factual information** about Dimitris Nanopoulos
- **Explain** why he counts as one of the “famous Greeks”

It is understood that the first requirement in each activity targets C1 level of language proficiency and the second targets C2.

3.1.2 KPG Marking scheme

The marking grid which is used for assessing KPG C level scripts is reproduced below in Figures 1 and 2. A variation of this marking grid can be found online as well (https://rcel2.enl.uoa.gr/kpg/gr_script_train.htm), the only difference being that since the latest update (May 2014) a change has occurred. To be precise, there used to be a sixth language criterion in overall language performance, text organization, which has now been subsumed under Task Completion. In every other respect, C level has been marked consistently since its first administration.

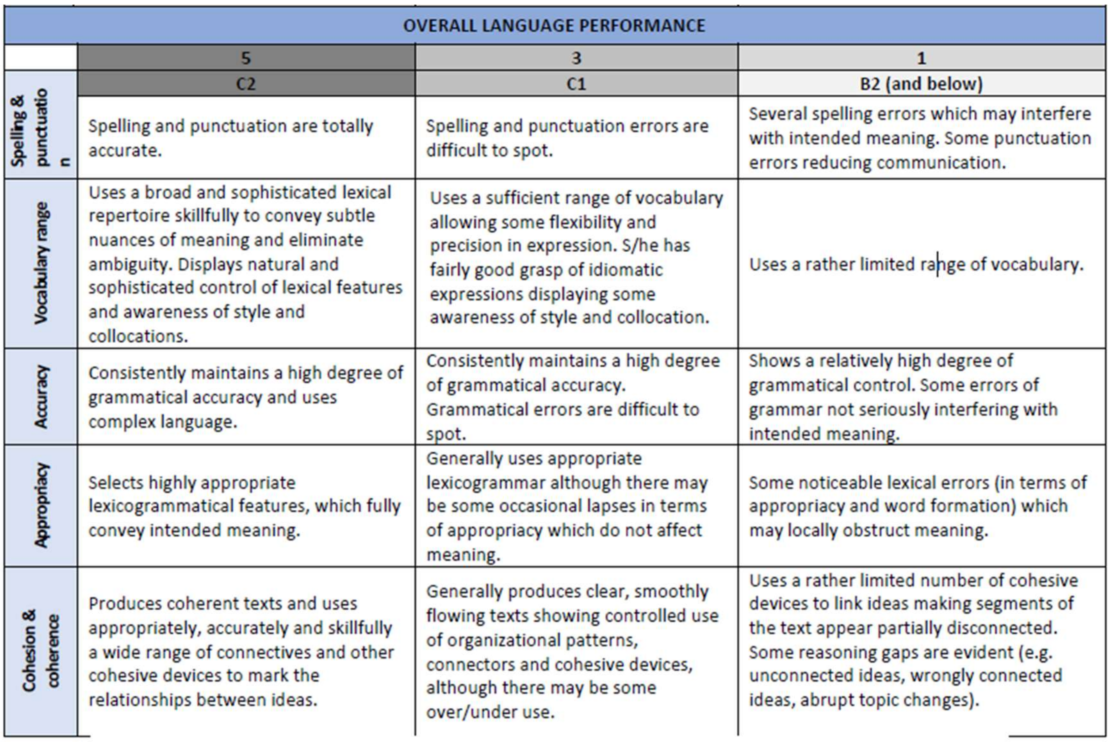
Figure 1. C Level Marking Grid, Task Completion



C LEVEL MARKING GRID

TASK COMPLETION			
	5	3	1
Activity 1	Fully appropriate text which fully achieves task communicative purpose. The text embodies the features of the required text type. A faultless sample of writing with appropriate paragraphing and a variety of organizational patterns. Output is fully satisfactory for C2 level.	More or less appropriate text, partly responding to the communicative purpose. Minor “violations” in terms of required text type/register/style. Appropriate paragraphing. Output is satisfactory for C1 level.	Text does not achieve communicative purpose or is inappropriate or irrelevant (in terms of topic or required text type). Organization and paragraphing are at times problematic. Output is generally unsatisfactory for C1 level. No response at all or irrelevant text.
Activity 2	Fully appropriate text which fully achieves task communicative purpose. The text embodies the features of the required text type. Pertinent source text information effectively relayed. A faultless sample of writing with appropriate paragraphing and a variety of organizational patterns. Output is fully satisfactory for C2 level.	More or less appropriate text, partly responding to the communicative purpose. Minor “violations” in terms of required text type/register/style. Appropriate paragraphing. Pertinent source text information not always relayed appropriately. Output is satisfactory for C1 level.	Text does not achieve communicative purpose or is inappropriate or irrelevant (in terms of topic or required text type). Organization and paragraphing are at times problematic. Source text information marginally used or inappropriately relayed. Output is generally unsatisfactory for C1 level. No response at all or irrelevant text.

Figure 2. C Level Marking Grid, Overall Language Performance



OVERALL LANGUAGE PERFORMANCE			
	5	3	1
	C2	C1	B2 (and below)
Spelling & punctuatio n	Spelling and punctuation are totally accurate.	Spelling and punctuation errors are difficult to spot.	Several spelling errors which may interfere with intended meaning. Some punctuation errors reducing communication.
Vocabulary range	Uses a broad and sophisticated lexical repertoire skillfully to convey subtle nuances of meaning and eliminate ambiguity. Displays natural and sophisticated control of lexical features and awareness of style and collocations.	Uses a sufficient range of vocabulary allowing some flexibility and precision in expression. S/he has fairly good grasp of idiomatic expressions displaying some awareness of style and collocation.	Uses a rather limited range of vocabulary.
Accuracy	Consistently maintains a high degree of grammatical accuracy and uses complex language.	Consistently maintains a high degree of grammatical accuracy. Grammatical errors are difficult to spot.	Shows a relatively high degree of grammatical control. Some errors of grammar not seriously interfering with intended meaning.
Appropriacy	Selects highly appropriate lexicogrammatical features, which fully convey intended meaning.	Generally uses appropriate lexicogrammar although there may be some occasional lapses in terms of appropriacy which do not affect meaning.	Some noticeable lexical errors (in terms of appropriacy and word formation) which may locally obstruct meaning.
Cohesion & coherence	Produces coherent texts and uses appropriately, accurately and skillfully a wide range of connectives and other cohesive devices to mark the relationships between ideas.	Generally produces clear, smoothly flowing texts showing controlled use of organizational patterns, connectors and cohesive devices, although there may be some over/under use.	Uses a rather limited number of cohesive devices to link ideas making segments of the text appear partially disconnected. Some reasoning gaps are evident (e.g. unconnected ideas, wrongly connected ideas, abrupt topic changes).

The KPG Script Rater Guide (2017: 7) also offers a word of caution regarding the numbers 1-5 of the Likert scale used, namely that they are to be treated as values, not scores or grades. It is also underlined that each variable has a different gravity status depending on language level (ibid). This is one of this dissertation's limitations. Without inside knowledge as to the exact way these values are assessed, any attempt at generalization of predictive models is futile. Nevertheless, the current research findings could be used by KPG stake holders to confirm, or question, the existing marking scheme. What is more, the predominance, for example, of one language criterion over another can be compared to relevant literature findings, regardless of the unknown final grade(s), mainly because this research focuses on perceived and actual use of practice, not on certification comparability.

This study is in agreement with the guidelines set by the KPG, where C1 level performance corresponds to a ’3’ while C2 corresponds to a ‘5’. However, with the aforementioned limitations in mind and expecting 2-digit values when calculating the means, the 1-5 Likert scale was used rather loosely for KPG standards, though consistently throughout this thesis. See Table 3. for the researcher’s initiatives in translating the given values into language level equivalents.

Table 3. Dissertation-specific correspondence of values to language levels based on the C level marking grid

Values	***1***	***1,5***	***2***	***2,5***	***3***	***3,5***	***4***	***4,5***	***5***
Language Level	B2-	B2	B2+	C1-	C1	C1+	C2-	C2	C2+

3.2 Research Questions

This study has been based on exploring the following three main research questions and a number of sub-questions for the first two. The conducted analysis in Chapter 4, which is also presented in 3.3. below, has been designed along the lines of these questions.

Research Question 1: How does task completion interact with the overall language performance criteria (Spelling & Punctuation, Vocabulary Range, Grammaticality/ Accuracy, Appropriacy and Cohesion & Coherence)?

- RQ1a: How do the task completion means (for both activities and for each one separately)correlate with those for overall language performance?
- RQ1b: Is there significant statistical difference between Activity 1 and Activity 2 cumulative task completion and overall language performance marks? Where can differences between C1 and C2 level of language proficiency be traced?
- RQ1c: What predictive model(s) can explain most of the variation in the sample?

Research Question 2: To what extent do rater judgements overlap/vary and where can this be attributed to?

- RQ2a: How do Rater 1 and Rater 2 quality judgements correlate with one another and with their cumulative means for task completion and overall language performance?
- RQ2b: Are interrater quality judgements different to a statistically significant degree?
- RQ2c: What predictive models can explain most of the variation in each rater's marking profile? Do these map on cumulative predictive models?

Research Question 3: Are there any indicators regarding cohesion and coherence which can potentially allow for distinctions between C1 and C2 level candidate performance?

3.3 Methods of Data Analysis

Two types of analysis were implemented, quantitative and qualitative. The quantitative analysis consisted of two parts, one focusing on the candidates' performance and the other on interrater variation, using correlations, various t-tests, ANOVAs and regressions. Based on the results of the quantitative analysis, the study proceeded to qualitatively analyze the Cohesion & Coherence language criterion, which proved to be prevalent in all calculations.

The first step, after the transcription of all scripts, was to transfer all the 66-scripts awarded marks to a Microsoft Excel sheet (2010). A preliminary statistical analysis included calculating the means for all marks which resulted in thirty categories as depicted in Table 4. All calculations were made using the respective formulas.

Table 4. Data categories for the 66-script corpus

Label	***Description***	***Label***	***Description***
Candidate	Number	GRAMACCR1	Grammaticality/ Accuracy Rater 1
WCA1	Word Count Activity 1	GRAMACCR2	Grammaticality/ Accuracy Rater 2
WCA2	Word Count Activity 2	Mean3	Grammaticality/ Accuracy Mean
A1R1	Activity 1, Rater 1	APPR1	Appropriacy Rater 1
A1R2	Activity 1, Rater 2	APPR2	Appropriacy Rater 2
MEANA1	Task Completion mean Activity 1	Mean4	Appropriacy Mean
A2R1	Activity 2, Rater 1	COHR1	Cohesion & Coherence Rater 1
A2R2	Activity 2, Rater 2	COHR2	Cohesion & Coherence Rater 2

MEANA2	Task Completion mean Activity 2	Mean5	Cohesion & Coherence Mean
SPPUNCTR1	Spelling & Punctuation Rater 1	R1TC	Rater 1 Task Completion Mean
SPPUNCTR2	Spelling & Punctuation Rater 2	R2TC	Rater 2 Task Completion Mean
Mean1	Spelling & Punctuation Mean	R1OLP	Rater 1 Overall Language Performance Mean
VRR1	Vocabulary Range Rater 1	R2OLP	Rater 2 Overall Language Performance Mean
VRR2	Vocabulary Range Rater 2	MTC	Task Completion Mean
Mean2	Vocabulary Range Mean	MOLP	Overall Language Performance Mean

The results of this preliminary analysis were then used for further statistical analysis with Microsoft SPSS, version .20 and .25. Correlations targeted inter-rater reliability and underlying connections among marking criteria, which in turn fueled the computation of various t-tests (paired-samples and one-sample) along with t-test based analyses of variance (ANOVAs). The objective was to pinpoint the differences, or narrow down the area where the variance originated, between raters, activities and language criteria. Finally, regressions were implemented to combine small-range findings so as to explore the possibility of deriving wider-range effects.

Specifically, in terms of scripts analysis, correlations, a paired-samples t-test, ANOVAs and regressions were employed. For the interrater comparison, correlations, a one-sample t-test and regressions were selected.

For the qualitative analysis a sub corpus was extracted from the initial 66-scripts corpus to retrieve qualitative information focusing on the Coherence and

Cohesion language criterion. Its purpose was to identify differences and potential indicators of variance within C level of language proficiency (C1-C2). To do so, eight candidates' scripts (16 scripts in total) were chosen based on the awarded scores. To further explain, two sub-corpora were further created, one for those scripts that had been awarded with a score of 5 by Rater 1 and 3 by Rater 2, and one for those scripts awarded with a score of 3 by both. The null hypothesis was that if the difference of opinion on the raters' part was inconsequential, no significant differences would be found, at least at a surface level, which the researcher believed to be easier for raters to identify. One drawback of this corpus design is the limited number of scripts. However, this was the most that could be done since only 33 test-takers' scripts were surveyed. Another point is that no automated qualitative distinction between activities can be made, with each rater giving a cumulative score for both activities. As a result, the analysis of findings was done with both activities in mind.

The automatic analytic tool that was used for the two sub-corpora was Coh-Metrix 3.0, created on September 1st, 2012, and last updated on August 16th, 2017 (http://tool.cohmetrix.com/). Coh-Metrix offers 106 indices measuring various levels of cohesion and coherence. More specifically the indices are categorized as: Descriptive, Text Easability Principle Component Scores, Referential cohesion, LSA (Latent Semantic Analysis), Lexical Diversity, Connectives, Situation Model, Syntactic Complexity, Syntactic Pattern Density, Word Information and Readability. Further information on relevant criteria can be found in the corresponding Results and Analysis section (4.3.2).

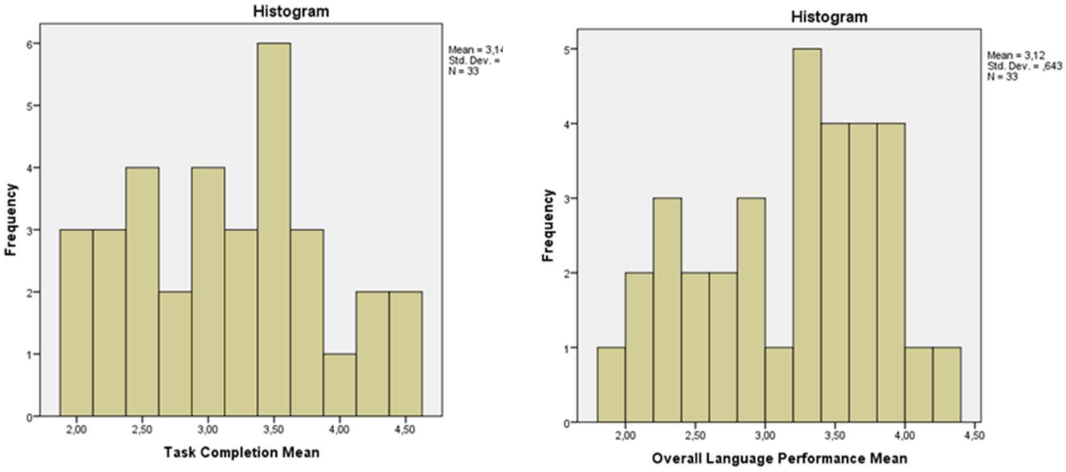
3.4 Statistical analysis and corpora characteristics.

With validity issues in mind, various facets such as normality of the sample, skewness and kyrtosis, collinearity and multicollinearity were addressed. Starting with sample size and distribution, Field (2009: 42) notes that the larger the sample, the more normal the distribution is. Since 'large' in this field corresponds to a measurement of N>30, this research sample, with 33 candidates and 66 scripts, can be considered borderline. In fact, both skewness and kyrtosis are observed in certain variables (see Appendices B.1.for Descriptives). However, descriptive analysis on said variables in juxtaposition to means and standard deviations shows that it is kyrtosis that slightly affects this sample, but without significantly impacting on its overall normal distribution. Table 5. illustrates the descriptive statistics of the means for cumulative task completion and overall language performance and Figure 3 their respective plot distribution. There is no skewness since its statistic is smaller than its standard error, but there is negative kyrtosis with both, indicating a sample with not many high peaks, high values in this case. Nevertheless, the Shapiro-Wilk normality test did not yield significant results, which suggests that the kyrtosis does not affect the normality of the sample distribution. Nonetheless, it should be noted that significance was found by the Kolmogorov-Smirnov test for Overall Language Performance, *p*= .038. It was not taken into consideration, though, as research into normality tests has found that the Shapiro-Wilk test is stronger than the Kolmogorov-Smirnov, although not in very small samples (Razali & Wah, 2011). The aforementioned results enhance confidence about the validity of the paired-samples t-tests and the ANOVAs.

Table 5. Normality check for Task Completion and Overall Language Performance.

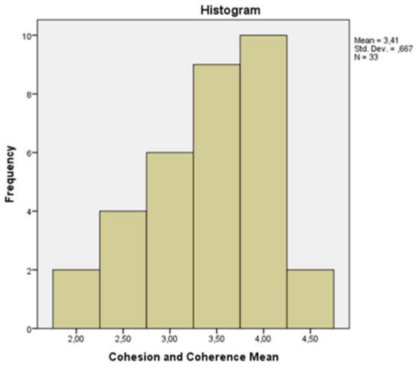
Index	*Mean*	*Mean Std Error*	*Skewness*	*Skewness Std Error*	*Kyrtosis*	*Kyrtosis Std Error*	*Shapiro-Wilk*	*Shapiro-Wilk Sig.*
MTC	3,14	,18	, 121	,409	-,858	,798	,956	,202
MOLP	3,12	,11	-,326	,409	-,889	,798	,955	,185
Mean5	3,41	,12	-, 488	,409	-,473	,798	,920	,018

Figure 3. Plot distribution for Task Completion and Overall Language Performance



On the other hand, the same cannot be claimed for the one-sample t-test employed in the qualitative analysis. The source sample size, the Coherence & Cohesion mean, is not normal, with Shapiro-Wilk attesting to it. In particular, Figure 4. shows that there is skewness, which is supported by the fact that its skewness value is larger than its standard error. Another drawback is that the size of the two designed sub-corpora is, even statistically, too small. With these in mind, it was decided to attend in detail only to the highly significant results, those with *p*< .005, expecting some degree of representativeness against the overall results.

Figure 4. Plot distribution for Cohesion & Coherence mean



With respect to the reliability of the regression analyses, the point of interest is collinearity, an effect created when one variable strongly correlates with another, and likewise, multicollinearity, when several variables are taken into consideration concurrently (Field, 2009: 223). As it is in detail explained in the respective Results and Analysis sections, this was circumvented to a large degree by decreasing the number of variables per calculation. However, since some degree of multicollinearity could not be avoided, other measures such as confidence intervals and tolerance levels were also examined, resulting in acceptable configurations.

Chapter 4

Results and Analysis

This chapter consists of three main parts and presents the findings of the conducted research. The first part focuses on candidate performance analyzing the means of marks trying to identify significant assessment criteria for different activities. Similarly, the second one focuses on the raters in order to pinpoint potential instances of statistically significant variation. The third part relates qualitative analysis findings regarding the Cohesion & Coherence language criterion and, in particular, possible indicators of differentiation between C1 and C2 level of language proficiency.

4.1 Quantitative script comparison

This section deals with quantitative analysis of the awarded marks for task completion and overall language performance.

