BRITISH
COUNCIL

ENGLISH LANGUAGE
ASSESSMENT RESEARCH GROUP

INVESTIGATING THE DISCOURSE PRODUCED AT SCORE LEVELS B2.2 TO C2 ON THE APTIS ADVANCED WRITING TEST

AR-G/2021/5

Ute Knoch, Jason Fan, Cathie Elder, Ksenia Zhao & Andrew Pitman
Language Testing Research Centre, University of Melbourne

ARAGs RESEARCH REPORTS ONLINE
SERIES EDITORS: RICHARD SPIBY AND CAROLYN WESTBROOK
ISSN 2057-5203

ABSTRACT

Adopting an exploratory sequential mixed-methods design, this study explored the features of test-takers' writing samples on the two tasks in the Aptis Advanced writing test, that is, email response and website writing that distinguished the three levels of B2.2, C1 and C2 on the CEFR.

The study consisted of a qualitative and quantitative phase. During the qualitative phase, we conducted focus groups with ESL experts (n = 5) and Aptis raters (n = 6) where they commented on the features of the writing samples at the three CEFR levels in focus. During the quantitative phase, we performed discourse analysis of the email response and website writing samples (n = 120).

Regarding the email response task, the two groups of participants identified five discourse features that distinguished writing samples at the three levels: a) vocabulary and grammar; b) sociolinguistic features; c) content; d) cohesion and coherence; and e) orthographic control. Most of these findings, however, were not supported by the discourse analysis results which indicated that the email writing samples at the three levels in focus failed to demonstrate significantly different discourse features.

When it came to the website writing task, very similar findings emerged from the focus groups with the two groups of participants. However, the discourse analysis results for both tasks indicated that, except for lexical sophistication and spelling errors, other discourse features did not distinguish the writing samples at the three levels in focus.

The findings of this study are not entirely surprising, given that: a) the differences in writing samples at advanced levels are often nuanced and not easily quantifiable through discourse analysis; and b) the writing samples that we drew on in this study were evaluated by the Aptis raters using the existing rating criteria which lack details at high levels. In view of the findings of this study, we provide a few recommendations for the possible revisions of the rating scales for the two tasks in the Aptis Advanced writing test.

Authors

Associate Professor Ute Knoch is the Director of the Language Testing Research Centre at the University of Melbourne. Her research interests are in the areas of writing assessment, rating processes, assessing languages for academic and professional purposes, and placement testing. She is the co-author of *Fairness, Justice and Language Assessment* (2019, OUP, with Tim McNamara and Jason Fan), and *Assessing English for Professional Purposes* (2020, with Susy Macqueen). Ute was the co-president of the Association for Language Testing and Assessment of Australian and New Zealand (ALTAANZ) from 2015–16 and has been serving on the Executive Board of the International Language Testing Association (ILTA) from 2011 to 2014 and from 2017 to 2019.

Dr Jason Fan is the Deputy Director and Senior Research Fellow at the Language Testing Research Centre at the University of Melbourne. His research interests include language test validation, language assessment literacy, and research methodology. He is the co-author of *Fairness, Justice and Language Assessment*' (2019, OUP, with Tim McNamara and Ute Knoch), and *Development and Validation of Standards in Language Testing* (2018, Fudan University Press). Jason is the Chair of the Nominating Committee of the Asian Association for Language Assessment (AALA) (from 2018) and on the Executive Board of the International Language Testing Association (ILTA) (2020–21). He is the co-editor (with John Read) of the journal, *Papers in Language Testing and Assessment* (from 2021).

Cathie Elder is Principal Fellow in the School of Languages and Linguistics at the University of Melbourne and former Director (2007–2012) of the Language Testing Research Centre. She has held senior academic positions at Monash University (2004–2006) and the University of Auckland (2000–2004). Cathie served as chair of the TOEFL Committee of Examiners at the Educational Testing Service in Princeton from 2006 to 2008 and was co-editor of *Language Testing* from 2007 to 2011. She was a founding member and co-president (2013) of the Australian Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) and president of the International Language Testing Association (ILTA) (2017–18). Cathie has been teaching and researching in the areas of language testing assessment and program evaluation for nearly 30 years, publishing widely in peer-reviewed journals and edited volumes, and contributing to major reference works in the field.

Ksenia Zhao is a test development specialist at the Language Testing Research Centre (LTRC), University of Melbourne. Her academic and professional interests include language assessment design and analysis, rating scales development and validation, as well as rater behaviour. She completed her TESOL MA at Brigham Young University, Utah, USA in 2017 with a high GPA award. Her MA thesis focused on rater bias in speaking assessment. Ksenia has been working in language testing since 2016. Her current research at the LTRC includes assessment development and analysis for various internal and external projects, as well as research into scale and self-assessment development. Ksenia will commence her PhD in 2021.

Andrew Pitman is a research fellow at the Language Testing Research Centre, University of Melbourne. He has experience as an English language teacher in first language, additional/foreign language, multicultural and culturally homogenous contexts. From 2016 to 2018 he taught in Japan at a private language school and a prestigious high school, while examining for Cambridge PET and KET. He returned to Australia in mid-2018 and completed his Master of Applied Linguistics at the University of Melbourne, receiving first-class honours for his minor thesis and all coursework components. His minor thesis focused on language testing validation and successful, graduate-level L2 university writing. Current research activities include language test review, validation and development projects. Andrew will commence his PhD in 2021.