Main Research Question 1: How does task completion interact with the overall language performance criteria (Spelling & Punctuation, Vocabulary Range, Grammaticality/ Accuracy, Appropriacy and Cohesion & Coherence)?

Sub-Research Questions:

RQ1a: How do the task completion means (for both activities and for each one separately)correlate with those for overall language performance?

RQ1b: Is there significant statistical difference between Activity 1 and Activity 2 cumulative task completion and overall language performance marks? Where can differences between C1 and C2 level of language proficiency be traced?

RQ1c: What predictive model(s) can explain most of the variation in the sample?

4.1.1 RQ1a- Correlations

A Pearson product moment correlation coefficient was computed to assess the relationships between the task completion mean for both of the activities (MTC), and their individual ones (MEANA1and MEANA2), and those awarded for overall language performance, in sum (MOLC) and separately (Mean1, Mean2, Mean3, Mean4 and Mean5) (Table 6.).

Table 6. Script marks comparison, Correlations

Variable Groupings		*r*	*p*
MTC	MOLP	.846	.000
MEANA1	Mean1	.387	.026
MEANA1	Mean2	.850	.000
MEANA1	Mean3	.634	.000
MEANA1	Mean4	.738	.000
MEANA1	Mean5	.803	.000
MEANA2	Mean1	.471	.006
MEANA2	Mean2	.659	.000
MEANA2	Mean3	.754	.000
MEANA2	Mean4	.759	.000
MEANA2	Mean5	.757	.000

1.1 MTC and MOLP

There was a very strong positive correlation *r*(33) = .85, *p*<.001 between MTC (task completion mean) and MOLP (overall language performance mean).

1.2. MEANA1, Mean1, Mean2, Mean3, Mean4 and Mean5

Activity 1 (MEANA1) weakly correlated with the mean for Spelling & Punctuation (Mean1) *r*(33) = .39, *p*= .026, while there was moderate correlation with the mean for Grammaticality/ Accuracy (Mean3) *r*(33) = .63, *p*< .001 and with that for Appropriacy (Mean4) *r*(33) = .74, *p*< .001. There was also a high degree of correlation with the mean for Vocabulary Range (Mean2) *r*(33) = .85, *p*< .001 and that for Cohesion & Coherence (Mean 5) *r*(33) = .80, *p*< .001. All correlations were positive.

1.3. MEANA2, Mean1, Mean2, Mean3, Mean4 and Mean5

Activity 2 (MEANA2) moderately correlated with the mean for Spelling & Punctuation (Mean1) *r*(33) = .47, *p*= .006. In addition, there was a strong correlation with the mean for Vocabulary Range (Mean2) *r*(33) = .66, *p*= .006, Grammaticality/ Accuracy (Mean3) *r*(33) = .75, *p*< .001, Appropriacy (Mean4) *r*(33) = .76, *p*< .001, and Cohesion & Coherence (Mean 5) *r*(33) = .76, *p*< .001. All correlations were positive.

4.1.2 Analysis

What has been found so far is that there is an effect between task completion and overall language performance marks (MTC and MOLP). This was to be expected as they are the sides of the same coin; they stand for the candidates' performance in Module 2 of the KPG C level language proficiency examination and reflect the exam designers' decisions about what should be given an evaluative mark. If these are run up against the descriptive marking rubric for C level they can shed some light about what is taken for granted and what is required at this level. However, at this point, it would be ill-advised to assume, for example, that Spelling & Punctuation is inconsequential for Activity 1 and thus candidates need not pay that much attention to it.

One interpretation could be that there is little difference between B and C level of language proficiency in terms of Spelling & Punctuation, with C level presupposing knowledge expected at B level, but not significantly adding to it. From another perspective, test-takers do not usually target higher levels until they have first secured previous ones. If Spelling & Punctuation holds strong for previous levels, it is only to be expected that candidates sitting a C level examination, would not have much difficulty with it. If the two previous hypotheses are combined, it could be claimed that Spelling & Punctuation does not, or should not, possess the same predictive strength as the other criteria at C level.

Moreover, despite the fact that these results are indicative of interaction among variables, there is still the issue of causality (Field, 2009: 173). There is not enough information to rule out the effect of other variables being at play concurrently (third-variable problem), or enough to say with conviction that it is this variable and not the other that causes the correlation (direction of causality). In this case this affects our view of the interaction between task requirements and language. Is it indeed the task that informs the linguistic choices and thus situated performance, as affirmed by the KPG developers, or is the unaffected by communicative context test takers' linguistic competence that affords for the same quality performance no matter the task? To address these questions more analysis is required using ANOVAs and Regressions.

4.1.3 RQ1b- Paired-samples t-tests and ANOVAs

1. Paired-samples t-tests

A paired-samples t-test comparing the Activity 1 cumulative task completion marks (MEANA1) with those of Activity 2 (MEANA2) did not yield any statistically significant results. However, when comparing the cumulative overall language performance marks, meaning script rater 1's and script rater 2's marks for both of the rated scripts, R1OLP and R2OLP respectively, the scores were significantly higher for R1OLP (*M*= 3.28, *SD*= .75) than for R2OLP (*M*= 2.95, *SD*= .64), *t*(32)= 3.30, *p*= .002, *d*= .57. Cohen's *d* indicates a moderate effect size.

2. ANOVAs

A series of one-way ANOVAs (Table 7.) were run to determine whether cumulative and individual task completion marks differed significantly in terms of overall language performance and, similarly, whether cumulative and individual overall language performance marks differed significantly in terms of task completion marks.

Table 7. Script marks comparison, ANOVAs

Dependent Variables	*Independent Variables*	*SS*	*MS*	*F*	*p*
MTC	MOLP	13.821	.813	3.687	.007
MOLP	MTC	11.182	1.118	12.014	.000
MEANA1	Mean1	4.918	.820	1.362	.267
MEANA1	Mean2	15.219	3.044	15.385	.000
MEANA1	Mean3	8.978	1.796	4.185	.006

MEANA1	Mean4	12.007	3.320	9.826	.000
MEANA1	Mean5	14.718	2.944	13.602	.000
MEANA2	Mean1	8.408	1.401	2.852	.029
MEANA2	Mean2	9.882	1.976	4.722	.003
MEANA2	Mean3	12.759	2.552	8.180	.000
MEANA2	Mean4	13.278	3.320	11.760	.000
MEANA2	Mean5	14.061	2.812	10.663	.000
Mean1	MEANA1	4.382	.876	1.467	.233
Mean2	MEANA1	16.255	3.251	25.169	.000
Mean3	MEANA1	10.174	2.035	5.336	.002
Mean4	MEANA1	6.796	1.359	7.420	.000
Mean5	MEANA1	10.085	2.017	13.146	.000
Mean1	MEANA2	10.473	2.095	5.632	.001
Mean2	MEANA2	11.597	2.319	7.688	.000
Mean3	MEANA2	12.813	2.563	9.037	.000
Mean4	MEANA2	8.086	1.617	11.943	.000
Mean5	MEANA2	10.081	2.016	13.131	.000

In some cases post hoc analysis could not be performed because at least one group had fewer than two cases. This is one of the limitations of this dissertation, which could be averted in future similar research by increasing the number of scripts.

1. MTC by MOLP

With respect to MTC, an analysis of variance showed that the effect of MOLP was significant at the $p< .05$ level, $F(17,15) = 3.69$, $p= .007$. Post hoc analysis could not be performed because at least one group had fewer than two cases.

2. MOLP by MTC

MTC had a greater effect on MOLP, $F(10,22) =12.01$, $p< .001$. As above, post hoc analysis could not be performed because at least one group had fewer than two cases.

3. MEANA1 and MEANA2 by Mean1

An analysis of variance only found a significant effect of Spelling & Punctuation (Mean1) on Activity 2 (MEANA2) at *p*< .05 level*, F*(6, 26)= 2.85*, p*= .029. No post hoc analysis could be performed.

4. MEANA1 and MEANA2 by Mean2

An analysis of variance showed that Vocabulary Range (Mean2) had a significant effect on both Activity 1 (MEANA1), *F*(5, 27)= 15.38*, p*< .001, and a weaker one on Activity 2 (MEANA2), *F*(5, 27)= 4.72*, p*= .003.

As regards Activity 1, post hoc analysis using Scheffe indicated that mean mark 2.00 (*M*= 2.00, *SD*= .00) was significantly different from that of 4.00 (*M*= 3.80, *SD*= .54), *p*< .001 and 4.50, *p*< .001. Mean mark 2.50 (*M*= 2.33, *SD*= .29) significantly differed from mean mark 4.00, *p*= .002, and 4.50*, p*= .001, while mean mark 3.00 (*M*= 3.10, *SD*= .22) was also significantly different from 4.00, *p*= .006, and 4.50, *p*= .003. No other mean combinations were significant for Activity 1. No significant mean combinations were found in Activity 2 either.

5. MEANA1 and MEANA2 by Mean3

An analysis of variance showed that the effect of Grammaticality/ Accuracy (Mean3) was significant on both Activity 1 (MEANA1) and Activity 2 (MEANA2) at *p*< .05: *F*(5, 27)= 4.18*, p*= .006 for the former and *F*(5, 27)= 8.18*, p*< .001 for the latter.

Scheffe post hoc analysis indicated that no mean mark combination was significantly different from the others for Activity 1. In Activity 2, however, there was enough evidence to suggest that mean mark 3.50 (*M*=3.87, *SD*=.58) significantly differed from mean mark 1.00 (*M*= 2.25, *SD*= .35*), p*= .04, 1.50 (*M*= 2.30, SD= .67), *p*= .003 and 2.00 (*M*= 2.40, *SD*= .42), *p*= .005.

6. MEANA1 and MEANA2 by Mean4

An analysis of variance showed that the effect of Appropriacy (Mean4) was significant at *p*< .05 on Activity 1 (MEANA1), *F*(4, 28)= 9.83, *p*< .001, and on Activity 2 (MEANA2), *F*(4, 28)= 11.76*, p*< .001.

Scheffe post hoc analysis for Activity 1 indicated that mean mark 2.00 (*M*= 2.21, *SD*= .27) was significantly different from mean mark 3.00 (*M*=3.50, *SD*= .71), *p*= .001, mean mark 3.50 (*M*= 3.75, *SD*= .64), *p*= .004, and mean mark 4.00 (*M*= 4.17, *SD*= .58), *p*= .001. Regarding Activity 2 similar difference was found between the same groups: mean mark 2.00 (*M*=2.14, *SD*= .24) with 3.00 (*M*= 3.35, *SD*= .51), *p*= .002, 3.50 (*M*= 3.88, *SD*= .75), *p*= .001, and 4.00 (*M*= 3.83, *SD*= .58), *p*= .003. Moreover, mean mark 2.50 (*M*= 2.50, *SD*= .24) was significantly different from 3.50, *p*= .011, and 4.00*, p*= .029.

7. MEANA1 and MEANA2 by Mean5

An analysis of variance showed that the effect of Cohesion & Coherence (Mean5) was significant on both Activity 1 (MEANA1) and Activity 2 (MEAN2) at *p*< .05: *F*(5, 27)= 13.60*, p*< .001 for the former and *F*(5, 27)= 10.66, *p*< .001 for the latter.

Regarding Activity 1, Scheffe post hoc analysis indicated that mean mark 4.00 (*M*= 4.05, *SD*= .55) was significantly different from mean mark 2.00 (*M*= 2.00, *SD*= .00), *p*< .001, 2.50 (*M*= 2.37, *SD*, .48), *p*< .001, 3.00 (*M*= 2.91, *SD*= .38), *p*= .004 and

3.50 (*M*= 3.05, *SD*= .39), *p*= .005. Additionally, mean mark 4.50 (*M*= 4.00, *SD*= .71) was similarly different from mean mark 2.00, *p*= .011, and 2.50, *p* < .020.

As for Activity 2, Scheffe post hoc analysis indicated that there is a significant difference between mean mark 3.50 (*M*=3.33, *SD*= .50) and those of 2.50 (*M*= 2.12, *SD*= .25 *p*= .026, and 3.00 (*M*= 2.33, *SD*= .41), *p*= .040. Mean mark 4.00 (*M*= 3.70, *SD*= .67) was also significantly different from 2.00 (*M*= 2.00, *SD*= .00), *p*= .012, 2.50, *p*= .001, and 3.00, *p*= .002.

8. Mean1, Mean2, Mean3, Mean4, Mean5 by MEANA1

An analysis of variance showed that, except Spelling & Punctuation (Mean1), the task completion Mean for Activity 1 (MEANA1) did have a significant effect on the remaining four overall language performance means at *p*< .05 level: Vocabulary Range (Mean2) *F*(5, 27)= 25.17, *p*< .001, Grammaticality/ Accuracy (Mean3) *F*(5, 27)= 5.34, *p*= .002, Appropriacy (Mean4) *F*(5, 27)= 7.42, *p*< .001 and Cohesion & Coherence (Mean5) *F*(5, 27)= 13.15, *p*< .001.

With respect to Vocabulary Range, Scheffe post hoc analyses indicated that mean mark 2.00 (*M*= 2.12, *SD*= .25) was significantly different from those of 3.00 (*M*= 3.50, *SD*= .50), 3.50 (*M*= 4.00, *SD*= .29), 4.00 (*M*= 4.25, *SD*= .35) and 4.50 (*M*= 4.25, *SD*= .27) at *p*< .001. Furthermore, mean mark 2.50 (*M*= 2.80, *SD*= .27) differed significantly from 3.50, *p*< .001, 4.00, *p*= .003, and 4.50, *p*< .001. Lastly, 3.00 was significantly different from 4.50, *p*= .023.

Grammaticality/ Accuracy Scheffe post hoc analysis indicated that mean mark 2.00 (*M*= 1.25, *SD*= .29) significantly differed from 3.00 (*M*= 2.61, *SD*= .74), *p*= .042, 3.50 (*M*= 2.64, *SD*= .69), *p*= .049 and 4.50 (*M*= 3.25, *SD*= .42), *p*= .002.

Appropriacy Scheffe post hoc analysis indicated that mean mark 2.00 (*M*= 2.00, *SD*= .00) was significantly different from those of 3.50 (*M*= 3.14, *SD*= .47), *p*= .012, and 4.50 (*M*= 3.41, *SD*= .49), *p*= .002. Similar difference was found between 2.50 (*M*= 2.40, *SD*= .55) and 4.50, *p*= .025.

Finally, with respect to Cohesion & Coherence, mean mark 2.00 (*M*= 2.25, *SD*= .29) was found to be significantly different from 3.00 (*M*= 3.28, *SD*= .44), *p*= .010, 3.50 (*M*= 3.71, *SD*= .49), *p*< .001, 4.00 (*M*= 4.00, *SD*= .00), *p*= .002 and 4.50 (*M*= 4.08, *SD*= *.20*), *p*< .001. Mean mark 4.50 was significantly different from 2.50 (*M*= 3.10, *SD*= .42*), p*= .016, and 3.00*, p*= .026, as well.

9. Mean1, Mean2, Mean3, Mean4, Mean5 by MEANA2

An analysis of variance showed that all overall language performance means had a significant effect at *p*< .05 level: Spelling & Punctuation (Mean1) *F*(5, 27)= 5.63, *p*= .001, Vocabulary Range (Mean2) *F*(5, 27)= 7.69, *p*< .001, Grammaticality/ Accuracy (Mean3) *F*(5, 27) = 9.04, *p*< .001, Appropriacy (Mean4) *F*(5, 27)= 11.94, *p*< .001 and Cohesion & Coherence (Mean5) *F*(5, 27)= 13.13, *p*< .001.

With regard to Spelling & Punctuation, Scheffe post hoc analyses indicated that mean mark 3.00 (*M*= 3.94, *SD*= .73) was significantly different from those of 2.00 (*M*= 2.69, *SD*= .46*), p*= .017, and 2.50 (*M*= 2.37, *SD*= .48), *p*= .014.

Vocabulary Range Scefffe post hoc analyses indicated that mean mark 2.00 (*M*= 2.69, *SD*= .53) was significantly different from those of 3.00 (*M*= 3.81, *SD*= .59), *p*= .017, and 3.50 (*M*= 4.17, *SD*= .26), *p*= .002. Furthermore, mean mark 3.50 significantly differed from 2.50 (*M*= 2.88, *SD*= .75), *p*= .045.

As regards Grammaticality/ Accuracy, Scheffe post hoc analysis indicated that mean mark 2.00 (*M*= 1.69, *SD*= .46) significantly differed from 3.00 (*M*= 2.75, *SD*= .46), *p*= .022, 3.50 (*M*= 3.00, *SD*= .77), *p*= .006, 4.00 (*M*= 3.00, *SD*= .41), *p*= .020, and 4.50 (*M*= 3.50, *SD*= .00), *p*= .002. Moreover, mean mark 4.50 was significantly different from 2.50 (*M*= 1,88, *SD*= .63), *p*= .022.

Appropriacy Scheffe post hoc analysis indicated that mean mark 2.00 (*M*= 2.19, *SD*= .26) was significantly different from those of 3.00 (*M*= 3.00, *SD*= .27), *p*= .009, 3.50 (*M*= 3.33, *SD*= .60), *p*< .001, 4.00 (*M*= 3.00, *SD*= .00), *p*= .048 and 4.50 (*M*= 3.67, *SD*= *.29*), *p*< .001. Similar difference was found between 2.50 (*M*= 2.37, *SD*= .48) and 3.50, *p*= .020, and 4.50, *p*= .006.

Lastly, regarding Cohesion & Coherence, mean mark 2.00 (*M*= 2.56, *SD*= .42) was found to be significantly different from 3.00 (*M*= 3.69, *SD*= .37), *p*< .001, 3.50 (*M*= 3,92, *SD*= .49), *p*< .001, 4.00 (*M*= 3.75, *SD*= .029), *p*= .003 and 4.50 (*M*= 4.00, *SD*= .00), *p*= .001. In addition, mean mark 2.50 (*M*= 3.00, *SD*= .41) was significantly different from 3.50, *p*= .046.

4.1.4 Analysis

1. Paired t-tests

The fact that there was no indication of a statistically significant difference between the task completion means for Activity 1 and Activity 2, but there was one between the overall language performance marks which allows for a number of explanations. As for the former, it could be that either the marking criteria are more clear-cut, or that the script raters, who are experienced and well-trained, can follow the marking rubric closely. Of course, both of these could apply concurrently. Either way, these results point towards a continuously valid language examination scheme.

In this line, although it would be easy to claim rater bias or question the validity of the marking rubric for the statistically significant difference in judging overall language performance, there are other task-related avenues to explore as well. Looking at the writing examination as a whole and investigating the interactions between task completion and overall language performance might reveal other factors worth considering in tandem. Detailed comparison between script raters will be explored in 4.2.

2. ANOVAs

1(MTC by MOLP) and 2 (MOLP by MTC) analyses of variance of cumulative mean scores for task completion and overall language performance (MTC and MOLP) show that they both have an effect on one another, but because the analyses are incomplete (due to the small number of scripts) we still cannot make any distinctions between combinations of marks and, thus, between levels of language proficiency. It is interesting to note, though, that the *F* value for MTC is three times that for MOLP, suggesting that the effect of task requirements on language is greater. In other words, it would appear that in the functional theoretical framework informing this exam suite, a top-down approach (from task to language) is favoured over a bottom-up (from language to task). In this sense, text types and text processes bring to the surface appropriate linguistic criteria respectively, attesting to the prevalence of communicative purpose and situational context in this language assessment, all the while highlighting the need for better understanding of the symbiotic relationship between genre and language. As Hyland (2004: 1-2) puts it,

> [i]t is through genres that individuals develop relationships, establish communities, and achieve their goals. Without the familiar structure that genres give to social events, we would be unable to conduct the most basic interactions of everyday life.

It is also validating, in the sense that what is shown is that it is the task that drives the language, the communicative purpose, when engaging in a social practice, and not grammar or vocabulary, for example, which are in turn thought of as means to an end in the genre-based approach of the KPG (KPG Script Rater Guide, 2017, 4).

From 3 (MEANA1 and MEANA2 by Mean1) to 7 (MEANA1and MEANA2 by Mean5) the weaker, bottom-up approach was examined, meaning the impact of the choice and accuracy of linguistic resources on task completion, an interaction which has been found to be very strong correlation-wise (*r*(33)= .85, *p*<.001). Statistically significant differences suggest that these results were not random and can be generalized in larger similar groups. What is more, because this corpus, though adequate, is rather small, 66 scripts, 33 scripts per activity, non-significant findings are so at least for this corpus, without excluding the possibility of being significant if more scripts are examined, (Field, 2009).

With these in mind, as far as Activity 1 is concerned, it is observed that it is significantly affected by Vocabulary Range, Grammaticality/ Accuracy, Appropriacy and Cohesion & Coherence. Particularly, in this corpus, the raters were able to draw a distinct line between B and C level scripts, wherever such differences occur across language criteria, and even differentiate within C level in terms of Vocabulary Range (C1 VS $C2^{-}$) and Cohesion & Coherence (C1/ $C1^{+}$ VS $C2^{-}$).

Likewise, Activity 2, which was found to be significantly affected by all of the language criteria, also allowed for distinctions between B and C level, but only Cohesion & Coherence yielded a significant difference within C level (C1 VS $C1^{+}$ and C1 VS $C2^{-}$)

At this point, another interpretation which can be put on these findings, is that Vocabulary Range and Cohesion & Coherence are significant C level criteria because of their connection to the specific text processes required. In Activity 1, both C1 and C2 questions involved the text process of ‘explaining’, thus being a fully expository task. In Activity 2, the narrative C1 question required that the candidates ‘give factual information’ and the expository C2 was as above to ‘explain’. Based on their common ground, it can be claimed that expository questions are more demanding with respect to the use of cohesive devices than narrative texts. Qualitative research using Coh-Metrix could shed some more light into potential differences, an attempt at which is included in this dissertation (4.3.).

As for Vocabulary Range, the interpretation can be two-fold. On the one hand, in Activity 1 candidates were provided with substantial vocabulary from the source text, which they were to use accordingly so as to explain their view, first, at a personal level (C1) and, second, at a societal level by extending it to other wider social groups (C2). It follows from this that the instantiation of more complex vocabulary means for the second aspect, could have allowed for distinctions to be made. Actually, Wind et al (2017: 11), argued that rater judgements can be influenced by textual characteristics of student compositions (expository essays) both with L1 and L2 writers, with the latter registering more variation. Even when automated scoring is performed, the lexical diversity values of the target text has been found to be greatly influence by the that of the source text (Gebril & Plakans, 2016).

On the other hand, the interlingual mediation task, where the source text is in Greek, might have applied equal strain on the candidates in tackling the narrative and expository aspects within the task, which is supported by relevant research. Jeong (2016), for example, found that there was a genre effect across proficiency levels and more specifically that more advanced lexical competence was fostered in the expository essays rather in the narrative texts employed in their research. Nevertheless, unlike Jeong (2016), who examined two different text processes (narrative and expository) in two different essays, the two text processes in Activity 2 co-existed in the encyclopedic entry, forming yet another hybrid genre. In this sense, the difference could be attributed to internal conflict between processes and the candidates' inability to equally attend to them, resulting in yielding to narrativity (the C1 task requirement) and displaying less linguistic sophistication.

8 (Mean1, Mean2, Mean3, Mean4, Mean5 by MEANA1) and 9 (Mean1, Mean2, Mean3, Mean4, Mean5 by MEANA2) investigated the stronger, top-down interaction between task completion and overall language assessment. The results are similar to those of the previously presented reverse investigation. Spelling and Punctuation is not a significant factor for Activity 1 and the vast majority of groupings involve distinguishing between B and C level.

As far as Spelling & Punctuation is concerned, results suggest a similar interpretation to the one in the Correlations section (4.2.2.). Nevertheless, the fact that it was indeed significant for Activity 2 can be explained by the task requirements. As mentioned before, Activity 2 is a cross-language mediation task involving the transformation of the content of a Greek text into another one of different type in L2. In this light, the discrepancy can be explained either by differences in punctuation between the two languages, or too much of inaccurate spelling because the candidates could not use the L1 source text to that end, or both.

As for the other language criteria, Vocabulary Range and Cohesion & Coherence retain their standing among them, allowing for distinctions between C1 and C2 level, although only in Activity 1 and with an observed longer distance between scales (see Methodology 3.1.2., Table 3. for the dissertation-specific adaptation of the C level marking scale).

Putting everything together, it could be claimed that, in this corpus, quality judgements between B and C level were the most common, with some exceptions regarding variance within C level. These results can also be interpreted in terms of genre effect and rater bias combined. Activity 1 had a significant impact on the expository decisions made by the candidates, but Activity 2, by incorporating both narrativity and exposition, nullified the predominance of the one over the other, according to rater judgements.

What is more, it can be claimed that the bottom-up approach allows for more finely grained linguistic differences to come to the surface than the top-down. In particular, with the adjusted scale allowing for 8 moves (1 move= 0,5), the language to task design effect could be pinpointed even down to 1-move difference and across the scale from 1(B1$^-$) to 4,5 (C2), whereas the task to language condition required at least 2 moves (with the exception of the effect of Activity 2 on Spelling & Punctuation) and ranged between 2 (B2) and 4,5 (C2). Again, it is not unlikely that with a bigger corpus more, or more precise, differences could come to light.

All these findings corroborate what is stated in the KPG Script Rater Guide (2017: 6), that

> when meanings are produced with more or less appropriate language use –i.e., language which follows rules of language required by the context of situation and the genre— grammar errors that do not create problems of intelligibility are considered unimportant during the process of evaluation and marking of candidate scripts.

The achievement of the communicative purpose trumps that of language realization, as reflected by the performed cross-checked analyses of variance.

4.1.5 RQ1c- Regressions

A series of multiple regressions were conducted to investigate the predictive strength of variable combinations both in terms of task completion (MEANA1, MEANA2 and MTC) and overall language performance (Mean1, Mean2, Mean3, Mean4, Mean5 and MOLP). Each regression design was checked for collinearity or multicollinearity, using the Variance Inflation Factors (IVF) and their respective Tolerance levels. No problematic cases were encountered when two variables were combined. However, on account of certain means being strongly correlated with others, some potential multicollinearity problems could arise with designs with more than two variables. Nevertheless, only two cases (see designs 1.3. and 1.4., this section) had two means with a VIF < .20 and tolerance> 5.0. Therefore, and with the values of unacceptable levels varying among experts or fields, these designs were not discarded as non-generalizable, since their values fall within acceptable ranges, VIF>.1or VIF>.2 and Tolerance <10 (Field, 2009: 224). See Table 8. for exact regression values.

Table 8. Script marks comparison, Regressions

Dependent Variables	***Independent Variables***	***r***	***r^2***	***B***	***SE***	***β***
MTC	MOLP	.846	.716	.963	.109	.846
MOLP	MTC	.846	.716	.744	.084	.846
MEANA1		.875	.766			
	Mean1	.387		-.178	.123	-.178
	Mean2	.850		.688	.212	.674
	Mean3	.634		-.254	.185	-.254
	Mean4	.738		.337	.317	.254
	Mean5	.803		.352	.287	.293
MEANA2		.808	.654			
	Mean1	.471		-.025	.152	-.025
	Mean2	.659		-.241	.262	-.233
	Mean3	.754		.370	.229	.364
	Mean4	.759		.332	.391	.247
	Mean5	.757		.574	.355	.471

MEANA1	Mean2	.850	.723	.868	.097	.850
MEANA1		.857	.735			
	Mean2	.850		.655	.204	.642
	Mean5	.803		.284	.240	.236
MEANA2		.797	.635			
	Mean3	.754		.422	.185	.414
	Mean5	.757		.521	.222	.427
MEANA2		.787	.620			
	Mean3	.754		.399	.216	.392
	Mean4	.759		.573	.285	.427
MEANA2		.783	.613			
	Mean4	.759		.558	.314	.415
	Mean5	.757		.480	.285	.394

B:unstandardized Beta; *SE*: standard error unstandardized Beta; β: standardized Beta

In order to accommodate the reader, the chosen final descriptive designs are provided as headings for each subsection. It should be pointed out that a number of regressions were calculated, but only the following were chosen since these were thought to be the most predictive.

*1. MTC= .141 + (.963*MOLP)*

The results indicated that the model explained 71.6% of the variance and that it was a significant predictor of the Mean for Task Completion (MTC), *F*(1, 31)= 78.20, *p*< .001. The contributing factor was the Mean for Overall Language Performance (MOLP) (*b*= .963, *p*< .001).

*2. MOLP= .780 + (.744*MTC)*

The results indicated that the model explained 71.6% of the variance and that it was a significant predictor of the Mean for Overall Language Performance (MOLP), *F*(1, 31)= 78.20, *p*< .001. The contributing factor was the Mean for Task Completion (MTC) (*b*= .744, *p*= .007).

3. *MEANA1= -.105 + (-.178*Mean1)+ (.688*Mean2)+ (-.254*Mean3)+ (.337*Mean4)+ (.352*Mean5)*

This linear regression yielded a significant model, *F*(5, 27)= 17.67, *p*< .001, which explained 76,6% of variation. One contributing factor was found to be statistically significant, Vocabulary Range (Mean2) *(b*= .69, *p*= .003). The remaining variables (see Table 6.) were not significant predictors.

4. *MEANA2= .134 + (-.025*Mean1)+ (-.241*Mean2)+ (.370*Mean3)+ (.332*Mean4)+ (.574*Mean5)*

Similarly, a linear regression yielded significant results for MEANA2, *F*(5, 27)= 10.19, *p*< .001. The model explained 65.4% of variation. However, none of the variables was found to be a significant contributing factor.

5. *MEANA1= .193 + (.868*Mean2)*

A linear regression was calculated to investigate whether Vocabulary Range is a good predictor of MEANA1. The design was significant, *F*(1, 31)= 80,83*, p*< .001 and explained 72.3% of the observed variance.

6. *MEANA1= -.027+ (.655*Mean2)+ (.284*Mean5)*

This model was significant, *F*(2, 30)= 41.62, *p*< .001, and explained 73.5% of variance. The contributing factor was Vocabulary Range (*b*=.655, *p*= .003), unlike Cohesion & Coherence (*b*= .284*, p*=.247)

7. *MEANA2= .204+ (.422*Mean3)+ (.521*Mean5)*

In this linear regression, *F(*2, 30)= 26.14, *p*< .001, it was found that 63.5% of variance in MEANA2 can be predicted by Grammaticality/Accuracy (*b*=.422, *p*= .030) and Cohesion & Coherence (*b*= .521, *p*= .026)

*8. MEANA2= .403 + (.399*Mean3)+ (.573*Mean4)*

This design explained significantly explained 62% of variance, *F*(2, 30)= 24.46, *p*< .001. Neither Grammaticality/Accuracy (*b*= .399, *p*= .075) nor Appropriacy (*b*=.573, *p*= .054) were significant contributors.

*9. MEANA2= -.181 + (.558*Mean4)+ (.480*Mean5)*

This linear regression model explained 61, 3% of variance and was significant, *F*(2, 30)= 23.80, *p*< .001. Again, neither of the variables, Apporpriacy *(b*= .558, *p*= .085) and Cohesion & Coherence (*b*= .480, *p*= .102), were significantly contributing factors.

4.1.6 Analysis

Designs 1. (MTC= .141 + (.963*MOLP)) and 2. (MOLP= .780 + (.744*MTC)) are able to explain a large amount of variance in this corpus. The fact that the task completion mean can explain such a large percentage of variance of the overall language performance, and vice-versa, highlights how closely-knit the relationship between the text as a whole and its language is, with ‘text’ treated in the KPG context as “as the material configuration of various aspects of the communicative context in which language functions” (https://rcel2.enl.uoa.gr/kpg/research_ler2.htm). Moreover, the largely similar results could stand as a good argument for task completion to hold the same gravity status as all the language criteria combined.

Some statistical reservations lie in that these two variables strongly correlate with one another, but this is not the case when the language criteria are examined against the two constituents of MTC, MEANA1 (task completion Mean for Activity 1) and MEANA2 (task completion Mean for Activity 2) (Table 6.), a course of investigation which was followed thereafter.

The results of regression 3. (MEANA1= -.105 + (-.178*Mean1)+ (.688*Mean2)+ (-.254*Mean3)+ (.337*Mean4)+ (.352*Mean5) beg the question of what the exact design parameters of (this) Activity 1, the integrated read-to-write task, are that allow, or target, sophisticated use of language on the part of the candidates in order to attain a C level language proficiency certification. From a genre perspective, it could be the expository requirements, as discussed previously (see analysis of ANOVAs in 4.1.4). Additionally, it could be argued that in intralingual mediation of a message, what is necessary is flexibility in the use of a wide range of vocabulary choices in order to relay it. The use of vocabulary supplied in the source text alone is not enough. What is also required is transformation of said vocabulary with task requirements in mind (Knoch & Sitajalabhorn, 2013). Source text inspection, which showed that candidates were provided with a large amount of sophisticated vocabulary, raises the possibility that (over-)dependence on said vocabulary items might have influenced the respective marks.

As far as Activity 2 is concerned (4. MEANA2= .134 + (-.025*Mean1)+ (-.241*Mean2)+ (.370*Mean3)+ (.332*Mean4)+ (.574*Mean5)), it is understood that not finding significantly contributing variables does not mean that these were not significant but that they might not have been significantly different from each other in effecting a change on Activity 2. This was corroborated by the implementation of regression designs juxtaposing both MEANA2 and MEANA1 (where a significant difference had been found) to each of the other variables separately (see Table 9.). Results showed that for MEANA2 the difference between Grammaticality/ Accuracy (Mean3), Appropriacy (Mean4) and Cohesion & Coherence (Mean5) was almost negligible, while Vocabulary Range (Mean2) came very close to their percentages. The Spelling & Punctuation (Mean1) design, although significant, was somewhat apart from the rest. With respect to MEANA1, the second most predictive design was with Mean5.

Table 9. Task completion means against individual language criteria, Regressions

Dependent Variables	***Independent Variables***	***r***	***r^2***	***B***	***SE***	***β***	***F***	***p***
MEANA1	Mean1	.387	.150	.387	.166	.387	5.459	.026
	Mean2	.850	.723	.868	.097	.850	80.832	.000
	Mean3	.634	.402	.635	.139	.634	20.841	.000
	Mean4	.738	.544	.976	.160	.738	37.008	.000
	Mean5	.803	.644	.965	.129	.803	56.124	.000
MEANA2	Mean1	.471	.222	.479	.161	.471	8.837	.006
	Mean2	.659	.434	.683	.140	.659	23.804	.000
	Mean3	.754	.569	.767	.120	.754	40.896	.000
	Mean4	.759	.577	1.020	.157	.759	42.244	.000
	Mean5	.757	.573	.923	.143	.757	41.535	.000

B:unstandardized Beta; *SE*: standard error unstandardized Beta; *β*: standardized Beta

With validation issues in mind, the above designs informed the following combinations of variables.

5. (MEANA1= .193 + (.868*Mean2)) and 6. (MEANA1= -.027+ (.655*Mean2)+ (.284*Mean5)show the overwhelming influence of Vocabulary Range on Activity 1, since even by adding the second most predictive variable to the model, very little difference is observed. In fact, only when three other variables were combined was there a similar predictive strength to that of Vocabulary Range on its own, e.g. MEANA1= .065+ (-.183*Mean1)+ (.329*Mean4)+ (.834*Mean5), *F*(3, 29)= 19, 57, *p*< .001, r^2 =. 669.

As noted before, regarding Activity 2, the predictive model in 4.(MEANA2= .134 + (-.025*Mean1)+ (-.241*Mean2)+ (.370*Mean3)+ (.332*Mean4)+ (.574*Mean5) did not yield any statistically significant contributors. Although this can be true without nullifying the model's strength, the last three models, namely 7. (MEANA2= .204+ (.422*Mean3)+ (.521*Mean5), 8. (MEANA2= .403 + (.399*Mean3)+ (.573*Mean4)) and 9. (MEANA2= -.181 + (.558*Mean4)+ (.480*Mean5)), allow for some conclusions to be drawn. Again the spotlight falls on Activity 2, the mediation task, and its requirements. The most predictive of all three is that of Grammaticality/ Accuracy and Cohesion & Coherence, with Grammaticality/ Accuracy and Appropriacy second and Appropriacy with Cohesion & Coherence third. The inference to be made here is that these three, and especially Grammaticality/ Accuracy, are very important in tackling the cross-language mediation task.

This is not that surprising since candidates, apart from a number of cognitive strategies that they need to apply, they also need to transform their message from

Greek to English. In doing so, they are also expending effort on fighting interference from their L1 as (negative) transfer is inextricably linked to sentence processing based on the Competition Model (Bates & MacWhinney, 1982; MacWhinney, 2002), a bilingual sentence processing model.

Even the Shallow Structure Hypothesis (SSH) (Clahsen & Felser, 2006a; 2006b; 2017), which among bilingual sentence processing models is not a proponent of transfer, it is argued that "L1 transfer [should] influence L2 processing only indirectly, as a consequence of one or more of the knowledge sources that feed the processing system being affected by properties of the L1" (Clahsen & Felser, 2006b:118). This is considered to be relevant to the interlingual mediation task especially at this level where the source text is lengthier.

Another contributing reason could be the familiarity level of the source text's text-type and, particularly, conventions employed within, which could affect interference not only in terms of sentence processing or text production, but also on a pragmatic level. For example, in this corpus the mediation task required retrieving information from an online magazine article, whose sensationalism-oriented mechanisms triggered affective and/or (anti)patriotic reactions from most of the candidates in the corpus, although these were inappropriate for the target text, an objective encyclopedic entry. For instance, the source text showcasing Dimitris Nanopoulos career in high energy physics read (see Appendices, A.2.):

> Ωστόσο, όταν δόθηκε η ευκαιρία να επιστρέψει στην Ελλάδα ως καθηγητής στο Πανεπιστήμιο Αθηνών, η εκλογή του καταψηφίστηκε κατά πλειοψηφία.

> [However, when the chance arose for him to return to Greece as a professor in the University of Athens, he was blackballed by the majority.]