CONTENTS

List of tables

List of figures

1. INTRODUCTION

Differentiating between advanced L2 writers at the higher levels of the Common European Framework of Reference for Languages (CEFR) presents a challenge for language testers given the relatively slight differences in the descriptors for these levels compared to those at the lower levels of the scale, and the fact that higher level performances are often differentiated based on rhetorical skills and expression of complex ideas which may be less well understood or defined. Developers of tests of advanced writing proficiency which reference these upper levels of the CEFR face the daunting task of developing rating criteria which discriminate between levels in the absence of both explicit guidance from the CEFR and substantial research targeting the later stages of L2 writing development. The problem is compounded by the fact that the top C2 level of the scale is not consistently achieved by native speakers, whose performance cannot therefore be used as a benchmark for judging the highest level of performance (Hamilton, Lopes, McNamara & Sheridan, 1993).

Rating scales are often seen as the *de facto* test constructs in writing or speaking assessments (McNamara, Hill & May, 2002). However, because most scales are developed using intuitive methods (Fulcher, 2014), it is important to validate the scoring levels on a rating scale empirically. Language proficiency scales are linear and therefore should refer to aspects of candidates' writing that develop in a linear fashion. However, there is sufficient evidence from studies in second language acquisition that not all aspects of interlanguage develop in this linear way (e.g., Meisel, Clahsen, & Pienemann, 1981) and this is particularly true for writing skills which are generally acquired in the highly variable contexts of formal education. In addition, at the upper levels of proficiency the boundary between language skills and cognitive abilities is unclear. Given this lack of clarity, Banerjee, Franceschina and Smith (2007) argue that the aim should be to identify an ideal group of measures that, when applied, produce a learner profile that could reliably be classified as being at a given level in a pre-determined scale. While previous studies conducted by Cumming et al. (2006), Banerjee et al. (2007), Banerjee, Yan, Chapman and Elliott (2015), Knoch (2011), and Knoch, Macqueen and O'Hagan (2014) have made some headway in this regard, there is, as noted above, a dearth of research targeting the abilities of more advanced learners. This is all the more pressing given recent research showing that raters may experience difficulty in making distinctions between learners at the upper end of the proficiency scale (e.g., Banerjee et al., 2015; Isbell, 2017). Greater specification of these upper scale levels is therefore needed.

Building on the relatively small body of test-driven research into the writing ability of second language learners at higher proficiency levels, this study explored the features of writing produced on the Aptis Advanced writing test at different scoring levels, particularly targeting the B2.2 to C2 range. The study has the aim to establish what features of writing are typical of the B2.2, C1 and C2 levels of performance on the Aptis Advanced writing test and make specific recommendations back to test developers in terms of what changes could be made to the task-specific rating scales. The results of the analysis can be used to verify the scoring levels at the upper levels of the CEFR scale, where differences in text quality may be hard to define and discern and can also play a role in the refinement of the scale for rater training and reporting purposes.

Specifically, this study set out to investigate the following two research questions:

1. What discourse features distinguish writing score levels at the upper end of the CEFR scale (i.e., B2.2, C1 and C2)?
2. Do any of these features relate to specific Aptis Advanced tasks?

2. REVIEW OF LITERATURE

2.1 Features of advanced L2 writing

Despite the abundance of research into L2 writing, limited research is currently available focusing specifically on advanced L2 writing and/or writers, or those at B2.2, C1, and C2 on the CEFR (the group of L2 learners that this project targets). In particular, there is a dearth of studies comparing the writing of learners at the very highest proficiency levels. The bulk of the existent research compares the linguistic features of the writings produced by L2 writers at different score levels of major proficiency tests. These studies have usually grouped students into broader categories to conduct the comparisons (e.g., three groupings spanning the whole range of possible score levels on tests such as IELTS and TOEFL iBT).

Despite the broader groupings, research along this line can still provide us with an indication of the features of advanced writing. For example, research indicates that the writing produced by advanced L2 writers is syntactically more complex, with fewer errors. For instance, Mayor, Hewings, North, Swann and Coffin (2007) investigated the linguistic features of test-takers' writing on the Academic Writing Task 2 component in the International English Language Testing System (IELTS). One of the primary objectives of this study was to identify the linguistic features which could effectively distinguish high- and low-scoring scripts. Five types of analysis were conducted of the scripts, including error analysis, analysis of sentence structure, analysis of argument structure at both sentence and discourse level, and analysis of tenor and interpersonal meaning. Results indicate that high-scoring scripts were significantly longer with fewer errors than low-scoring scripts. Grammar errors were the most frequent among low-scoring scripts while errors in lexis/idiom were the most frequent among high-scoring scripts. In addition, the study revealed that high-scoring candidates tended to draw on different combinations of grammatical structure to demonstrate effective use of English, whereas low-scoring candidates tended to use 'incomplete' argument structures with limited evidence in support of their arguments.

These findings generally resonate with Guo, Crossley and McNamara (2013), who investigated the predictive power of different linguistic features of human judgment of essay quality in the context of both the integrated and independent writing task in the Test of English as a Foreign Language (TOEFL iBT). Results indicate that, in the case of integrated essays, text length, the use of past participle verbs and third person singular verbs, and word familiarity (a measure of lexical sophistication) strongly predicted essay score assigned. Similar to what was found about integrated essays, text length, the use of participle verbs, and two measures of lexical sophistication (i.e., average syllables per word and noun hypernymy values) proved to be strong predictors of independent essay scores.

Knoch, Macqueen and O'Hagan (2014), drawing on the same data set as Guo et al (2013), examined the linguistic features of TOEFL iBT test-takers' writing samples at different score levels and across the two writing tasks (independent and integrated tasks). They drew on a large number of discourse measures to compare score levels and tasks. The findings showed that measures of accuracy (e.g., proportion of error-free clauses and T-units), fluency (number of words), accurate use of lexical patterns (co-occurring lexical bundles), lexical sophistication, and number of ideas were the best at differentiating different proficiency levels. In terms of differentiating between different task types, measures of fluency, lexical sophistication, and the use of voice were most effective. The authors made several recommendations for possible revisions to the TOEFL iBT writing rubrics.

Thewissen (2013) provided more empirical evidence regarding syntactic complexity and error types of L2 writers. The author analysed the errors made by L2 learners at four levels on the CEFR: B1 (lower intermediate), B2 (higher intermediate), C1 (advanced) and C2 (near-native), with a view to capturing the developmental patterns of L2 learners' accuracy. An overall pattern identified by this study is that L2 learners' accuracy improved with the increase of their language proficiency levels. This was indicated by the smaller number of errors made by more proficient learners. The study also found that lexis developed most strongly from the intermediate (i.e. B1 and B2) to advanced levels (i.e. C1 and C2). Having said that, the study also indicated that certain error types tended to persist with the development of language proficiency. More specifically, spelling errors, morphological errors, uncountable noun errors, verb number errors and erroneously pluralised adjectives were the typical errors which tended to affect the writing of learners at B2, C1 and C2 levels.

Lexical sophistication represents one of the crucial elements in advanced L2 writing and has been recurrent in the literature. A corpus-based study by Staples, Egbert, Biber and McClair (2013) examined one aspect of lexical sophistication – the use of three or four word recurrent sequence, known as lexical bundles, in L2 writing. The corpus of this study included a total of 960 written responses across three proficiency levels (low, intermediate and high) on the writing part of the TOFEL iBT. The study was designed to explore whether the frequency, function, and fixedness of lexical bundles used by EFL learners vary across proficiency levels. Results show that learners at the lowest proficiency level used significantly more bundles overall; however, it was revealed that these learners tended to use the bundles which were identical to those in the prompts, while learners at the higher proficiency levels relied less on these unanalysed chunks and were able to draw on formulaic language from outside of the test prompts. Learners at the highest proficiency levels used fewer bundles overall, but they were more skilled in developing their own formulaic sequences in their writing.

Cohesion is another feature which has been investigated. Crossley, Kyle and McNamara (2016), for instance, investigated the development of local, global and text cohesion in the writings produced by 57 L2 learners attending an upper-level English for Academic Purposes (EAP) course. The corpus used in the analysis included 171 descriptive essays which students wrote on a timed writing task. To study the development of cohesion in students' writing, the writing samples were collected at the beginning, middle, and end of the EAP course, with each student writing one essay at each time point. The study identified the growth in the use of cohesive devices in L2 writing over the course of the semester. In addition, the study indicates that the use of cohesion devices significantly predicted both the ratings of organisation and the overall quality of the essays.

Yang and Sun (2012) investigated the use of cohesive devices in argumentative writing by Chinese learners of English in second- and fourth-year at university, representing learners at intermediate and advanced level respectively. The study identified much variation in the use of cohesive devices among intermediate and advanced L2 learners. For instance, compared with advanced learners, intermediate learners used significantly more personal references and demonstrative pronouns. Furthermore, it was also found that the errors in the use of cohesive devices generally decreased as learners' proficiency increased, which concurred with the studies we reviewed earlier (Mayor et al., 2007). Similar to Crossley et al. (2016), the study revealed a significant positive relationship between the overall use of cohesive devices and students' writing quality. The findings suggested that cohesion should be adopted as a crucial criterion when it comes to the evaluation of L2 writing, and such a criterion is particularly applicable to more advanced L2 writers.

Specifically focused on advanced level learners, Banerjee et al. (2015) reported a study to revise a rating scale of a large-scale writing assessment designed to target L2 learners at C2 on the CEFR. Four linguistic features were identified based on a review of literature which could consistently distinguish written texts at different proficiency levels: fluency, lexical sophistication, cohesion, and syntactic complexity. Discriminant function analysis revealed that length, lexical diversity, lexical frequency, and cohesion made important contributions to classifying writing samples into different proficiency levels. The study also indicated that the amount of basic morphosyntactic errors distinguished writers at more advanced levels, and that sociolinguistic features (e.g., audience awareness) should be included in the rating scale for more advanced writers, as they needed to demonstrate their writing ability through meaningfully engaging readers in their composing process.