One of the many examples in the corpus showing the effect of the use of affective language related to patriotism was the following (see Appendices, A.5):

> "It's depressing the fact that in country where philosophy and democracy were born, we refuse to accept minds like these because they are better than us."

Coming back to the regression findings, it could be claimed that they are inconsistent with KPG previous research conducted by Stathopoulou (2009). While researching B level mediation (Greek source text to English target text) and English cue tasks (English source and target text), she found that it was semantic errors that candidates struggled with in mediated writing production, while syntax was found to be more challenging for the English cue task. Nevertheless, there is the issue of comparability between samples. The B level examination differs from C level in terms of task design and requirements in both tasks. For example, although at B level candidates are only given 'cues' to activate engagement with the task and are not required to incorporate information in any way, this is not the case with C level writing production activity, as it has been presented earlier. Therefore, if the results are not directly comparable, it can be accepted that lexis and syntax are important for both levels but they affect a different activity across these levels.

Another variable that could be responsible for the different findings might be the time which has elapsed since Stathopoulou's (2009) research. During this time the face validity of the KPG examinations has increased, more research has been done and publicized, more educational material has been designed and put to use by the KPG and publishing companies and more teachers have been educated in mediation practices. In other words, it could be claimed that test-takers now are more knowledgeable, thus it could be easier for them to avoid making vocabulary mistakes, for example, by focusing on mediating and not translating the source text. Similarly, more candidates know not to inappropriately transfer genre characteristics of the source text to the target text and, in this way, avoid modelling their syntactic choices to that of the source text.

Either way, the findings of this research are based on candidate performance as perceived by KPG raters and so are those in Stathopoulou (2009), at least for the first phase when sorting through satisfactory and moderately satisfactory scripts. This renders script-rater comparison necessary, which is the focus of the following section.

4.1.7 Summary

What the correlations mainly indicated was that there is a strong connection between the criteria of task completion and overall language performance, and that the language criteria mapped on the two integrated tasks to varying degrees.

From the paired-samples t-test it was inferred that in terms of communicative purpose candidates were judged to have performed similarly in both of the activities, although the raters' quality judgements diverged with respect to language. This was further corroborated by the ANOVA and regression computations.

The ANOVAs suggested that the task exerts a stronger influence on the language used to implement it, while the multiple bidirectional analyses of variance identified Vocabulary Range and Cohesion & Coherence to be in a position to afford differences within C level to come to light. This is especially true for the read-to-write task, while the mediation task was only found to provide such answers only for Cohesion & Coherence and even that only when the effect of language on the activity was calculated.

The regression designs that followed confirmed the predictive strength of Vocabulary Range, both on its own and in combination with Cohesion & Coherence, for the intralingual mediation activity. As for the interlingual mediation task, combinations of Grammaticality/ Accuracy with either Appropriacy or Cohesion & Coherence were the strongest predictors, along with the combined force of the latter two previously mentioned variables.

4.2. Quantitative Analysis: Interrater Comparison

This section focuses on the script raters' quality judgements and employs quantitative measures. The relevant research questions are as follows:

Main Research Question 2: To what extent do rater judgements overlap/vary and where can this be attributed to?

Sub-Research Questions:

RQ2a: How do Rater 1 and Rater 2 quality judgements correlate with one another and with their cumulative means for task completion and overall language performance?

RQ2b: Are interrater quality judgements different to a statistically significant degree?

RQ2c: What predictive models can explain most of the variation in each rater's marking profile? Do these map on cumulative predictive models?

4.2.1. RQ2a- Correlations

A Pearson product moment correlation coefficient was calculated to assess the relationships between the means of the script raters' quality judgements (R1TC, R2TC, R1OLP, R2OLP), the task completion mean for both of the activities (MTC) and those for overall language performance (MOLP) (see Table 10.).

Table 10. Correlations within and between script raters and against cumulative means.

Variable Groupings		*r*	*p*
R1TC	R2TC	.705	.000
R1OLP	R2OLP	.672	.000
R1TC	R1OLP	.866	.000
R2TC	R2OLP	.770	.000
R1TC	MTC	.922	.000
R2TC	MTC	.925	.000
R1OLP	MOLP	.930	.000
R2OLP	MOLP	.898	.000

1. *R1TC & R2TC , R1OLP & R2OLP*

Rater 1 task completion mean (R1TC) had a strong positive correlation to that of Rater 2 (R2TC), *r*(33)= .70, *p*< .001. Similarly, Rater 1 overall language performance mean (R1OLP) strongly correlated to that of Rater 2 (R2OLP), *r*(33)= .67*, p*< .001.

2. *R1TC and R1OLP, R2TC and R2OLP*

Moreover, R1TC had a very strong positive correlation to its respective R1OLP, *r*(33)= .87, *p*< .001, while Rater 2 task completion mean (R2TC) strongly correlated to Rater 2 overall language performance mean (R2OLP), *r*(33)= .77, *p*< .001.

3. R1TC and MTC, R1TC and MOLP, R2TC and MOLP

Lastly, R1TC very strongly correlated with both the overall mean for task completion (MTC), *r*(33)= .92, *p*< .001, and the mean for overall language performance (MOLP), *r*(33)= .83, *p*< .001. In addition, R2TC very strongly correlated with MTC, *r*(33)= .92, *p*< .001, and a strong positive correlation was observed with MOLP as well, *r*(33)= .73, *p*< .001.

4.2.2 Analysis

The above results, as others before (4.1.1.), can attest to the internal validity of this examination battery, as there is evident correlation of the raters' marks both in terms of task completion and overall language performance. Moreover, there is indication that Rater 2 is less influenced by language criteria when marking for task completion (see 1. and 2. this section).

4.2.3. RQ2b- Paired-samples t-tests for Rater 1 and Rater 2 Comparison

Multiple paired-samples *t*-tests (Table 11.) indicated that regarding overall language performance the scores of Rater 1 (*M*= 3.20, *SD*= .80) were significantly higher than those of Rater 2 (*M*= 2.95, *SD*= .64), *t*(32)= 3.30, *p*= .002, *d*= .57. With respect to specific language performance criteria, Spelling and Punctuation scores for Rater 1 (*M*= 3.52, *SD*= .83) were significantly higher than Rater 2 (*M*= 3.06, *SD*= .97), *t*(32)= 3.14, *p*= .004, *d*= .55. Grammaticality/ Accuracy followed in the same line with Rater

1 (*M*= 2.67, *SD*= .92) scores being significantly different from Rater 2 (*M*= 2.39, *SD*= .83), *t*(32)= 2.18 , *p*= .037, *d*= .38. Finally, with respect to Cohesion and Coherence, Rater 1 (*M*= 3.79, *SD*= .86) scores were significantly higher than those of Rater 2 (*M*= 3.03, *SD*= .68), *t(*32)= 5.50, *p*< .001, *d*= .95.

Table 11. Paired-samples t-tests

Pairs	***r***	***Mean***	***Standard Deviation***	***95% CI of the difference***	***t***	***df***	***p***	***d***
R1OLP-R2OLP	.672	.33333	.58023	.13, .54	3.300	32	.002	.57
SPPUNCTR1-SPPUNCTR2	.581	.455	.833	.159, .750	3.136	32	.004	.55
GRAMACCR1-GRAMACCR2	.668	.273	.719	.018, .528	2.179	32	.037	.38
COHR1-COHR2	.491	.758	.792	.477, 1.038	5.496	32	.000	.95

Interrater comparison using paired-samples *t*-tests did not yield any statistically significant differences between the Rater 1 task completion mean (R1TC) and that of Rater 2 (R2TC), or between two overall language performance means, namely Vocabulary Range (VRR1 and VRR2) and Appropriacy (APPR1 and APPR2). Likewise, no significant differences were found between them with respect to marks individually awarded for Activity 1 and Activity 2 separately (A1R1 and A1R2; A2R1 and A2R2).

4.2.4. Analysis

The above results indicate that the means of the two raters' quality judgements converge when marking both writing activities for Task Completion, Vocabulary Range and Appropriacy, but diverge in assessing Spelling & Punctuation, Grammaticality/ Accuracy and Coherence & Cohesion, with the last one registering the strongest significance and the largest effect size.

In section 4.1.3 the analyses of variance indicated that judgements between B and C level were prevalent throughout the language criteria, Spelling & Punctuation and Grammaticality/ Accuracy included. However, there were no statistically significant indications of any differences within C level. Taking these into consideration along with the results of the present section, certain inferences can be made.

According to the marking grid for C level (see Fig. 1 and Fig. 2.), what changes between levels as regards Spelling & Punctuation is the extent to which the existing mistakes interfere in conveying meaning. At B2 level there is a good chance that they will, whereas at C1 they do not, to the extent that they become imperceptible. Similarly, when judging Grammaticality/ Accuracy, for C1 level, where candidates are expected to maintain close control over grammatical accuracy throughout, any mistakes are difficult to discern. In contrast, at B2 level (at least), where less but effective control is expected, such mistakes are few and far between and it is intelligibility that distinguishes B2 from B1.

The point to be made here is that the differences between levels, although distinct, are very small, therefore some degree of disagreement between raters would be understandable. What is more, if intelligibility is the defining point, then the synergy of other language criteria might have affected the outcome. In the next section, where qualitative analysis on Cohesion & Coherence is addressed, it is shown that there are a number of other means to ensure communication of a message.

There is yet another reason why it would be negligent to claim that raters could not differentiate within C level at least on these two criteria. In the respective Methodology section it was noted that the sample is adequate but small, especially for overall language performance where there are only 33 sets of marks, not to mention that a slight kyrtosis could be observed. These suggest that the corpus which probably did not contain any/ enough of fully satisfactory grades to compare, might have affected the statistical results.

Of course, this calls for more investigation as these marks could negatively reflect on the candidates' final score for the writing module in the KPG examination battery. To this end, it would be useful to determine how these three language criteria map on the two different writing activities and to what extent they are integral to the rating outcomes. However, the former requires detailed qualitative script analysis, which falls outside the scope of this dissertation. In the next section, an effort is made to predict the possible extent of the effect of these language criteria. .

4.2.5. RQ2c- Regressions

1. What predictive models can explain most of the variation in each rater's marking profile?

If the results of the t-tests presented in the previous section were of any indication, then it would be expected to find that in this corpus the combination of Spelling and Punctuation, Grammaticality/Accuracy and Coherence and Cohesion would explain most of the variation in each rater's marking profile (see Table 12.)

Table 12. Regression analysis results of the most predictive models for this corpus

Dependent Variables	*Independent Variables*	***r***	***r²***	***B***	***SE***	***β***
R1OLP		.987	.974			
	SPPUNCTR1	.811		.253	.038	.275
	GRAMACCR1	.872		.301	.037	.363
	COHR1	.924		.418	.044	.469
R1OLP		.967	.935			
	VRR1	.929		.462	.064	.567
	APPR1	.906		.461	.080	.450
R2OLP		.981	.963			
	SPPUNCTR2	.621		.225	.025	.340
	GRAMACCR2	.863		.323	.042	.417
	COHR2	.874		.421	.051	.449
R2OLP		.945	.893			
	VRR2	.854		.328	.062	.446
	APPR2	.891		.600	.088	.575

B:unstandardized Beta; *SE*: standard error unstandardized Beta; *β*: standardized Beta

1. *R1OLP= .010+ (.253*SPPUNCTR1)+ (.301*GRAMACCR1)+ (.418*COHR1)*

The results indicated that the model explained 97.4% of the variance and that it was a significant predictor of Rater 1 Overall Language Performance (R1OLP), *F*(3, 29)= 358.09, *p*< .001. All three variables were contributing factors: Spelling and Punctuation (SPPUNCTR1), *b*= .253, *p*< .001; Grammaticality/ Accuracy

(GRAMACCR1), *b*= .301, *p*< .001; Cohesion and Coherence (COHR1), *b*= .418, *p* < .001.

2. _R10LP= .306+ (.462*VRR1)+ (.461*APPR1)_

The results indicated that the model explained 93.5% of the variance and that it was a significant predictor of Rater 1 Overall Language Performance (R1OLP), *F*(2, 30)= 215.42, *p*< .001. Both variables were contributing factors: Vocabulary Range (VRR1), *b*= .462, *p*< .001; Appropriacy (APPR1), *b*= .461, *p*< .001.

3. _R20LP= .215+ (.225*SPPUNCTR2)+ (.323*GRAMACCR2)+ (.421*COHR2)_

The results indicated that the model explained 96.3% of the variance and that it was a significant predictor of Rater 2 Overall Language Performance (R2OLP), *F*(3, 29)= 253.77, *p*< .001. All three variables were contributing factors: Spelling and Punctuation (SPPUNCTR2), *b*= .225, *p*< .001; Grammaticality/ Accuracy (GRAMACCR2), *b*= .323, *p*< .001; Cohesion and Coherence (COHR2), *b*= .421, *p* < .001.

4. _R2OLP= .144+ (.328*VRR2)+ (.600*APPR2)_

The results indicated that the model explained 93.5% of the variance and that it was a significant predictor of Rater 2 Overall Language Performance (R2OLP), *F*(2, 30)= 125.44, *p*< .001. Both variables were contributing factors: Vocabulary Range (VRR2), *b*= .328, *p*< .001; Appropriacy (APPR2), *b*= .600, *p*< .001.

2. Do the Rater 1- and Rater 2-specific designs map on cumulative predictive models?

From section 4.1.5., designs 5 and 6, it was shown that the mean for Activity 1 could be most predicted (apart from the obvious mean of all respective language marks) by Vocabulary Range, when taking one variable into consideration, and Vocabulary Range and Coherence and Cohesion combined, when looking at two. Similarly, for Activity 2 the most predictive language criteria were, more or less equally, double combinations of Grammaticality/ Accuracy, Appropriacy and Cohesion and Coherence (4.1.5., designs 7, 8 and 9).

However, the results from these designs were derived by and can apply to Activity 1 and 2 when both of the raters' marks are considered. Therefore, in order to see how they mapped on each rater individually (with a view to producing generalizable rater-focused models), there were two angles to follow.

Inter-rater comparison

The first one was to calculate the predictive strength of each language criterion for each rater's mean for overall language performance (R1OLP and R2OLP) and then pick the two, or three, variables that were most predictive for each one and run them against each rater (see Table 13.) .

Table 13. Interrater comparison on the predictive strength of individual language criteria

Independent Variables	***Dependent Variables***	
	R1OLP r^2	R2OLP r^2
Spelling & Punctuation	.658*	.386*
Vocabulary Range	.863*	.729*
Grammaticality/ Accuracy	.761*	.746*
Appropriacy	.820*	.794*

Coherence & Cohesion	.854*	.764*

*all models significant at p< .001

The series of linear regressions which were calculated resulted in identifying Vocabulary Range, Cohesion and Coherence and Appropriacy as the three first most predictive variables for Rater 1, and Appropriacy, Cohesion and Coherence and Grammaticality/ Accuracy for Rater 2. The choice to look at the three first was because there was evident overlap between the two raters and because there were no issues with multicollinearity when combining them.

In this light, four more regressions were run to assess the predictive strength of said variables for both raters (see Table 12).

Table 12. Predictive strength of research-derived regression designs applied to both tasks

Independent Variables	*Independent Variables*	r^2	*B*	*SE*	*Β*	*p*
R1OLP		.954				
	GRAMACCR1		.376	.072	.421	.000
	APPR1		.291	.083	.285	.002
	COHR1	.	.376	.072	.421	.000
R1OLP		.954				
	VRR1		.355	.062	.436	.000
	APPR1		.296	.083	.290	.001
	COHR1		.277	.079	.310	.001
R2OLP		.906				
	GRAMACCR2		.204	.081	.263	.017
	APPR2		.405	.110	.389	.001
	COHR2		.356	.088	.380	.000
R2OLP		.913				
	VRR2		.219	.071	.297	.005
	APPR2		.494	.091	.474	.000
	COHR2		.254	.099	.271	.016

*B:*unstandardized Beta*; SE:* standard error unstandardized Beta*; β:* standardized Beta

Their final predictive models and significance are as follows:

R1OLP= .288+ (.291*GRAMACCR1)+ (.291*APPR1)+ (.376*COH1), *F*(3, 29)= 200.74, *p*< .001.

R1OLP= .116+ (.355*VR1)+ (.296*APPR1)+ (.277*COHR1), *F*(3, 29)= 201.83, *p*< .001.

R2OLP= .268+ (.204*GRAMACCR2)+ (.405*APPR2)+ (.356*COH2), *F*(3, 29)= 92.68, *p*< .001.

R2OLP= .051+ (.219*VR2)+ (.494*APPR2)+ (.254*COHR2), *F*(3, 29)= 101.42, *p*< .001.

Inter-activity comparison

The second avenue to explore, regarding the question of whether derived rater-specific designs can map on previously retrieved cumulative designs, was to look at each activity by each rater's marks (A1R1, A2R1, A1R2 and A1R2), find the predictive strength of each variable for all four instances and then divide them per activity into two groups to find the two, or three, with the most predictive strength, irrespective of rater (see Table 15.). The newly produced models (see Table 16.) were then compared to the predictive strength of cumulative language designs for the two activities (as before) (see Table 17.).

Table 15. Inter-activity comparison on the predictive strength of individual language criteria

Independent Variables	*Dependent Variables*		*Dependent Variables*	
	A1R1		***A2RI***	
	r^2	*p*	r^2	*p*
Spelling and Punctuation	.273	.002	.389	.000
Vocabulary Range	.485	.000	.492	.000
Grammaticality/ Accuracy	.320	.001	.604	.000
Appropriacy	.553	.000	.562	.000
Coherence/ Cohesion	.513	.000	.620	.000
	A1R2		***A2R2***	
	r^2		r^2	
Spelling and Punctuation	.029	.340	.092	.086
Vocabulary Range	.679	.000	.197	.010
Grammaticality/ Accuracy	.431	.000	.368	.000
Appropriacy	.437	.000	.379	.000
Coherence/ Cohesion	.584	.000	.234	.004

Table 15. shows that the linear regressions run for Activity 1 resulted in Vocabulary Range, Appropriacy and Cohesion and Coherence being the three most variance predicting variables. As for Activity 2, it was found that Grammaticality/ Accuracy, Appropriacy and Cohesion and Coherence explained most of the variance.

The next step was regressing the set of variables pertinent to each activity to retrieve their potential cumulative predictive strength so as to compare it to the overall language performance models for each activity (Table 16.).

Table 16. Activity-specific overall language performance predictive models

Dependent	*Independent Variables*	r^2	*B*	*SE*	*B*	*p*

Variables						
A1R1		.587				.000
	VRR1		.184	.209	.202	.386
	APPR1		.486	.278	.423	.092
	COHR1		.184	.266	.184	.493
A1R2		.707				.000
	VRR2		.663	.203	.580	.003
	APPR2		.102	.259	.063	.697
	COHR2		.356	.281	.245	.216
A2R1		.708				.000
	GRAMACCR1		.392	.152	.403	.015
	APPR1		.141	.246	.117	.572
	COHR1		.409	.214	.390	.066
A2R2		.415				.001
	GRAMACCR2		.354	.265	.345	.192
	APPR2		.543	.364	.394	.146
	COHR2		-.094	.291	-.076	.748

*B:*unstandardized Beta*; SE:* standard error unstandardized Beta*; β:* standardized Beta

The final step was to calculate the overall language performance designs for each activity and rater, and compare them to the curtailed ones from the previous step (Table 17.).