The review of literature identifies the following features that characterise advanced L2 writing.

- **Text length**: For timed writing tasks, text length is an important feature. Text length is found to be significantly correlated with human judgement of writing quality (e.g., Guo et al., 2013; Knoch et al, 2014; Mayor et al., 2007).
- **Syntactic complexity:** Advanced writers tend to use a combination of grammatically complex structures to demonstrate their writing ability (e.g., Banerjee et al., 2015; Mayor et al., 2007).
- **Lexical sophistication:** Compared with basic or intermediate writers, advanced writers tend to use more sophisticated and a wider range of lexicon in their written discourse (e.g., Laufer & Waldman, 2011; Staples et al., 2013). They also tend to rely less on the formulaic sequences in the test prompt; instead, they are more skilled in creating their own formulaic sequences in their writing (e.g., Knoch et al., 2014; Staples et al., 2013).
- **Cohesive devices:** The use of cohesive devices has been found to be positively correlated with human judgement of writing (e.g., Crossley et al., 2016; Yang & Sun, 2012). As argued by Yang and Sun (2012), it is a criterion which is particularly applicable to advanced L2 writers.
- **Sociolinguistic features:** Advanced writers tend to demonstrate their writing ability through meaningfully engaging readers through their composing process. Sociolinguistic features such as audience awareness is therefore an important feature of advanced L2 writing (e.g., Banerjee et al., 2015).

While the findings from the studies reported here generally point to the same features differentiating advanced writing from that of lower levels, it is important to remember that there is a circularity in all these studies. The grouping of the test-takers is always based on the scores awarded by raters (or scoring engines), and therefore based on the features represented in these scales. Similarly, the raters are typically teachers of ESL or EFL from various contexts, and it is therefore not surprising that features such as linguistic accuracy are often found as the strongest predictors of advanced writing.

2.2 Rating scale development and revision

Rating scale development and revision methods have traditionally been divided into two major approaches (e.g., Fulcher, 2003; Fulcher et al., 2011; Fulcher, 2012): 1) those that draw on intuitive methods (e.g., scale developer intuition, existing rating scales); and 2) those that draw on performance data to inform the scale. In recent years, however, this dyadic approach has been questioned by various authors (e.g., Montee & Malone, 2014; Knoch et al.). Knoch et al. reviewed rating scale development and revision studies and identified 11 different sources of scale construct. They grouped these sources into test-external and test-internal sources. Among the sources they identified were expert intuition, language proficiency frameworks (such as the CEFR), target language use domain analysis, existing rating scales, curriculum, and theory/literature review. Test-internal sources included the review/analysis of performance samples, input from raters, and test task features. Some of these sources could be further broken down into various aspects. In their discussion, they argued that while rating scales are generally considered as an operational representation of the test construct, the sources scale developers draw on when developing these scales did not always reflect the wider test construct. For example, their review identified rating scales that were developed for specific academic purposes, that drew on previous rating scales that were not developed with academic language performance in mind. As a result, they argued for a closer consideration of the source of scale construct at the outset of scale development and revision projects.

The Aptis writing rating scales were developed closely mirroring the CEFR writing descriptors. While leaning on a language framework for the Aptis writing sub-test makes sense, due to the tests' general proficiency approach, for the Aptis Advanced, which was developed later, the CEFR descriptors at higher levels provide less detail. In many cases, the descriptors at levels C1 and C2 are identical, and the most recent revision of the CEFR (Council of Europe, 2018) has not added much detail to the writing descriptors at higher levels either; nor does the review of the literature described above provide much more guidance for scale developers, as few studies have examined differences at the very highest proficiency levels (e.g., C1 and C2). For these reasons, we argue that drawing on test-internal sources of scale construct, including rater feedback, performance samples and task types, may provide valuable information to explicate the score criteria at the highest levels.

3. METHODOLOGY AND INSTRUMENTS

The focus of the current study is to ascertain whether the Aptis Writing criteria fully capture the writing at these advanced levels, or whether more detailed, differentiated criteria may be needed to support scoring decisions.

3.1 Context of the study

The Aptis Advanced writing test was developed to provide an accurate assessment of written language production at levels of B2.2 on the CEFR and above. Test-takers respond to three writing tasks detailed in Table 1 below. It should be noted that this study focused on Task B (email response) and Task C (article for website publication) only, as Task A requires relatively short responses from test-takers which does not allow the display of advanced writing ability.

A holistic scale is used to rate both Tasks B (email response) and C (website article). The rating criteria of Task B focus on the following aspects of writing: task fulfilment, register, grammatical range and accuracy, vocabulary range and accuracy, punctuation and spelling, and cohesion. The rating criteria for Task C focus on similar aspects, except that punctuation and spelling are not clearly indicated. Rating criteria for both tasks at CEFR levels B2.2, C1, and C2 are provided in Appendix 1.

Table 1: Task descriptions of the Aptis Advanced writing test

Test design	Task description
Task A Three written parts of a text, all of which require responses.	In this task, test-takers will have a social network-type interaction. They will receive three questions and need to respond.
Task B Email response	In this task, test-takers will read an email from an authority. They need to respond to the email in 120–150 words, using the notes provided, expressing how they feel about the situation.
Task C Article for website publication	In this task, test-takers will write an article for a website that is both informative and interesting. Notes about the topic are provided and test-takers need to use the information to help write the article. The article needs to be between 180 and 220 words.

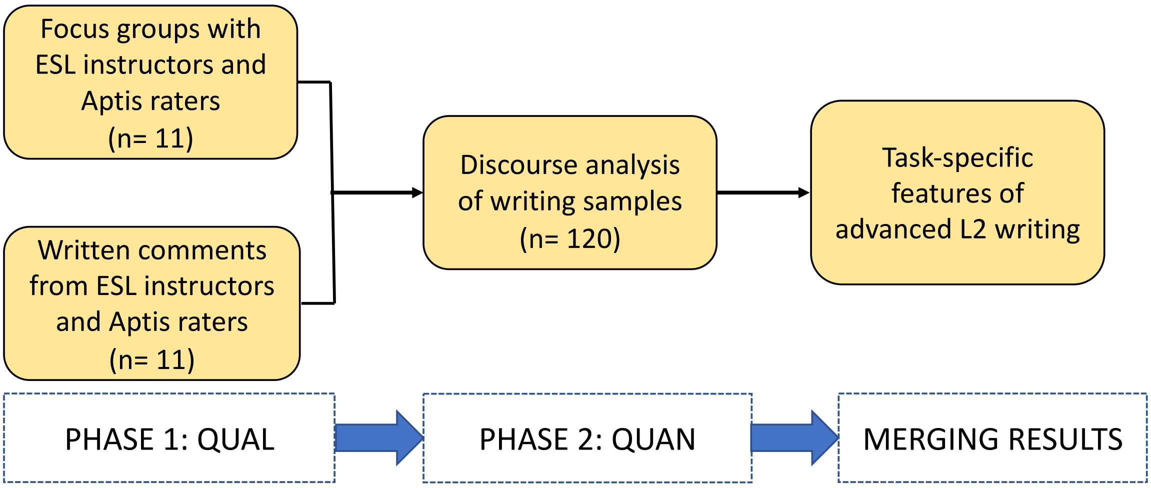
Adapted from British Council (2020, p. 10)

3.2 Overview of the research procedures

As illustrated in Figure 1, this study employed an exploratory sequential mixed-methods design (Creswell, 2013), consisting of a qualitative (Phase 1) and quantitative component (Phase 2).
In Phase 1, we conducted focus groups with ESL experts and Aptis raters (n = 11) to explore the features of email and website writing samples that distinguished the three levels in focus, namely, B2.2, C1, and C2; in addition, we also asked the same participants to write comments on the discourse features of email and website writing samples across the three levels. This phase of the study aimed at: 1) exploring the features of email and website writing that could potentially distinguish performances at the three levels; and 2) informing the selection of features to be included in the discourse analysis of the email and website writing samples at Phase 2. During Phase 2, based on the findings that emerged from Phase 1, we analysed the discourse features of the writing samples (n = 120) to ascertain whether the features identified at Phase 1 were generalisable to a larger number of writing samples.

Findings from both phases, as well as our findings from the literature review, were then integrated to throw light on the revisions of the rating scales for email and website writing tasks in the Aptis Advanced writing test.

Figure 1: Overview of the research procedures and data sources of this study


Focus groups with ESL instructors and Aptis raters (n= 11)
Written comments from ESL instructors and Aptis raters (n= 11)
Discourse analysis of writing samples (n= 120)
Task-specific features of advanced L2 writing
PHASE 1: QUAL
PHASE 2: QUAN
MERGING RESULTS

3.3 Phase 1: Focus groups and written comments

As described above, Phase 1 was designed to inform the discourse analysis study in Phase 2. In the following section, we describe the participants, instruments and procedures employed during Phase 1.

3.3.1 Participants

Five ESL experts (denoted as E1 to E5) and six Aptis raters (denoted as R1 to R6) participated in this study. Profiles of the five participating ESL experts are presented in Table 2. As indicated, except for E5, each participating ESL expert had over five years of experience in teaching university-level ESL writing courses (mean = 16.20; standard deviation = 11.94). With English as their first language, all six Aptis raters had completed their Master's study, with five in either TESOL or Applied Linguistics and one in Business Administration (R6, see Table 3). They all had over five years of experience in teaching ESL writing courses (mean = 16.33, standard deviation = 8.38), and had been working as Aptis raters for over two years (mean = 3.67; standard deviation = 2.07). It is worth noting that all raters except R4 reported that they had been involved in rating other writing assessments apart from Aptis. For example, R2 had five years' and R3 13 years' experience in rating the writing component in IELTS.