Table 17. Activity- and rater- specific regression designs in two conditions

Dependent Variables	***5-variable model***	***3-variable model***
	r^2	***r^2***
A1R1	.593	.587
A1R2	.722	.707
A2R1	.719	.708
A2R2	.418	.415

In this corpus, overall language performance criteria predicted 59.3% of the variance observed in Rater 1's marks for Activity 1 (A1R1). The model was significant, *F*(5, 27)= 7.85*, p*< .001. No significant predictors were found.

Rater 2's marks for Activity 1 (A1R2) were explained by overall language performance up to 72.2%. The significant model, *F*(5, 27)= 14.01, *p*< .001, yielded a statistically significant factor, Vocabulary Range (VRR2) (*b*=.654, *p*= .004).

Rater 1's marks for Activity 2 (A2R1) were also significantly predicted by overall language performance, *F*(5, 27)= 13.81, *p*< .001. The model explained 71.9% of variance and brought forward Grammaticality/ Accuracy as a significant contributor, *b*= .431, *p*= .016. However this design may be liable to multicollinearity as the Tolerance level for Cohesion and Coherence was .195, when < .20 is generally considered acceptable, and its VIF was 5.138.

Lastly, the scores for Activity 2 by Rater 2 (A2R2) were explained by overall language performance at 41.8%. The significant model *F*(5, 27)= 3.88, *p*= .009 did not yield any significant contributors.

Comparison between the complete overall language performance model and the 3-variable ones retrieved from inter-rater and intra-activity investigation suggests that the latter models can explain most of the variation as well.

4.2.6. Analysis

In the first half of this section the focus was on quantifying the extent of incongruence and concurrence characterizing the raters' quality judgements of candidate overall language performance. Spelling & Punctuation, Grammaticality/ Accuracy and Cohesion & Coherence were indeed found to explain almost all variance observed, with the combination of Vocabulary Range and Appropriacy scoring a little lower.

Although, as noted before, qualitative research is in order (see section 4.3.), the questions that were addressed thereafter were who (which rater), where (which activity) and what (which of the language criteria).

Inter-rater comparison highlighted the predictive strength of two three-variable combinations, one for Rater 1 (Vocabulary Range, Appropriacy and Cohesion & Coherence) and one for Rater 2 (Grammaticality/ Accuracy, Appropriacy and Cohesion & Coherence). Actually, for Rater 1 both models explained the same percentage of variance, while with Rater 2, Rater 1's model was more predictive. This could suggest that the Vocabulary Range, Appropriacy and Coherence & Cohesion design is more applicable at the C level for expository compositions.

Inter-activity comparison yielded the same results but also connected Vocabulary Range with Activity 1 and Rater 2, and Grammaticality/ Accuracy with Activity 2 and Rater 1. This could prove to be an important piece of information fueling further qualitative research.

Last but not least, even though the rater's judgements differed, within and between them, their marks were complementary to one another. Where Rater 1 was found to predict around 70% of the task completion mean, Rater 2 did the same with the mean for overall language performance. Likewise, where Rater 1 predicted approximately 50% of variance with the overall language performance mean, Rater 2 followed suit with the task completion mean. Although this might suggest a sense of balance, it should be noted that it cannot have been planned as the script bundles after being marked by the first rater they are returned to the script bundle pool and are randomly allocated to a second rater.

Another potentially significant point is that two variables were always in the top three in their quality judgements, Appropriacy and Cohesion & Coherence. If Vocabulary Range and Grammaticality/ Accuracy can be regarded, and thus disregarded, as activity- or rater-specific, then it can be claimed that these two are in no such way restricted, but language-proficiency-level-dependent.

Raters seem to favour/ expect highly cohesive, coherent and appropriate use of language at C level of language proficiency. In particular, according to the marking Grid for C level (see Fig. 1 and Fig. 2), as regards Cohesion & Coherence, the difference between C2 and C1 is that the former consistently (as well as appropriately, accurately and skillfully) use both a wide range of connectives and other cohesive devices, while the latter do so most of the times and tend to overuse some or underuse others, in all probability on account of their (lack of) familiarity with them.

With respect to Appropriacy, again, the difference between levels lies in the observed frequency of the relevant criteria. At C2 level candidates are expected to make appropriate lexicogrammatical choices in terms of communicative purpose, while at C1 candidate performance registers fluctuations, although without impeding on meaning-making.

A final point to be made here is that Appropriacy appears to function as a safety net between task completion and overall language performance. Even if the candidate is highly proficient, unless they succeed in meeting the task requirements, then the language they can possibly provide is a moot point. In this sense, it could be claimed that it overrides the other language criteria, even though it cannot be broken down and analyzed as easily as, for example, Cohesion & Coherence. This is why it is felt that it needs to be awarded a high gravity status in the KPG marking scheme, or every other examination it is likewise employed in.

4.2.7. Summary

To sum up, what interrater comparison has brought forth is that, although there might be some degree of rater bias, inconsistencies are within acceptable standards. Furthermore, the two most predictive models of task completion were: 1. Vocabulary Range-Appropriacy-Cohesion & Coherence, and 2. Grammaticality/Accuracy-Appropriacy- Cohesion & Coherence. Actually, the first one was found to be even more applicable to both activities than the second. The significance of Vocabulary Range was able to be traced to Rater 2 and Activity 1, while Rater 1 significantly influenced the Grammaticality/ Accuracy result for Activity 2. What is more, one might argue that the research-derived intra-rater differences can be explained by the different demands of each activity (see 4.1.4. and 4.1.6.) and that they reflect varied candidate performance. Nevertheless, although acceptable, interrater differences is very important especially in an examination that relies on rater reliability/ internal validity. What is proposed, if not already implemented, is that the training programmes are continuously recalibrated to adjust to the multiple task requirements in all their variations.

4. 3. Qualitative Analysis: Cohesion and Coherence

The quantitative research presented in the previous sections, whether candidate- (section 4.1) or rater-oriented (section 4.2), revealed that in all cases and all types of analyses the criterion of Cohesion & Coherence is strong indicator when assessing writing quality. In order to investigate the extent of its impact within C level (C1 and C2) on candidates' scripts, a sub-corpus of the initial corpus was formed which consisted of: 8 scripts inconsistently rated in terms of Cohesion & Coherence ,with one rater awarding them with '5' and the other with '3', and 8 scripts which were given a mark of '5' by both. Of course, the ensuing results are not to be generalized to larger populations, due to the small size of the sample. However, they do provide some insight into linguistic choices that can potentially differentiate advanced learners within C level.

The research question to be answered is as follows:

Main Research Question 3: Are there any indicators regarding cohesion and coherence which can potentially allow for distinctions between C1 and C2 level candidate performance?

4.3.1. One-sample t-test

To supplement the quantitative analysis thus far performed, qualitative measures using Coh-metrix 3.0 and targeting Cohesion & Coherence were employed. A one-sample *t*-test calculated that, out of the 106 available indices, the sub-corpus of scripts awarded with 5 marks was significantly different to the 3-marks one on 55 of said indices (see Appendices B.2. for the complete 106-indice list).

Subsequent categorization (Table 18.) based on previous research (Crossley & McNamara, 2012; McNamara et al, 2014; Aryadoust & Liu, 2015) sorted these features into three types: Surface Code (25), Textbase (11) and Situation Model (10) measures. The 9 remaining reported on descriptive aspects (5), text easability (3) and readability (1) (Table 19.).

Table 18. Coh-metrix indices initial categorization- mental representations

SURFACE CODE	***TEXTBASE***	***SITUATION MODEL***
Syntactic Complexity	***Referential Cohesion***	***LSA***
68- Number of modifiers per noun phrase, mean	32- Argument overlap, all sentences, binary, mean	39- LSA overlap, adjacent sentences, standard deviation
69- Minimal Edit Distance, part of speech	34- Content word overlap, adjacent sentences, proportional, mean	40- LSA overlap, all sentences in paragraph, mean
70- Minimal Edit Distance, all words	35- Content word overlap, adjacent sentences, proportional, standard deviation	41- LSA overlap, all sentences in paragraph, standard deviation
71- Minimal Edit Distance, lemmas	36- Content word overlap, all sentences, proportional, mean	43- LSA overlap, adjacent paragraphs, standard deviation
72- Sentence syntax similarity, adjacent sentences, mean	***Lexical Diversity***	44- LSA given/new, sentences, mean

73- Sentence syntax similarity, all combinations, across paragraphs, mean	46- Lexical diversity, type-token ratio, content word lemmas	45- LSA given/new, sentences, standard deviation
Syntactic Pattern Density	47- Lexical diversity, type-token ratio, all words	***Situation Model***
74- Noun phrase density, incidence	48- Lexical diversity, MTLD, all words	59- Causal verbs and causal particles incidence
75- Verb phrase density, incidence	49- Lexical diversity, VOCD, all words	61- Intentional verbs incidence
77- Preposition phrase density, incidence	50- All connectives incidence	63- Ratio of intentional particles to intentional verbs
81- Infinitive density, incidence	53- Adversative and contrastive connectives incidence	66- Temporal cohesion, tense and aspect repetition, mean
Word Information	54- Temporal connectives incidence	
82- Noun incidence		

83- Verb incidence	97- Concreteness for content words, mean
84- Adjective incidence	98- Imagability for content words, mean
85- Adverb incidence	99- Meaningfulness, Colorado norms, content words, mean
92- CELEX word frequency for content words, mean	100- Polysemy for content words, mean
93- CELEX Log frequency for all words, mean	101- Hypernymy for nouns, mean

95- Age of acquisition for content words, mean	102- Hypernymy for verbs, mean
96- Familiarity for content words, mean	103- Hypernymy for nouns and verbs, mean

Table 19. Uncategorized Coh-Metrix indices: descriptives, text easability, readability

Descriptives	***Text Easability***	***Readability***
2- Sentence count, number of sentences	13- Text Easability PC Narrativity percentile	106- Coh-Metrix L2 Readability
8- Word length, number of syllables, mean	19. Text Easability PC Referential Cohesion, percentile	
9- Word length, number of syllables, standard deviation	23- Text Easability PC Verb Cohesion, percentile	
10- Word length, number of letters, mean		
11- Word length, number of letters, standard deviation		

The scope and design limitations of this research (see Methodology 3.3) warranted that the list of indices be condensed, allowing for the most statistically significant ones (N=13, p< .005) to come to light. As before, they were sorted according to measure type: Surface Code (10), Textbase (2) and Situation Model (1) (Table 20.). None of the uncategorized indices was found to be highly statistically significant (p< .005)

Table 20. Coh-metrix highly significant indices- p< .005

Indice	*Label*	*5 marks*	*3 marks*	*p*
	Lexical Diversity			
46. LDTTRc	Lexical diversity, type-token ratio, content word lemmas	0.476	0.480	.003
47. LDTTRa	Lexical diversity, type-token ratio, all words	0.263	0.260	.004
	Situation Model			
61. SMINTEp	Intentional verbs incidence	15.391	15.436	.001
	Syntactic Complexity			
69. SYNMEDpos	Minimal Edit Distance, part of speech	0.663	0.662	.000
70. SYNMEDwrd	Minimal Edit Distance, all words	0.886	0.876	.004
71. SYNMEDlem	Minimal Edit Distance, lemmas	0.864	0.854	.004
	Syntactic Pattern Density			
74. DRNP	Noun phrase density, incidence	389.182	390.665	.001
	Word Information			
93.WRDFRQa	CELEX Log frequency for all words, mean	3.218	3.205	.001
95. WRDAOAc	Age of acquisition for content words, mean	362.040	362	.000
96. WRDFAMc	Familiarity for content words, mean	584.449	580.736	.002
97. WRDCNCc	Concreteness for content words, mean	347.116	349.106	.002
98. WRDIMGc	Imagability for content words, mean	394.150	395.146	.001
99. WRDMEAc	Meaningfulness, Colorado norms, content words, mean	432.086	434.311	.002

As regards Surface Code, the derived indices related to syntactic complexity (3), syntactic pattern density (1) and word formation (6). The results indicate that 5-marks judgements involved achieving a higher score in minimal edit distance in parts of speech, all words and lemmas and word frequency for all words, but a lower score in noun density incidence. Furthermore, indices for age of acquisition, concreteness, imagability and meaningfulness of content words were higher than those found in the 3-marks sub-corpus, while the opposite was noted for the familiarity of content words incidence.

The Textbase features related to lexical diversity and, in particular, linked 5-marks judgements with a lower type to token ratio of content word lemmas, but a higher one for all words.

Lastly, with respect to Situation Model measures, a higher intentional verbs incidence was registered with the 3-marks judgements.

4.3.2. Analysis

These results could allow for some distinction between C2 (5-marks) and C1 (3-marks) level of proficiency to be made. First, a juxtaposition of the categories and the way the indices are distributed will be presented. Following, each category will be dealt with separately.

Indices Distribution

The fact that the selected indices had this distribution, Surface Code (10), Textbase (2) and Situation Model (1), is very significant in itself, not only in terms of candidate performance but also in delineating the script raters' quality judgements and comparing them against expected practice. Before continuing with examination of the highly (statistically) significant findings, it should additionally be pointed out that the curtailed distribution is somewhat representative of the findings in total, Surface Code (25), Textbase (11) and Situation Model (10), and as such can apply to the initial corpus (16 scripts) to some extent.

With the Situation Model containing the smallest number, it can be inferred that within C level candidates are able to mentally recreate, and convey to the raters, a mental representation of the situation context with relative ease. According to Zwaan (2001: 14137) situation models stand for "mental representations of events, people, objects, and their relations," which are generated when "we combine the ideas derived from understanding words, clauses and sentences." He continues that they are bound by time and space and they integrate textual information with background knowledge, while they are needed in order to arrive at a coherent translation of a text and explain learning from it. What is more, since discourse "is a coherent description of a sequence of events that are related on several dimensions," the situation model weaves these into a coherent mental representation (ibid.: 14139). In this sense, it would not be amiss to claim that at C level, at least these candidates, they are knowledgeable enough to adequately synthesize the rubric into a coherent product to a great extent.

Not many indices referred to the textbase level either. This suggests that advanced learners (both groups included) are able to successfully convey a "mental representation of the semantic meaning of what was explicitly stated in the text" (ibid.: 14138). However, it should be noted that the highly significant indices related to lexical diversity, which suggests a lexicogrammatically richer text (Yasuda, 2011), for the more proficient learners. The fact, though, that referential cohesion, which could be claimed that it suggests the use of lexical items contextually meaningful, was not highly significant, raises the hope that even less proficient, but still advanced learners, can produce texts with lexical sophistication (ibid., 2011: 125).

Lastly, having more than a third of the surface code indices registering as highly significant is quite interesting. According to the Shallow Structure Hypothesis (SSH) (Clahsen & Felser, 2006a; 2006b; 2017), a model of bilingual sentence processing, L1 and L2 speakers, who are thought to "have available the same processing architecture and mental processing mechanisms" (ibid.: 2017: 1), are "being guided by lexical-semantic and pragmatic information to at least the same extent as adult native speakers" (ibid.: 2006a: 35). This does corroborate the Situation Model and Textbase findings, with the only reservation being that it explicitly refers to adult L2 speakers, when in the KPG C level examination candidates are expected to be, more or less, 16 and above.

The even more interesting point they make is that on-line 'the sentential representations adult L2 learners compute for comprehension contain less syntactic detail than those of native speakers" (ibid.). Could the time-constricted assessment conditions be bringing about a similar effect in written production? The findings indicate that it might be so, since the majority of differences within C level were syntactic complexity, syntactic pattern density and word formation. On the one hand, this could mean that syntactic errors are to be expected even with more proficient L2 learners when under the pressure of on-line communication. On the other hand, it can also be claimed that the more proficient L2 writers seem to be experienced and/ or knowledgeable enough to avoid surface code mistakes.

In sum, the fact that these advanced L2 writers did not differ so much in Situation Model or Textbase measures as they did in the "mental representation of the actual wording of the text" (Zwaan, 2001: 14138), supports the genre-based approach adopted by the KPG, where the communicative purpose and situation context override grammatical mistakes that don't interfere with intelligibility (KPG Script Rater Guide, 2017: 6). Of course, this leaves the surface level responsible for distinguishing within C level (at least). Xie (2015: 24), who argues that "surface features may be considered as observable and reliable indicators of the deeper and latent (invisible) competence, which is the target construct to be measured", agrees with relevant research that surface textual features can be conducive for writing assessment in an L2 context where "language mechanics and writing conventions are legitimate aspects of the constructs of L2 essay tests" (ibid.) In this sense, both the validity of the language examination (or any other for that matter in the same context) and the reliability of the raters are safeguarded.

Surface Code

Writers judged as more language proficient favour complex syntactic structures and it could be claimed that they seem to be able to make mistakes that interfere less in comprehension. The indices for minimal edit distance (Indices 69, 70, 71) are higher with the 5-marks group, suggesting that they need fewer operations (insertion, deletion and substitution), to transform one string of words into another (https://en.wikipedia.org/wiki/Edit_distance). As Aryadoust and Liu (2015: 39) explain, higher scores in these indices indicate high sentential dissimilarity. They also enrich previous research (McCarthy et al., 2009: 685, as cited in Aryadoust and Liu, 2015) by reporting that these indices are related to both writing quality and the other two levels of mental representation, textbase and situation model (Aryadoust & Liu, 2015: 51).

On a phrasal level, there is a tendency to refrain from heavily modifying their nouns succeeding in avoiding making their messages informationally dense (Index 74). Biber and Gray (2010) note that the extent to which a sentence structure is compressed, suggesting density of information, largely depends on the number of modifiers per noun phrase (as cited in Guo et al., 2013: 226). On the other hand, in this corpus less proficient writers tend to overuse the informational power a noun phrase can afford. Could it be for informational economy purposes as Biber, Grieve and Ibberi- Shea (2009) propose? Heavy noun modification puts a strain on the reader, affecting the texts readability and, potentially, even constraining instead of explicating meaning. Actually, Parkinson and Musgrave (2014: 49) report on studies on the development of L1 and L2 writers which highlight the increased level of phrasal and not clausal complexity acquired with the passing of time. By also quoting Halliday (1993, as cited ibid.) it is understood that complex embeddedness of nominal groups is not only developmental but also a feature of written speech. This is in accordance with previous research supporting heavy noun formations to reflect advanced level of language proficiency (e.g., Biber et al., 2011; McNamara, Scott &McCarthy, 2010, as cited in Parkinson and Musgrave, 2014) in academic contexts. It could also, in fact, explain the less advanced learners' choices in this corpus, who being at least senior high schoolers, meaning 16 and above (Dendrinos & Karavas, 2013), have been extensively trained in academic writing. It is not that surprising that they would transfer this knowledge to yet another formal examination context, forsaking actual task requirements (for what they might have believed to be the most 'appropriate' way to answer.