Table 2: Profiles of the participating ESL experts (n = 5)

Expert No.	Highest qualification	When completed	ESL writing teaching experience (in years)
E1	PhD	2010	5
E2	MA	1994	30
E3	PhD	2014	18
E4	PhD	2013	25
E5	MA	1998	3

Table 3: Profiles of the participating Aptis raters (n = 6)

Rater No.	Language background	Highest qualification	When completed	Aptis writing rating experience (in years)
R1	English	MA	2018	2
R2	English	MA	2007	2
R3	English	MA	1983	5
R4	English	MA	2015	2
R5	English/Swahili	MA	2009	4
R6	English/Hindi	MA	1997	7

3.3.2 Procedures and instruments

Two researchers conducted focus groups with the ESL experts and Aptis raters separately, in two groups of ESL experts (denoted as EG1 and EG2) and Aptis raters (denoted as RG1 and RG2) respectively (see Table 4).

The procedures for each focus group session are delineated in Figure 2 below. At the start, the ESL experts or Aptis raters were asked to complete a short background questionnaire (see Appendix 2). This was followed by a brief explanation of the purpose of this study and the Aptis Advanced writing test (this was mostly for the ESL experts). Next, the researchers showed the participants two email writing samples at C1 and C2 respectively and asked them to write comments on the discourse features that they perceived as differentiating the two writing samples, and after that, discussed their comments with other participants. The levels of the scripts assigned by the Aptis raters were not made known to the participants. In their discussion, the participants were asked to focus on the features which they believed could distinguish the email writing samples at the two levels. This step was designed to familiarise the participants with the focus group procedures.

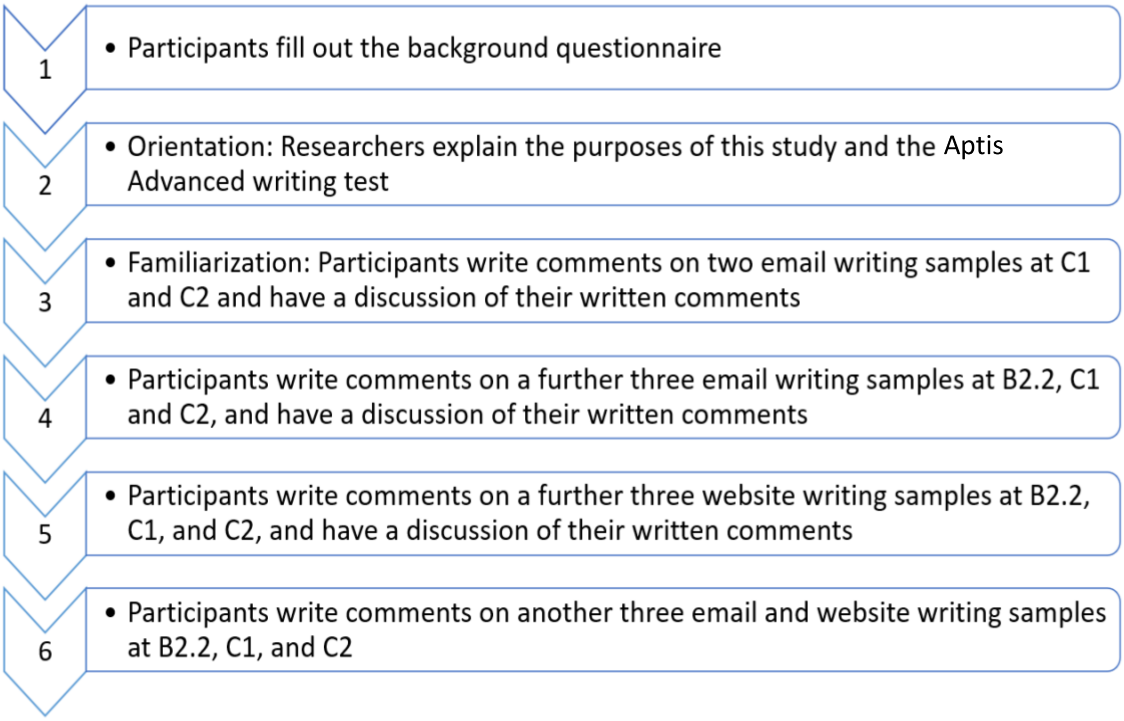
Table 4: The four focus groups in this study (n = 11)

Group ID	Participants	Number of participants	Format
EG1	ESL experts	2	Face-to-face
EG2	ESL experts	3	Face-to-face
RG1	Aptis raters	3	Online
RG2	Aptis raters	3	Online

Once the participants were familiar with the procedures, the researchers asked them to write comments on a further three email writing samples, at B2.2, C1 and C2 respectively, and each group discussed their written comments. Next, the same procedures were repeated, involving a further three website writing samples at the same three levels. Each focus group session lasted for about one hour. As indicated in Table 4, the two focus groups with the ESL experts were conducted face-to-face, as all of them were based at the University of Melbourne. The focus groups with the Aptis raters, however, were conducted on Zoom, an online video-conferencing platform, as they were recruited by the British Council and were located in different parts of the world. All focus group discussions were digitally recorded and subsequently transcribed verbatim.

Following each focus group session, the researchers sent another three email and three website writing samples, at the three levels of B2.2, C1, and C2 respectively to the participants through email so that they could comment on the discourse features of these writing samples on their own. In total, 14 writing samples were used during Phase 1 of this study, including eight email writing samples and six website writing samples. All writing samples used in Phase 1 were provided by the British Council, the owner of the Aptis Advanced writing test, together with their levels that were awarded by certified Aptis raters.