Furthermore, it is observed that texts with better quality assessment in this corpus use more frequent words (Index 93).This is opposed to previous research findings (Crossley, Salsbury, McNamara, & Jarvis, 2011), where writers rated as proficient tended to use more complex and thus less frequent words. However, in those cases the writing tasks under scrutiny were formal and impersonal answers to argumentative tasks. In this research the scripts involve an expository informal personal blog entry and a narrative and expository objective formal encyclopedic entry, which are given an individual mark for task completion and a cumulative mark for overall language, in accordance with the functional theoretical framework underlying the KPG battery, where language use is defined by socially purposeful contexts (Dendrinos & Karavas, 2013: 16). With genre and situational context in mind, the fact that a greater number of sophisticated word choices were not preferred and a more reader-oriented approach was adopted can explain this unexpected finding.

Despite the fact that the 5-marks groups used more frequent words overall, the higher age of acquisition indice for content words (Index 95) indicates that they did so while also using content words expected to be acquired at later stages by children (http://cohmetrix.com/). This proposition is also supported by the higher familiarity indice (Index 96), which suggests that the abstract ideas they were negotiating were at the same time highly recognizable by an adult (ibid).

The above mentioned findings might seem contradictory, but it is also plausible that simpler/frequent function words were intentionally employed to allow them to work their way around notions that required more sophisticated use of vocabulary. This is corroborated by the fact that they scored lower for concreteness (Index 97), imagability (Index 98) and meaningfulness (Index 99). Based on descriptions of the functions of these indices on the Coh-metrix website (ibid) it could be claimed that they indicate abstractness of ideas, which cannot easily be evoked as a mental image, although they are not characterized by semantic ambiguity.

Textbase

With respect to textbase features, more proficient-perceived writers produce texts with higher lexical diversity, avoiding repetition at text level (Index 47). As Guo et al. (2013: 226) explain, a higher lexical score indicates use of a wider range of words. Nevertheless, they appear to rely less on content word lemmas, unlike the 3-marks group (Index 46). This could probably be because more advanced learners are able to employ a variety of cohesive devices, including lexical ones, so as to render it coherent, in contrast to less knowledgeable ones who might overcompensate with content lemmas in order to signpost their ideas.

Situation Model

Finally, the fact that the (judged as) more competent writers used less intentional verbs can be explained either by examining this index (Index 61) taking into consideration the category it belongs to, the situation model, or by combining word count and overall task requirements information to propose a possible explanation. However, it is felt that the analysis required in order to validly identify all the intentional verbs (Tonelli et al., 2012: 45) would be worth another MA dissertation in its entirety and, thus, falls outside the scope of the present one.

Instead, an outside-the-box approach was adopted. This index provides the 'incidence' of intentional verbs, meaning "the number of classified units per 1000 words" (http://cohmetrix.com/). On account of the 5-marks sub-corpus being 2.257 words long (Activity 1: 1.126, Activity 2: 1.131) and the 3-marks one coming up to 2.710 words (Activity 1: 1.443, Activity 2: 1.267), it can be asserted that the latter affords more opportunities for such verbs to be employed. This is especially true for Activity 1, where at least the topic and communicative purpose of the expository read-to-write task encourage intentional verb use (see Materials section for more information). Of course, this does not exclude the possibility that the variance is (also) due to genre-related inappropriate overuse of such verbs in Activity 2. As mentioned in Methodology (3.2.1.) this interligual mediation task involves compiling a narrative biographical encyclopedic entry, followed by an expository task question. In this sense, overuse of intentional verbs could create a mental representation of D. Nanopoulos being a ‘voluntary’ agent of his accomplishments with the view to getting him where he is today. In other words, the synergy of intentional verb use, which is characteristic of narrative texts (Crossley & McNamara, 2012), with a chronological account of events expected in an encyclopedic entry would raise issues of causality and imply careful implementation of a pre-existing plan (ibid).

All in all, one of the many interpretations that can potentially be made of this finding is that being a competent writer at C level involves adhering to task requirements, word count included, which in this case could mean that the 5-marks group had a better control of the mental representations their texts produced.

Chapter 5

Discussion and Conclusion

This dissertation was a micro-scale investigation of both textual characteristics in rated scripts, and rater quality. The quantitative analysis attempted to provide insight into certain textual characteristics that pertain to two genres sharing a common expository dimension, albeit to a different extent. It proposed to identify inconsistencies in writing quality judgements between the two raters. Subsequent qualitative research into the Cohesion and Coherence language criterion, tried to provide an explanation for its prevalence at C level of language proficiency.

In this chapter, the main research questions are discussed separately and they are followed by overall comments, the limitations of the study and further research suggestions.

Main Research Question 1

How does task completion interact with the overall language performance criteria (Spelling & Punctuation, Vocabulary Range, Grammaticality/ Accuracy, Appropriacy and Cohesion & Coherence)?

As far as textual features are concerned, distinction within C level of language proficiency (C1-C2) registered with two of the five criteria for overall language performance in the KPG marking grid for this level: Vocabulary Range and Coherence & Cohesion. This is in agreement with findings in literature in which vocabulary sophistication and cohesion and coherence have been found indicative of advanced level of language proficiency (e.g., Crossley & McNamara, 2012).

What is interesting to note is that when looking at how the language criteria map on the two activities, it is Vacabulary Range that corresponds to the intralingual mediation task (Activity 1) and Grammaticality/ Accuracy to the interlingual mediation task (Activity 2). The first finding corroborates related research by Jeong (2016) in which advanced learners achieved high marks in expository essays and scored the highest in vocabulary, in contrast to grammar. These suggest that there is a genre effect, and in the case of the KPG, a text function effect ('explaining' in Activity 1), on language realization .

Regarding Grammaticality/ Accuracy, it can be argued that it was the task type (integrated interlingual mediation), that can explain this finding. In line with Yoon (2017), who found that that lexical and morphological complexity varies across levels, but was unable to draw clear distinctions when looking at adjacent proficiency levels, this criterion did not shed any light in this respect. However, it does support that, as far as construct coverage is concerned, the two activities might impose different cognitive strains on the candidate. In Activity 2, the candidates are explicitly asked to use the Greek source text in order to do the task. As a result, the source text could be functioning as a prime for L1, impeding L2 activation, and probing candidates to resort to L1 unacceptable formations, thus scoring low for Grammaticality/ Accuracy.

From another perspective, the Grammaticality/ Accuracy finding for Activity 2, which is a fully integrated task requiring summarization and synthesis of information, falls in line with Guo et al.'s (2013) research where it was postulated that raters may assess integrated tasks more severely in this respect, as they have more material readily available. It could be that this did not apply to Activity 1 because of the difference in task requirements.

Main Research Question 2

To what extent do rater judgements overlap/vary and where can this be attributed to?

The two script raters in this corpus were found to be in agreement when marking for task completion, but revealed some degree of inconsistency when marking specific language criteria although their marks positively correlated with one another. In particular, Rater 1 was found to value Grammaticality/ Accuracy more in Activity 2, whereas Rater 2 paid close attention to Vocabulary Range for Activity 1.

These findings echo those of Wind it al. (2017) who argued that rater judgements can be influenced by textual characteristics. However, with some degree of variance in rater agreement being acceptable in many examinations (Hyland, 2003 :217), holistically, these differences can be regarded as negligible. Goodwin (2016: 8) also found negligible discrimination at a holistic level, which was not though the case when checking internal validity. In this line, the findings of this study pose as a valid argument for continuous rater-training to reduce rater severity (Weigle, 1998; Shohamy et al., 1992, as cited in Gebril and Plakans, 2014) and further research into intra-rater reliability.

Main Research Question 3

Are there any indicators regarding cohesion and coherence which can potentially allow for distinctions between C1 and C2 level candidate performance?

The conducted quantitative analysis in this study has underscored the strength of the Cohesion & Coherence language criterion especially at this level. Regression analyses showed that it is a significant predictor not only on its own but in combination with other criteria as well, which suggests that it captures aspects of candidate performance that cannot be taken for granted if certain quality is displayed in another criterion.

Adding to these findings, the third Main Research Question, yielded some interesting results despite the small size of the sub-corpus that was qualitatively analyzed. The results indicate that cohesion and coherence indices are indeed an important text complexity feature which can differentiate between C1 and C2 level performance. This is partly in agreement with research in integrated tasks at advanced level performed by Plakans and Gebril (2016), who found that the bigger the overall score, the better the quality of the texts with respect to cohesion and coherence and organization was. However, as regards distinctions within C level, in this study, the results highlighted surface textual features mostly, with markedly very few exceptions at textbase and situation model level. This has also been derived in previous KPG studies, Blani's (2018) in particular, who only found surface features to afford such distinctions.

In a sense, this can be regarded as fortuitous for rater validity assessment. The fact that raters, who judge writing quality and not text complexity, are said to mostly attend to surface level features has been presented as quite concerning. However, if anything, what this study has shown is that it is this particular aspect of writing performance that can derive the desired outcome in at least this language examination, meaning writing quality distinctions within C level. It appears that, C level candidates are competent enough to effectively interact with task instructions and produce texts that are somewhat equally proficient in terms of communicative purpose and text organization. Problems arise in the final formatting of their work, which is not entirely unexpected, as in time-constrained language examinations there is very little time that can be spared for fine-tuning their text.

Overall Comments

An advantage of this dissertation was that it was based on rated test-takers' scripts on two tasks sharing features on two levels. The first relates to their format and constitutes both read-to-write integrated tasks, since in both instances candidates are provided with a source text to make use of in producing another. The second level has to do with the role the candidates are asked to play, which is that of a mediator. In Activity 1 the source text is supposed to be used as a point of departure for the production of a new text in a different genre, thus intralingual mediation. Likewise, in Activity 2 the target text is expected to relay information retrieved from a source text in the L2, also in a different genre, thus interlingual mediation. These similarities attest to the functional framework underlying this examination suite and its strong connection to social practice.

Guo et al. (2013: 234) conclude that the combination of an independent and an integrated task in the TOEFL examination is complementary as they "share construct coverage and, at the same time, tap into different elements of writing." Similarly, this dissertation argues for the synergy of the two integrated mediation tasks in the KPG examination. While the one fosters the communication of personal ideas, but expects their well-rounded development, the other constricts content, to allow for the extra cognitive resources to tackle the interlingual aspect of the task. This dual process corresponds both to the everyday and academic demands a plurilingual KPG candidate might need to meet. Personal ideas interact with those of others and lead to co-construction of meaning in all situational contexts. In academia, in particular, students are required to tackle a number of source texts simultaneously in order to filter out their own position towards them. Therefore, it would not be amiss to claim, as others have before about the value of integrated tasks (see Anmarkrud,, Braten & Stromso (2014) for both everyday life and academics and Gebril & Plakans (2014) for further references regarding academic contexts), that the KPG C level writing module prepares candidates for real-life writing production and, thus, caters for their needs as active members of a multilingual society.

Limitations

One of the limitations of this study has been the small size of the corpus. It complicated the statistical analysis and limited the scope of the results of the qualitative analysis. Another issue was the slight kyrtosis of the initial corpus which might otherwise have allowed for the other language features, namely Spelling & Punctuation, Grammaticality/ Accuracy and Appropriacy, to provide distinctions within C level. Finally, although the predictive strength of the regression models holds strong for this corpus, it might not possible to be generalized to other larger populations because the final score for its candidate was not available. Depending on the gravity status of the individual criteria other combinations might be more appropriate for this level.

Future Research

It is felt that a logical next step for further research would be to qualitatively address the two most prevalent language criteria, Cohesion & Coherence and Vocabulary Range, using Coh-Metrix and TAALES respectively. If also juxtaposed to rater judgements with think-aloud protocols and other intra- and inter-rater research methods, the results could inform rater training programmes and provide valuable input for the continuous evaluation and calibration of the marking critiria.

References

Anmarkrud, O., Braten, I., & Stromso, H. I. (2014). "Multiple-documents literacy: Strategic processing, source awareness, and argumentation when reading multiple conflicting documents". *Learning and Individual Differences* 30: 64–76.

Aryadoust, V., & Liu, S. (2015). "Predicting EFL writing ability from levels of mental representation measured by Coh-Metrix: A structural equation modeling study". *Assessing Writing* 24: 35–58.

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

Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Banerjee, J., Yan, X., Chapman, M. & Elliott, H. (2015). "Keeping up with the times: Revising and refreshing a rating scale". *Assessing Writing*, http://dx.doi.org/10.1016/j.asw.2015.07.001.

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). "Functional constraints on sentence processing: A cross-linguistic study". *Cognition* 11: 245-299.

Biber, D., & Gray, B. (2010). "Challenging stereotypes about academic writing: Complexity, elaboration, explicitness". *Journal of English for Academic Purposes* 9: 2–20.

Biber, D., Gray, B., & Poonpon, K. (2011). "Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?". *TESOL Quarterly* 45(1): 5–35.

Biber, D., Grieve, J., & Iberri-Shea, G. (2009). "Noun phrase modification". *One Language, Two Grammars?,* 182–193.

Blani, V. (2018). Text Grammar in KPG candidates' scripts: A corpus-based approach. *Unpublished doctoral dissertation*. Athens: National and Kapodistrian University of Athens.

Casal, J. E. & Lee, J. J. (2019). "Syntactic complexity and writing quality in assessed first-year L2 writing". *Journal of Second Language Writing* 44: 51-62.

Cheong, C. M., Zhu, X., Li, G. Y. & Wen, H. (2019). "Effects of intertextual processing on L2 integrated writing". *Journal of Second Language Writing* 44: 63-75.

Cho, Y., Rijimen, F. & Novak, J. (2013). "Investigating the effects of prompt characteristics on the comparability of TOEFL-iBT™ integrated writing tasks". *Language Testing* 30(4): 513-534.

Clahsen, H., & Felser, C. (2006a). "Grammatical processing in language learners". *Applied Psycholinguistics* 27: 3–42.

Clahsen, H., & Felser, C. (2006b). "Continuity and shallow structures in language processing". *Applied Psycholinguistics* 27: 107–126.

Clahsen, H., & Felser, C. (2017). "Critical Commentary: some notes on the Shallow Structure Hypothesis". In *Studies in Second Language Acquisition*. Cambridge University Press,1-17.

Coste, D. & M. Cavalli (2015). *Education, mobility, otherness: The mediation functions of schools*. Strasbourg: Council of Europe DGII – Directorate General of Democracy, Language Policy Unit.

Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment - Companion volume with new descriptors.* Strasbourg: Council of Europe. Available: https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/168074a4e2

Crossley, S. A. & McNamara, D. S. (2014). "Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners". *Journal of Second Language Writing* 26: 66-79.

Crossley, S. A. & McNamara, D. S. (2012). "Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication". *Journal of Research in Reading* 35(2): 115–135.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). "The development and use of cohesive devices in L2 writing and their relations to judgments of essay quality". *Journal of Second Language Writing* 32: 1–16.

Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). "What is lexical proficiency? Some answers from computational models of speech data". *TESOL Quarterly* 45: 182–193.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). “Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL”. *Assessing Writing* 10(1): 5–43.

De Groot, A. M. B. (2011). *Language and cognition in bilinguals and multilinguals: An introduction.* New York: Psychology Press.

Dendrinos, B. & Karavas, K. (Eds.) (2013). *The KPG handbook: performance descriptors and specifications.* Athens: © RCeL, National and Kapodistrian University of Athens.

Deremer, M. L. (1998). “Writing Assessment: Raters’ Elaboration of the Rating Task”. *Assessing Writing* 5(1): 7-29.

Dunsmuir, S., Kyriacou, M., Batuwitage, S., Hinson, E., Ingram, V., & O’ Sullivan, S. (2015). “An evaluation of the Writing Assessment Measure (WAM) for children’s narrative writing”. Assessing Writing 23: 1–18.

Field, A. (2009). *Discovering Statistics Using SPSS, Third Edition*. Los Angeles, London, New Delhi, Singapore, Washington DC: SAGE.

Gebril, A. & Plakans, L. (2014). “Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks”. *Assessing Writing* 21: 56-73.

Gebril, A. & Plakans, L. (2016). “Source-based tasks in academic writing assessment: Lexical diversity, textual borrowing and proficiency”. *Journal of English for Academic Purposes* 24: 78-88.

Goodwin, S. (2016). “A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes”. *Assessing Writing*, http://dx.doi.org/10.1016/j.asw.2016.07.004.

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

Halliday, M. A. K. (1993). “Some grammatical problems in scientific English”. In M. A. K. Halliday, & J. R. Martin (eds.), *Writing science*. London: The Falmer Press, 69–85.

Humphry, S. & Heldsinger, S. (2019). “Raters’ perceptions of assessment criteria relevance”. *Assessing Writing* 41: 1-13.

Hyland, K. (2004). *Genre and second language writing*. Ann Arbor, MI: University of Michigan Press.

Hyland, K. (2003). *Second Language Writing*. Cambridge: Cambridge University Press.

Jeong, H. (2016). “Narrative and expository genre effects on students, raters, and performance criteria”. *Assessing Writing* 31: 113–125.

Johnson, D. & VanBrackle, L. (2012). “Linguistic discrimination in writing assessment: How raters react to African American “errors,” ESL errors, and standard English errors on a state-mandated writing exam”. *Assessing Writing* 17: 35-54.

Karavas, E. & Mitsikopoulou, B. (2018). “Introduction: Issues and challenges in glocal language testing”. In E. Karavas & B. Mitsikopoulou (eds) (2018) Developments in Glocal Language Testing: The Case of The Greek National Foreign Language Proficiency Exam (p. 1-21). Oxford: Peter Lang.

Khalil, A. (1989). "A study of cohesion and coherence in Arab EFL college students' writing”, *System* 17(3), 359-371.

Kim, M. & Crossley, S. A. (2018). “Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing”. *Assessing Writing* 37: 39-56.

Knoch, U. & Sitajalabhorn, W. (2013). “A closer look at integrated writing tasks: Towards a more focused definition for assessment purposes”. *Assessing Writing* 18: 300-308.

Kormos, J. (2011). “Task complexity and linguistic and discourse features of narrative writing performance”. *Journal of Second Language Writing* 20(2): 148-161.

KPG Script Rater Guide, 14, 2017B (2017). Athens: © *RCeL National and Kapodistrian University of Athens.*

Kyle, K., & Crossley, S. (2016). “The relationship between lexical sophistication and independent and source-based writing”. *Journal of Second Language Writing* 34: 12–24.

Lu, X. (2010). “Automatic analysis of syntactic complexity in second language writing”. *International Journal of Corpus Linguistics* 15(4): 474–496.

Lumley, T. (2002). “Assessment criteria in a large-scale writing test: What do they really mean to the raters?”. *Language Testing* 19(3): 246–276.

MacWhinney, B., (2002). Extending the Competition Model. In *Bilingual Sentence Processing*, 1st Edition, Elsevier

Mazgutova, D., & Kormos, J. (2015). “Syntactic and lexical development in an intensive English for Academic Purposes programme”. *Journal of Second Language Writing* 29: 3–15.

McCarthy, P. M., Guess, R. H., & McNamara, D. S. (2009). “The components of paraphrase evaluations”. *Behavioral Research Methods* 41: 682–690.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). Automated evaluation of text and discourse with Coh-Metrix. Cambridge: *Cambridge University Press.*

McNamara, D. S. C., Scott, A., & McCarthy, P. M. (2010). “Linguistic features of writing quality”. *Written Communication* 27: 57–87.

Oikonomidou, V., (2016). Factors affecting read-to-write task difficulty. *Unpublished doctoral dissertation,* National and Kapodistrian University of Athens.

Parkinson, J., & Musgrave, J. (2014). “Development of noun phrase complexity in the writing of English for Academic Purposes students”. *Journal of English for Academic Purposes* 14: 48–59.

Plakans, L. (2009). “The role of reading strategies in integrated L2 writing tasks”. *Journal of English for Academic Purposes* 8: 252-266.

Plakans, L. & Gebril, A. (2012). “A close investigation into source use in integrated second language writing tasks”. *Assessing Writing* 17: 18-34.

Plakans, L. & Gebril, A. (2017). “Exploring the relationship of organization and connection with scores in integrated writing assessment”. *Assessing Writing* 31: 98-112.

Plakans, L. & Gebril, A. (2013). “Using multiple texts in an integrated writing assessment: Source text use as a predictor of score”. *Journal of Second Language Writing* 22: 217-230.

Plakans, L., Liao, J. T. & Wang, F. (2019). ““I should summarize this whole paragraph”: Shared processes of reading and writing in iterative integrated assessment tasks”. *Assessing Writing* 40: 14-26.

Rashid, S. M. D. & Rafik-Galea, S. (2007). “ESL writing variability: writing tasks, gender and proficiency level”. *Indonesian Journal of English Language Teaching* 3(2): 227-243.

Razali, N. M. & Wah, Y. B. (2011). “Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests”. *Journal of Statistical Modeling and Analytics* 2 (1): 21-33.

Rezazadeh, M., Tavakoli, M. & Eslami-Rakesh, A. (2011). "The role of task type in foreign language written production: focusing on fluency, complexity and accuracy". *International Education Studies* 4(2): 169-176.

Shohamy, E., Gordon, C. M., & Kraemer, R. (1992). "The effect of raters' background and training on the reliability of direct writing tests". *Modern Language Journal* 76(1): 27–33.

Stathopoulou, M. (2015). *Cross-language mediation in foreign language teaching and testing.* Multilingual Matters.

Stathopoulou, M. (2014). The linguistic characteristics of KPG written mediation tasks across proficiency levels, in *Major Trends in Theoretical and Applied Linguistics 3.* Sciendo Migration, 349-366.

Stathopoulou, M. (2019). "The reading-to-write construct across languages: Analysing written mediation tasks and performance". *Selected papers on theoretical and applied linguistics* 23: 414-428.

Stathopoulou, M. (2009). Written mediation in the KPG exams. Source text regulation resulting in hybrid formations. *Unpublished MA Thesis*, National and Kapodistrian University of Athens.

Tonelli, S., Manh, K. T., & Pianta, E. (2012). "Making readability indices readable. *NAACL-HLT 2012 Workshop on Predicting and Improving Text Readability for target reader populations"* (PITR 2012): 40-48.

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

Weigle, S. C. (2013). "English language learners and automated scoring of essays: Critical considerations". Assessing Writing 18(1): 85–99.

Weigle, S. C. (1998). "Using FACETS to model rater training effects". *Language Testing* 15(2): 263–287.

Wind, S. A. (2019). "Do raters use rating scale categories consistently across analytic rubric domains in writing assessment?". *Assessing Writing*, https://doi.org/10.1016/j.asw.2019.100416.

Wind, S. A., Stager, C. & Patil, Y. J. (2017). "Exploring the relationship between textual characteristics and rating quality in rater-mediated writing assessments: An illustration with L1 and L2 writing assessments". *Assessing Writing* 34: 1-15.

Xie, Q. (2015). ""I must impress the raters!" An investigation of Chinese test-takers' strategies to manage rater impressions". *Assessing Writing* 25: 22–37.

Yasuda, S. (2011). “Genre-based tasks in foreign language writing. Developing writers’ genre awareness, linguistic knowledge, and writing competence”. *Journal of Second Language Writing* 20: 111-133.

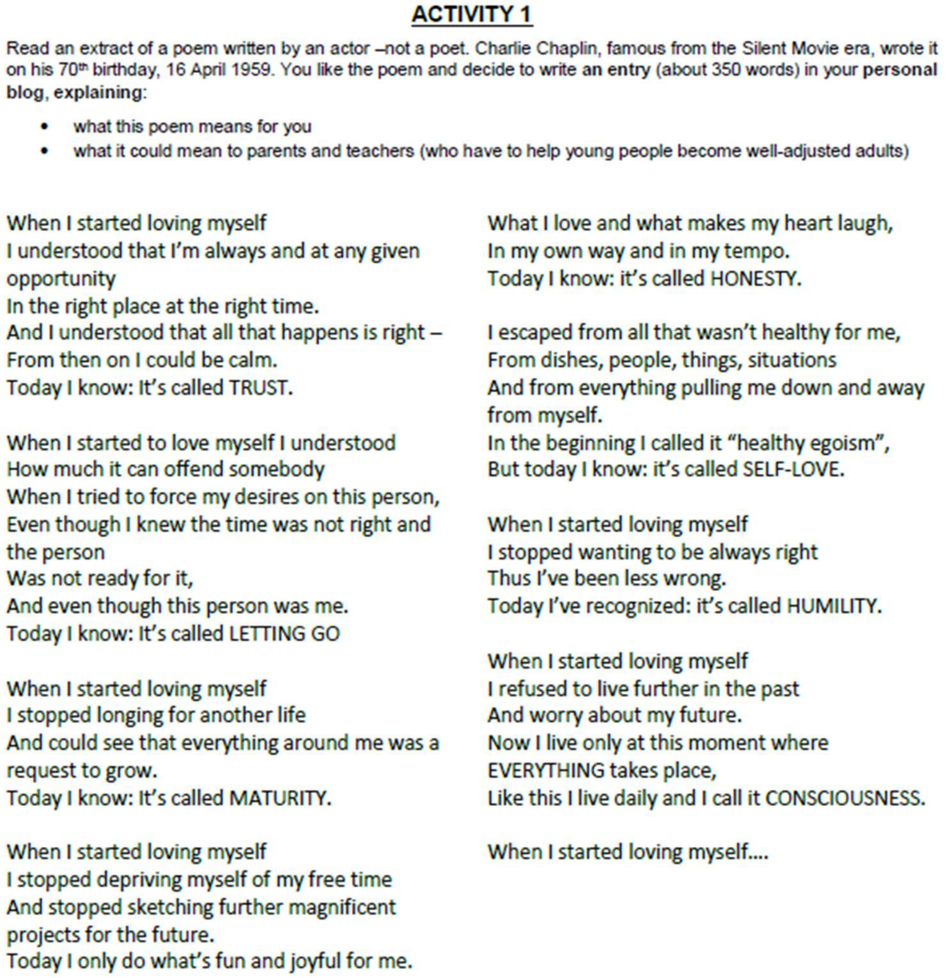
Yoon, H.-J., & Polio, C. (2016). “The Linguistic Development of Students of English as a Second Language in Two Written Genres”. *TESOL Quarterly* 51(2): 275–301.

Zwaan, R. A. (2001). “Situation Model: Psychological”. *International Encyclopedia of the Social & Behavioral Sciences*, 14137–14141.

Appendices

Appendix A

1. **Activity 1 source text**



ACTIVITY 1

Read an extract of a poem written by an actor –not a poet. Charlie Chaplin, famous from the Silent Movie era, wrote it on his 70th birthday, 16 April 1959. You like the poem and decide to write **an entry** (about 350 words) in your **personal blog, explaining**:

- what this poem means for you
- what it could mean to parents and teachers (who have to help young people become well-adjusted adults)

When I started loving myself
I understood that I'm always and at any given opportunity
In the right place at the right time.
And I understood that all that happens is right –
From then on I could be calm.
Today I know: It's called TRUST.

When I started to love myself I understood
How much it can offend somebody
When I tried to force my desires on this person,
Even though I knew the time was not right and the person
Was not ready for it,
And even though this person was me.
Today I know: It's called LETTING GO

When I started loving myself
I stopped longing for another life
And could see that everything around me was a request to grow.
Today I know: It's called MATURITY.

When I started loving myself
I stopped depriving myself of my free time
And stopped sketching further magnificent projects for the future.
Today I only do what's fun and joyful for me.
What I love and what makes my heart laugh,
In my own way and in my tempo.
Today I know: it's called HONESTY.

I escaped from all that wasn't healthy for me,
From dishes, people, things, situations
And from everything pulling me down and away from myself.
In the beginning I called it "healthy egoism",
But today I know: it's called SELF-LOVE.

When I started loving myself
I stopped wanting to be always right
Thus I've been less wrong.
Today I've recognized: it's called HUMILITY.

When I started loving myself
I refused to live further in the past
And worry about my future.
Now I live only at this moment where
EVERYTHING takes place,
Like this I live daily and I call it CONSCIOUSNESS.

When I started loving myself....

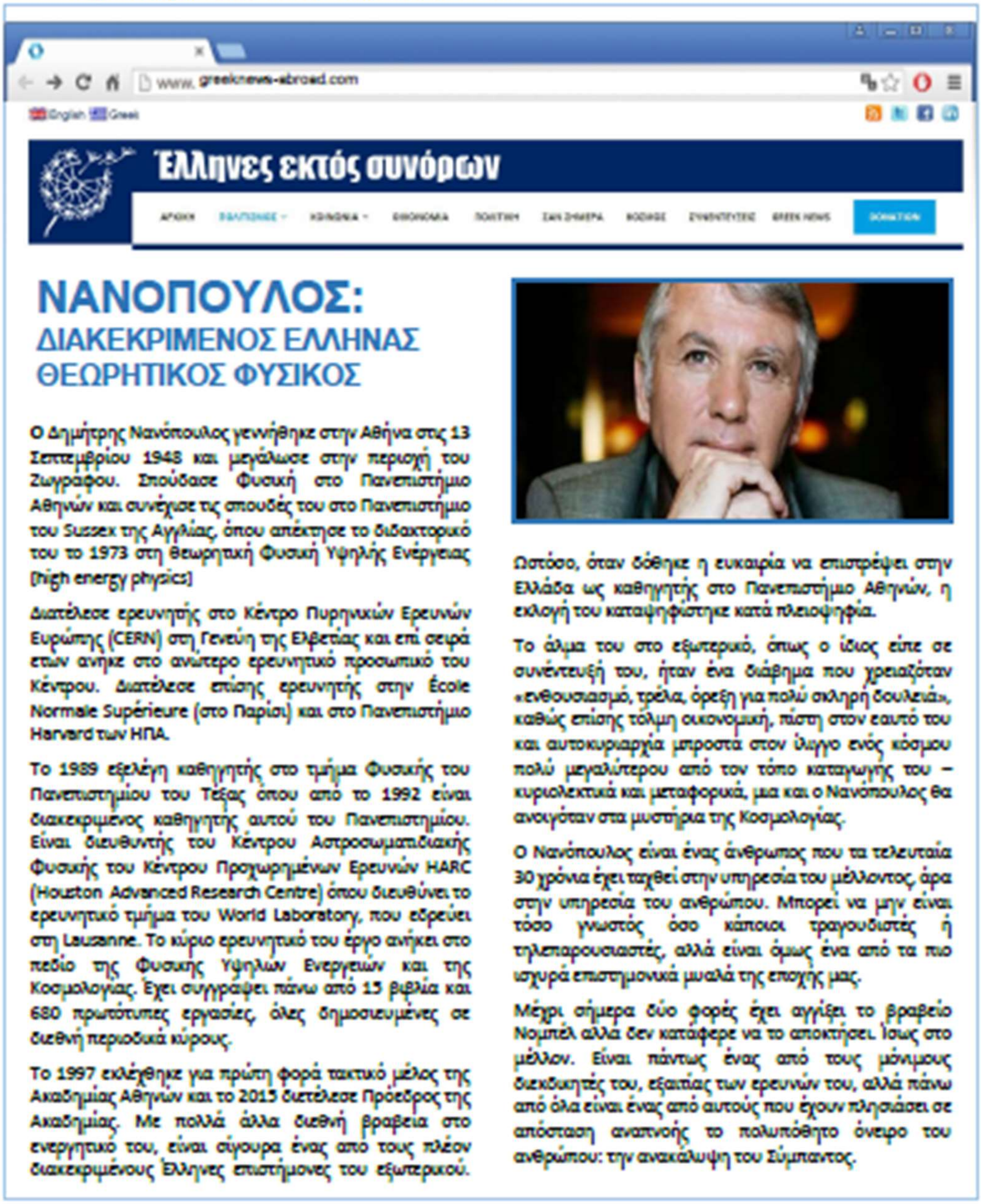
(KPG Script Rater Guide, 2017: 13)

2. Activity 2 source text

ACTIVITY 2

Using information from the text below, write an entry (about 300 words) for an electronic encyclopaedia:

- Provide factual information about Dimitris Nanopoulos
- Explain why he counts as one of the "famous Greeks"



ΝΑΝΟΠΟΥΛΟΣ:
ΔΙΑΚΕΚΡΙΜΕΝΟΣ ΕΛΛΗΝΑΣ ΘΕΩΡΗΤΙΚΟΣ ΦΥΣΙΚΟΣ

Ο Δημήτρης Νανόπουλος γεννήθηκε στην Αθήνα στις 13 Σεπτεμβρίου 1948 και μεγάλωσε στην περιοχή του Ζωγράφου. Σπούδασε Φυσική στο Πανεπιστήμιο Αθηνών και συνέχισε τις σπουδές του στο Πανεπιστήμιο του Sussex της Αγγλίας, όπου απέκτησε το διδακτορικό του το 1973 στη Θεωρητική Φυσική Υψηλής Ενέργειας (high energy physics)

Διατέλεσε ερευνητής στο Κέντρο Πυρηνικών Ερευνών Ευρώπης (CERN) στη Γενεύη της Ελβετίας και επί σειρά ετών ανήκε στο ανώτερο ερευνητικό προσωπικό του Κέντρου. Διατέλεσε επίσης ερευνητής στην École Normale Supérieure (στο Παρίσι) και στο Πανεπιστήμιο Harvard των ΗΠΑ.

Το 1989 εξελέγη καθηγητής στο τμήμα Φυσικής του Πανεπιστημίου του Τέξας όπου από το 1992 είναι διακεκριμένος καθηγητής αυτού του Πανεπιστημίου. Είναι διευθυντής του Κέντρου Αστροσωματιδιακής Φυσικής του Κέντρου Προχωρημένων Ερευνών HARC (Houston Advanced Research Centre) όπου διευθύνει το ερευνητικό τμήμα του World Laboratory, που εδρεύει στη Lausanne. Το κύριο ερευνητικό του έργο ανήκει στο πεδίο της Φυσικής Υψηλών Ενεργειών και της Κοσμολογίας. Έχει συγγράψει πάνω από 15 βιβλία και 680 πρωτότυπες εργασίες, όλες δημοσιευμένες σε διεθνή περιοδικά κύρους.

Το 1997 εκλέχθηκε για πρώτη φορά τακτικό μέλος της Ακαδημίας Αθηνών και το 2015 διετέλεσε Πρόεδρος της Ακαδημίας. Με πολλά άλλα διεθνή βραβεία στο ενεργητικό του, είναι σίγουρα ένας από τους πλέον διακεκριμένους Έλληνες επιστήμονες του εξωτερικού.

Ωστόσο, όταν δόθηκε η ευκαιρία να επιστρέψει στην Ελλάδα ως καθηγητής στο Πανεπιστήμιο Αθηνών, η εκλογή του καταψηφίστηκε κατά πλειοψηφία.

Το άλμα του στο εξωτερικό, όπως ο ίδιος είπε σε συνέντευξή του, ήταν ένα διάβημα που χρειαζόταν «ενθουσιασμό, τρέλα, όρεξη για πολύ σκληρή δουλειά», καθώς επίσης τόλμη οικονομική, πίστη στον εαυτό του και αυτοκυριαρχία μπροστά στον ίλιγγο ενός κόσμου πολύ μεγαλύτερου από τον τόπο καταγωγής του – κυριολεκτικά και μεταφορικά, μια και ο Νανόπουλος θα ανοιγόταν στα μυστήρια της Κοσμολογίας.

Ο Νανόπουλος είναι ένας άνθρωπος που τα τελευταία 30 χρόνια έχει ταχθεί στην υπηρεσία του μέλλοντος, άρα στην υπηρεσία του ανθρώπου. Μπορεί να μην είναι τόσο γνωστός όσο κάποιοι τραγουδιστές ή τηλεπαρουσιαστές, αλλά είναι όμως ένα από τα πιο ισχυρά επιστημονικά μυαλά της εποχής μας.

Μέχρι σήμερα δύο φορές έχει αγγίξει το βραβείο Νομπέλ αλλά δεν κατάφερε να το αποκτήσει. Ίσως στο μέλλον. Είναι πάντως ένας από τους μόνιμους διεκδικητές του, εξαιτίας των ερευνών του, αλλά πάνω από όλα είναι ένας από αυτούς που έχουν πλησιάσει σε απόσταση αναπνοής το πολυπόθητο όνειρο του ανθρώπου: την ανακάλυψη του Σύμπαντος.

(KPG Script Rater Guide, 2017: 15)

3. Expectations for Activity 1, C1+C2 (writing production)

Candidates are expected to produce a personal response to the poem so as to supposedly post the text as an entry on their personal blog (**genre**), in which they explain what this poem means to them (if they think it's important in some way, if they agree or disagree with it) perhaps providing examples from personal experience. In the same text, they explain what message(s) this poem sends out to parents and teachers who work with young people (**topic and communicative purpose)**. Their script should have an **informal style** and a **personal tone**. In terms of **organization**, we expect this script to consist of at least two two paragraphs – one explaining their personal response to the poem and the other explaining what the message the poem sends to parents and teachers about young people. Candidates are expected to base their script on the ideas presented in the source text – ideas which they should incorpotae into their own scripts creatively, appropriately and effectively. Candidates might also choose to use their own ideas in relation to the topic in question, so long as they do not exceed the word limit.

(KPG Script Rater Guide, 2017: 14)

4. Expectations for Activity 2, C1+C2 (written mediation)

Candidates are asked to use information from the Greek source text in order to write an online encyclopedic entry about the life and achievements of Dimitris Nanapoulos (**genre and topic**). In their script they should present basic facts about his life (date of brith, where he was born, studies, professional appointments, major achievements) and explain why he is famous (**communicative purpose**). As their script is expected to have the form of an encyclopedic entry, it must be precise, specific, objective and neutral. It must be written in a rather formal register. In terms of organization, one would expect 3

> paragraphs: one on basic life facts and studies, one on professional achievements and one on why he is considered a famous Greek. As this is a mediation activity, candidates are expected to extract the necessary information from the Greek text and to relay it, so as to achieve the communicative goal. Of course, candidates may resort to their own ideas and experiences as long as they are conducive to the topic in question and they do not exceed the set word limit. However, even a linguistically satisfactory text which has limited or no information from the Greek text must be considered partly satisfactory or unsatisfactory for the level in terms of the specific criterion.