Figure 2: Data collection procedures (Phase 1)



1
- Participants fill out the background questionnaire

2
- Orientation: Researchers explain the purposes of this study and the Aptis Advanced writing test

3
- Familiarization: Participants write comments on two email writing samples at C1 and C2 and have a discussion of their written comments

4
- Participants write comments on a further three email writing samples at B2.2, C1 and C2, and have a discussion of their written comments

5
- Participants write comments on a further three website writing samples at B2.2, C1, and C2, and have a discussion of their written comments

6
- Participants write comments on another three email and website writing samples at B2.2, C1, and C2

3.3.3 Coding and analysis of data: Focus groups and written comments

All qualitative data in this study was coded in NVivo 12 (QSR, 2012). In our data analysis, we noticed that in the focus group discussions, the participants were inclined to compare the discourse features of email and website writing samples either between two adjacent levels (e.g., B2.2 and C1) or across the three levels in focus; however, when it came to written comments, they tended to focus quite exclusively on the discourse features of the specific writing samples that were assigned to them. Due to the different foci of these two data sources, we decided to code the focus group and written comment data separately. The findings that emerged from coding these two sources of data were subsequently used to either triangulate or complement each other.

The open coding method (Miles, Huberman & Saldaña, 2014) was employed in NVivo 12 to explore the discourse features of email and website writing samples on which the participants commented across the three levels in focus. The coding schemes were generated through reading the data repeatedly to identify the recurrent themes or patterns in the data (see Appendix 3). These themes were then organised into a few broader categories based on their foci and also by referring to some relevant literature in L2 writing (e.g., Council of Europe, 2001; Cummings, et al., 2006). Two researchers coded the data in NVivo 12. Specifically, the two researchers independently coded the data from two focus groups (one with ESL experts and the other Aptis raters) and four written sets of comments (two from ESL experts and Aptis raters respectively). Inter-coder reliability estimates were calculated, using Cohen's kappa statistics (email response: k = 0.82; website article: k = 0.87). Disparities in the coding process were resolved through follow-up discussions. As a result, minor revisions were made of the coding schemes. Finally, one of the researchers coded the rest of the data with the revised coding schemes.

As mentioned previously, the participants' written comments focused on the discourse features of the writing samples that were allocated to them. An examination of their written comments data, however, revealed that compared with the verbal comments that they made in focus group sessions, their written comments appeared more detailed and focused, representing each participant's evaluation of a specific writing sample. This enabled us to perform additional analyses of the written comments data, on top of the discourse features that were extracted from the coding process. Specifically, we calculated the frequency statistics of each discourse feature that was extracted from the written comments data across the three levels. Then, based on the nature of the comments, we coded each comment into two categories of 'positive' and 'negative' (see examples in Appendix 4) and calculated their frequencies. In so doing, we were able to identify the patterns, if any, that emerged across the extracted discourse features and across the three levels in focus.

3.4 Phase 2: Discourse analysis of Aptis writing samples

Following the focus groups, we conducted detailed discourse analyses of Aptis writing samples, aiming at establishing whether the trends of discourse features across writing score levels identified in the focus groups also held for a larger sample of email and website writing performances. This was necessary, as the features that were commented on in Phase 1 might have just been due to features in the writing samples we selected, rather than being more broadly representative.

3.4.1 Writing samples

A total of 120 writing samples in two parallel test forms were included in our analysis with 40 samples at each level. The writing samples were provided by the British Council, the owner of the Aptis Advanced writing test. Details concerning the number of samples at each of the three levels (B2.2, C1, and C2) are delineated in Table 5.

Table 5: The writing samples for discourse analysis study

Level	Task 2: Email response (n = 60)	Task 3: Website article (n = 60)
B2.2	Form A (n = 10)	Form A (n = 10)
	Form B (n = 10)	Form B (n = 10)
C1	Form A (n = 10)	Form A (n = 10)
	Form B (n = 10)	Form B (n = 10)
C2	Form A (n = 10)	Form A (n = 10)
	Form B (n = 10)	Form B (n = 10)

3.4.2 Coding procedures

To explore the discourse features across the two writing tasks and three score levels, coding schemes were developed to code the email and website writing samples (see Appendix 5). The development of the coding schemes was informed by the findings that emerged from Phase 1 of this study. Table 6 summarises the discourse features that were included in our analysis. We established some broader categories (shown in column 2) which included a range of discourse features (column 3). The selection of these features was informed by the findings from Phase 1 of this study. In addition, for those features that entailed manual coding, we also took into consideration the difficulty of reaching consistent coding results. After piloting the coding schemes, only the features which were likely to generate reliable coding results were retained in our analysis.