As for the overall language performance, candidates are expected to produce a script that flows. Both accuracy and appropriacy errors are taken into account. Ideas in/among paragraphs should be cohesively and coherently linked with a variety of appropriate linking devices (coherence and cohesion). Additionally, candidates are expected to employ a wide range of sophisticated vocabulary and complex grammatical and syntactic structures appropriate for the communicative purpose of each task, while maintaining a high degree of grammatical accuracy (vocabulary range – accuracy and appropriacy). Finally, candidates are expected to maintain high levels of accuracy as far as their spelling and punctuation are concerned (spelling and punctuation).

(KPG Script Rater Guide, 2017: 14)

5. Affective candidate response to Activity 2 with (anti)patriotic connotations.

A Greek miracle.

Nowadays, it is a sad reality that we are most impressed by singes or actors and we ignore successful scientists like Dimitris Nanopoulos. His life seems imaginary and his skills supernatural, so if you don't believe, reade below.

He was born in Athens in 1948 and he was raised in Zografou. Then, he studied in University of Athens, in the department of physics. After that, he had a master degree in England and finally took a phD in high energy physics. He worked in many researches, like the world-known experiment in Cern, and also in Paris and USA. In 1989, he became professor in the University of Texas and he still teaches there. His brilliant career doesn't stop in these achievements. He is also director of the Houston Advanced Research Centre. To be more specific, he is in charge of the World Laboratory in Lausanne. The issues he investigates re the high energy physics and cosmology. Moreover, he has written fifteen books and has published many of his works in global scientific magazines. In 1997 he first became member in Academy of Athens and in 2015 he totally became president. Although, when he had the opportunity to return to Greece as a professor in a Greek University, his colleagues refused to vote in favour.

The undeniable truth is that he is not alone. Most scientific make huge careers abroad and Greece usually recognizes their offer after their death. That happens because Greece seems so "old-fashioned". It' is hard to accept new ideas and the most important to accept the disadvantages of not being a fair country. Nanopoulos has dominated twice to win the Oscar but is not capable enough to teach greek students. In Greece, it is common to admire people who doesn't deserve it because they managed to be on television. We are mostly impressed by the glorious and not by the worthy. It's depressing the fact that in country where philosophy and democracy were born, we refuse to accept minds like these because they are better than us.

It is undoubtedly truth that people who live in Greece love this country but they should keep on their minds that they have to honour everyone and everything that exerts a positive influence in the future of next generations. Innovative ideas will bring progress in Greece. We should not forget it.

Appendix B

1. Descriptive statistics, 66-script corpus

Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
Word Count Activity 1	33	136	559	11420	346,06	16,178	92,935	-,534	,409	,775	,798
Word Count Activity 2	33	167	456	9928	300,85	10,584	60,798	,090	,409	,919	,798
Activity 1, Rater 1	33	2	5	103	3,12	,149	,857	,390	,409	-,347	,798
Activity 1, Rater 2	33	2	5	111	3,36	,173	,994	,200	,409	-,922	,798
Task Completion mean Activity 1	33	2,00	4,50	107,00	3,2424	,13954	,80157	,199	,409	-,858	,798
Activity 2, Rater 1	33	2	5	101	3,06	,157	,899	,150	,409	-1,192	,798
Activity 2, Rater 2	33	2	5	100	3,03	,147	,847	,597	,409	,030	,798
Task Completion mean Activity 2	33	2,00	4,50	100,50	3,0455	,14163	,81359	,215	,409	-,987	,798
Spelling and Punctuation Rater 1	33	2	5	116	3,52	,145	,834	-,223	,409	-,395	,798
Spelling and Punctuation Rater 2	33	2	5	101	3,06	,168	,966	,537	,409	-,630	,798
Spelling and Punctuation Mean	33	2,00	5,00	108,50	3,2879	,13938	,80069	,170	,409	-,770	,798
Vocabulary Range Rater 1	33	2	5	116	3,52	,164	,939	-,287	,409	-,764	,798
Vocabulary Range Rater 2	33	2	5	116	3,52	,152	,870	-,049	,409	-,538	,798
Vocabulary Range Mean	33	2,00	4,50	116,00	3,5152	,13673	,78546	-,516	,409	-,766	,798
Grammaticality/ Accuracy Rater 1	33	1	4	88	2,67	,161	,924	-,019	,409	-,834	,798
Grammaticality/ Accuracy Rater 2	33	1	4	79	2,39	,144	,827	-,522	,409	-,737	,798

Grammaticality/ Accuracy Mean	33	1,00	3,50	83,50	2,5303	,13923	,79980	-,349	,409	-1,054	,798
Appropriacy Rater 1	33	2	5	97	2,94	,130	,747	,578	,409	,485	,798
Appropriacy Rater 2	33	2	4	91	2,76	,107	,614	,178	,409	-,427	,798
Appropriacy Mean	33	2,00	4,00	94,00	2,8485	,10545	,60576	,180	,409	-,582	,798
Cohesion and Coherence Rater 1	33	2	5	125	3,79	,149	,857	-,513	,409	-,068	,798
Cohesion and Coherence Rater 2	33	2	4	100	3,03	,119	,684	-,038	,409	-,726	,798
Cohesion and Coherence Mean	33	2,00	4,50	112,50	3,4091	,11607	,66679	-,488	,409	-,473	,798
Rater 1 Task Completion Mean	33	2,00	4,50	102,00	3,0909	,13668	,78516	,041	,409	-1,066	,798
Rater 2 Task Completion Mean	33	2,00	5,00	105,50	3,1970	,13923	,79980	,376	,409	-,486	,798
Rater 1 Overall Language Performance Mean	33	2,00	4,80	108,40	3,2848	,13315	,76490	-,116	,409	-,622	,798
Rater 2 Overall Language Performance Mean	33	1,80	4,00	97,40	2,9515	,11147	,64037	-,216	,409	-,864	,798
Task Completion Mean	33	2,00	4,50	103,75	3,1439	,12736	,73162	,121	,409	-,858	,798
Overall Language Performance Mean	33	1,90	4,20	102,90	3,1182	,11193	,64297	-,326	,409	-,889	,798
Valid N (listwise)	33										

2. Coh-metrix quantitative analysis of the 16 scripts corpus

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
DESPC	3.800	1	.164	38.00000	-89.0620	165.0620
DESSC	26.400	1	.024	132.00000	68.4690	195.5310
DESWC	11.174	1	.057	2497.50000	-342.3368	5337.3368
DESPL	4.394	1	.142	3.69500	-6.9909	14.3809
DESPLd	2.996	1	.205	2.83850	-9.2006	14.8776
DESSL	7.879	1	.080	19.01200	-11.6481	49.6721
DESSLd	5.377	1	.117	11.28450	-15.3795	37.9485
DESWLsy	37.667	1	.017	1.46900	.9735	1.9645
DESWLsyd	26.397	1	.024	.83150	.4313	1.2317
DESWLlt	36.387	1	.017	4.33000	2.8180	5.8420
DESWLltd	53.600	1	.012	2.41200	1.8402	2.9838
PCNARz	6.650	1	.095	.81800	-.7449	2.3809
PCNARp	22.116	1	.029	79.06500	33.6403	124.4897
PCSYNz	-2.619	1	.232	-.29600	-1.7318	1.1398
PCSYNp	8.774	1	.072	38.47500	-17.2417	94.1917
PCCNCz	-2.781	1	.220	-.50750	-2.8264	1.8114
PCCNCp	4.788	1	.131	30.98000	-51.2291	113.1891
PCREFz	1.207	1	.440	.10500	-1.0004	1.2104
PCREFp	15.137	1	.042	53.96500	8.6674	99.2626
PCDCz	3.265	1	.189	.87500	-2.5303	4.2803
PCDCp	10.860	1	.058	79.93000	-13.5877	173.4477
PCVERBz	4.241	1	.147	.50250	-1.0032	2.0082
PCVERBp	16.355	1	.039	69.02000	15.3998	122.6402
PCCONNz	-8.898	1	.071	-2.80300	-6.8055	1.1995
PCCONNp	1.316	1	.414	.37500	-3.2463	3.9963
PCTEMPz	-1.245	1	.431	-.48850	-5.4757	4.4987
PCTEMPp	2.379	1	.253	32.67500	-141.8447	207.1947
CRFNO1	3.230	1	.191	.26650	-.7818	1.3148
CRFAO1	12.208	1	.052	.61650	-.0252	1.2582
CRFSO1	3.264	1	.189	.37050	-1.0717	1.8127

CRFNOa	7.125	1	.089	.17100	-.1339	.4759
CRFAOa	34.517	1	.018	.50050	.3163	.6847
CRFSOa	4.770	1	.132	.26950	-.4484	.9874
CRFCWO1	33.000	1	.019	.11550	.0710	.1600
CRFCWO1d	15.714	1	.040	.11000	.0211	.1989
CRFCWOa	18.556	1	.034	.08350	.0263	.1407
CRFCWOad	12.600	1	.050	.09450	-.0008	.1898
LSASS1	11.200	1	.057	.16800	-.0226	.3586
LSASS1d	13.636	1	.047	.15000	.0102	.2898
LSASSp	19.429	1	.033	.13600	.0471	.2249
LSASSpd	18.200	1	.035	.13650	.0412	.2318
LSAPP1	5.407	1	.116	.29200	-.3941	.9781
LSAPP1d	23.500	1	.027	.14100	.0648	.2172
LSAGN	27.348	1	.023	.31450	.1684	.4606
LSAGNd	27.857	1	.023	.09750	.0530	.1420
LDTTRc	239.000	1	.003	.47800	.4526	.5034
LDTTRa	174.333	1	.004	.26150	.2424	.2806
LDMTLD	111.529	1	.006	64.96550	57.5641	72.3669
LDVOCD	30.201	1	.021	98.18200	56.8741	139.4899
CNCAll	16.598	1	.038	102.94850	24.1383	181.7587
CNCCaus	8.887	1	.071	30.72200	-13.2033	74.6473
CNCLogic	9.187	1	.069	44.41250	-17.0156	105.8406
CNCADC	25.213	1	.025	16.07350	7.9733	24.1737
CNCTemp	22.532	1	.028	17.94650	7.8260	28.0670
CNCTempx	3.169	1	.195	19.27550	-58.0100	96.5610
CNCAdd	9.337	1	.068	60.08500	-21.6794	141.8494
SMCAUSv	12.812	1	.050	17.54000	.1452	34.9348
SMCAUSvp	47.348	1	.013	29.28450	21.4257	37.1433
SMINTEp	685.044	1	.001	15.41350	15.1276	15.6994
SMCAUSr	7.043	1	.090	.66200	-.5324	1.8564
SMINTEr	25.226	1	.025	1.67750	.8325	2.5225
SMCAUSlsa	9.476	1	.067	.09950	-.0339	.2329
SMCAUSwn	10.625	1	.060	.42500	-.0832	.9332
SMTEMP	18.553	1	.034	.78850	.2485	1.3285
SYNLE	4.080	1	.153	3.16200	-6.6853	13.0093
SYNNP	23.407	1	.027	.69050	.3157	1.0653
SYNMEDpos	1325.000	1	.000	.66250	.6561	.6689

SYNMEDwrd	176.200	1	.004	.88100	.8175	.9445
SYNMEDlem	171.800	1	.004	.85900	.7955	.9225
SYNSTRUTa	15.727	1	.040	.08650	.0166	.1564
SYNSTRUTt	43.667	1	.015	.06550	.0464	.0846
DRNP	525.858	1	.001	389.92350	380.5018	399.3452
DRVP	30.680	1	.021	214.64000	125.7474	303.5326
DRAP	12.413	1	.051	30.61000	-.7235	61.9435
DRPP	17.273	1	.037	120.89500	31.9643	209.8257
DRPVAL	8.368	1	.076	6.67800	-3.4616	16.8176
DRNEG	3.979	1	.157	8.22350	-18.0339	34.4809
DRGERUND	3.516	1	.176	10.27150	-26.8497	47.3927
DRINF	24.999	1	.025	24.73650	12.1637	37.3093
WRDNOUN	25.136	1	.025	235.39800	116.4044	354.3916
WRDVERB	22.249	1	.029	125.83150	53.9716	197.6914
WRDADJ	17.601	1	.036	67.41000	18.7452	116.0748
WRDADV	18.879	1	.034	53.40050	17.4610	89.3400
WRDPRO	8.807	1	.072	116.49900	-51.5787	284.5767
WRDPRP1s	4.745	1	.132	18.16050	-30.4725	66.7935
WRDPRP1p	3.130	1	.197	16.15500	-49.4217	81.7317
WRDPRP2	2.035	1	.291	14.44900	-75.7523	104.6503
WRDPRP3s	6.205	1	.102	44.68850	-46.8279	136.2049
WRDPRP3p	8.888	1	.071	10.90150	-4.6827	26.4857
WRDFRQc	55.556	1	.011	2.50000	1.9282	3.0718
WRDFRQa	494.077	1	.001	3.21150	3.1289	3.2941
WRDFRQmc	10.872	1	.058	1.35900	-.2293	2.9473
WRDAOAc	18101.000	1	.000	362.02000	361.7659	362.2741
WRDFAMc	313.812	1	.002	582.59250	559.0034	606.1816
WRDCNCc	349.860	1	.002	348.11100	335.4683	360.7537
WRDIMGc	792.466	1	.001	394.64800	388.3203	400.9757
WRDMEAc	389.392	1	.002	433.19850	419.0628	447.3342
WRDPOLc	17.361	1	.037	3.75000	1.0055	6.4945
WRDHYPn	83.167	1	.008	5.48900	4.6504	6.3276
WRDHYPv	51.444	1	.012	1.38900	1.0459	1.7321

WRDHYPnv	51.364	1	.012	1.41250	1.0631	1.7619
RDFRE	11.005	1	.058	63.26050	-9.7811	136.3021
RDFKGL	6.537	1	.097	9.15900	-8.6424	26.9604
RDL2	14.269	1	.045	21.81700	2.3892	41.2448

Περίληψη

Ένα μεγάλο μέρος της έρευνας στον τομέα της αξιολόγησης γραπτών κειμένων στη δεύτερη γλώσσα (Γ2) έχει επικεντρωθεί στην αντιστοίχιση γλωσσικών χαρακτηριστικών του κειμένου με την απόδοση των υποψηφίων, όπως αυτή κρίνεται από τους βαθμολογητές. Τα θέματα που έχουν διερευνηθεί αφορούν στην ποιότητα του γραπτού κειμένου και στην αξιοπιστία μεταξύ βαθμολογητών.

Ωστόσο, εξαιτίας υφιστάμενων προβλημάτων, όπως η ασαφής διατύπωση της βαθμολογικής κλίμακας (Lumley, 2002) και η δυσκολία που συναντούν οι βαθμολογητές όταν τα γραπτά (π.χ. δοκίμια) βρίσκονται στο όριο ανάμεσα στα διάφορα επίπεδα (Gebril & Plakans, 2014), πολύ λίγα έχουν βρεθεί σχετικά με τις λεπτές διαφορές μεταξύ διπλανών επιπέδων γλωσσικής επάρκειας στην Αγγλική γλώσσα, και πιο συγκεκριμένα στο επίπεδο Γ (Γ1-Γ2).

Η παρούσα έρευνα είχε σκοπό να εξερευνήσει τη βαθμολογημένη απόδοση στην παραγωγή γραπτού λόγου υποψηφίων του Κρατικού Πιστοποιητικού Γλωσσομάθειας (ΚΠΓ), στο επίπεδο Γ (ΓΙ-Γ2). Ειδικότερα, σε αυτό το επίπεδο γίνεται χρήση δύο γραπτών δοκιμασιών 'read-to-write', μία ενδο-γλωσσικής διαμεσολάβησης και μία δια-γλωσσικής διαμεσολάβησης. Χρησιμοποιώντας 66 βαθμολογημένα γραπτά (33 υποψήφιοι), τρείς ήταν οι ερευνητικοί τομείς: α) η πιθανή επίδραση του κειμενικού είδους στη γλωσσική παραγωγή (ένα blog με ερμηνευτικό χαρακτήρα και μία εγκυκλοπαιδική καταχώρηση με ερμηνευτική διάσταση), β) οι βαθμολογικές διαφορές μεταξύ αξιολογητών και γ) η συνοχή και συνεκτικότητα ως

πιθανό διαφοροποιητικό γλωσσικό κριτήριο της απόδοσης των υποψηφίων σε επίπεδο Γ, δηλαδή μεταξύ των επιπέδων Γ1 και Γ2.

Τα αποτελέσματα της ποσοτικής ανάλυσης δείχνουν ότι τα γλωσσικά κριτήρια Συνοχή & Συνεκτικότητα και Εύρος Λεξιλογίου, σε μεγαλύτερο και μικρότερο βαθμό αντίστοιχα, προσφέρονται για την ανίχνευση διαφορών ανάμεσα στα δύο επιμέρους επίπεδα, Γ1 και Γ2. Μάλιστα, η προβλεπτική τους ισχύ ενισχύεται όταν συνδυάζονται με το γλωσσικό κριτήριο Καταλληλότητα Λόγου, αποφέροντας μία πιο ολοκληρωμένη εικόνα των παρατηρημένων αποκλίσεων (αναφορικά με τις δοκιμασίες).

Η ποιοτική ανάλυση επίσης επεσήμανε κάποιες διαφορές στις κρίσεις των βαθμολογητών, οι οποίες όμως κρίθηκαν αμελητέες. Ιδιαίτερο ενδιαφέρον ήταν, ωστόσο, το ότι κατά τη διερεύνηση των διαφορών μεταξύ δοκιμασιών, το Εύρος Λεξιλογίου αντιστοιχούσε στη δοκιμασία ενδο-γλωσσικής διαμεσολάβησης και το κριτήριο Γραμματικότητα/Ορθότητα Λόγου στη δεύτερη δοκιμασία, τη διαγλωσσική διαμεσολάβηση.

Το κριτήριο Συνοχή & Συνεκτικότητα αναλύθηκε και ποιοτικά. Σύμφωνα με τα ευρήματα τα οποία προέρχονται από την χρήση του Coh-Metrix, κρίνεται ότι είναι δυνατή η διαφοροποίηση μεταξύ Γ1 και Γ2 απόδοσης σε συνθήκες εξέτασης. Οι δείκτες που προσφέρονται για αυτή τη διερεύνηση αφορούν κυρίως σε επιφανειακές διαφορές και, ειδικότερα, σε πληροφορίες σχετικά με την ποιότητα των λέξεων.

Τα ευρήματα τα οποία πηγάζουν από αυτή την εργασία μπορούν να συνεισφέρουν στην έρευνα η οποία διεξάγεται από το ΚΠΓ σχετικά με την ανάλυση των δοκιμασιών, καθώς και με τα συνεχιζόμενα προγράμματα εκπαίδευσης βαθμολογητών. Επιπροσθέτως, τα ευρήματα υποστηρίζουν την πολιτική του ΚΠΓ να αποδίδει διαφορετικό συντελεστή βαρύτητας στα κριτήρια αξιολόγησης ανάλογα με το γλωσσικό επίπεδο. Αυτό που προτείνεται είναι να λαμβάνονται υπ' όψιν και οι υπό εξέταση κειμενικές λειτουργίες.