Table 6: Discourse features that were included in the discourse analysis study

Task	Category	Discourse feature	Coding method
Email	Email conventions	- Inclusion of opening	Manual
		- Inclusion of closing	Manual
		- Appropriateness of opening	Manual
		- Appropriateness of closing	Manual
Website		- Interestingness	Manual
Email & website	Accuracy	- Proportion of Error-free T-units (EFTs)	Manual
Email & website	Fluency	- Number of words	Automated
		- Number of T-units	Manual
		- Number of clauses	Manual
Email & website	Syntactic complexity	- Number of words per T-unit	Manual
		- Number of clauses per T-unit	Manual
		- Number of words per clause	Manual
Email & website	Lexical complexity	- Average word length	Automated
		- Lexical density	Automated
		- Lexical sophistication	Automated
Email & website	Coherence	- Referential cohesion	Automated
		- Argument overlap (local)	Automated
		- Argument overlap (global)	Automated
		- All connectives	Automated
Email	Content	- Number of idea units	Manual
Website		- Use of notes	Manual
		- Use of statistics	Manual
Email & website	Orthographic control	- Spelling	Manual
		- Punctuation	Manual
Website		- Paragraphing	Manual

As indicated in Table 6, some features were manually coded whereas others were coded with automated discourse analysis tools. For example, Coh-Metrix (Graesser, McNamara, Louwerse & Cai, 2004) was utilised to analyse average word length, and VocabProfile (Cobb, 2002) was used to analyse lexical density and lexical sophistication. Two researchers coded the features that entailed manual coding. For each of these features, the two researchers coded 16 samples (13.3%) independently and then compared their coding results. Inter-coder reliability estimates (Spearman's *rho*) for these features ranged from 0.78 to 0.95, suggesting reasonably satisfactory to satisfactory inter-coder consistency (Field, 2009). Disparities in the coding process were resolved through follow-up discussions. Finally, one of the researchers coded the rest of the samples.

3.4.3 Statistical analysis

Statistical analyses were performed of the coding results, aiming to explore whether the selected discourse features could distinguish email and website writing samples at the three score levels in focus. The only independent variable in this analysis was score level awarded by certified Aptis raters. Task was not included in our analysis as an independent variable because: a) we did not intend to investigate test-takers' language use across the two tasks of email and website writing; and b) the rating scales for email and website are different, and hence this analysis was not warranted.

In the Results section, we present the descriptive statistics of each selected discourse feature to explore emerging patterns, if any, in the findings. We then employed inferential statistics to ascertain whether a significant difference existed in each selected discourse feature for the writing samples at the three score levels. In this analysis, the score level was the independent variable and the selected discourse feature in the writing samples the dependent variable (DV). The statistical analysis method was selected based on the property of the DV. If the DV was a continuous variable, one-way analysis of variance (ANOVA) was employed; however, if the DV was not a continuous variable, the non-parametric ANOVA, that is, the Kruskal-Wallis test was performed. When a significant difference was detected, partial eta-squared value was computed as estimate of effect size (η^2), which was interpreted based on Cohen (1988):

- $\eta^2 = 0.02$ (small)
- $\eta^2 = 0.13$ (medium)
- $\eta^2 = 0.26$ (large)

Prior to the ANOVA procedure, statistical assumptions were checked. Two key assumptions to ANOVA are that the data in the samples are each normally distributed and have variances that are equivalent to each other (Field, 2009). If assumptions were violated, the Kruskal-Wallis test was performed. When the ANOVA result was significant, post hoc pairwise comparisons using least significant difference (LSD) were implemented to investigate whether a significant difference existed between any two levels (i.e. B2.2 versus C1, C1 versus C2, and B2.2 versus C2) with regards to a specific discourse feature. In the case of a significant Kruskal-Wallis test, post hoc comparisons were performed using the Dunn-Bonferroni method (Field, 2009).

In addition to ANOVA, we also applied independent-samples t-test (when the DV was continuous) or Mann-Whitney U test (when the DV was not continuous) to compare the selected discourse features across the two test forms (i.e. Forms A and B, see Table 5). In this analysis, when a significant difference was detected, Cohen's *d* was used as the effect size, which was interpreted based on Cohen (1988):

- d < 0.2 (no effect size)
- 0.2 ≤ d < 0.5 (small effect size)
- 0.5 ≤ d < 0.8 (medium effect size)
- d ≥ 0.8 (large effect size)

All analyses were performed in SPSS 21 (IBM, 2012) with the critical value set at 0.05. It should be noted, however, that we applied the Bonferroni correction procedure (Field, 2009) to adjust the critical value when necessary to avoid Type I error because multiple ANOVAs/t-tests were performed of the discourse features in some categories in Table 6. For example, fluency was captured by three measures (i.e. *numbers of words*, *T-units*, and *clauses*). Therefore, the critical value was adjusted to 0.05/3 = 0.017 accordingly.

4. RESULTS

In this section, we report the findings relating to email response and website article respectively. For each task, we present the findings from: a) the focus groups with ESL experts and Aptis raters; b) their written comments on the writing samples; and c) the discourse analyses of the writing samples with the statistical analyses.

4.1 Email response

4.1.1 Focus groups

Overall, the participants commented on the following five discourse features which they considered as distinguishing email writing samples either at two adjacent levels (e.g., B2.2 versus C1, C1 versus C2) or across the three score levels in focus (i.e. B2.2, C1, and C2, see also Appendix 3):

1. vocabulary and grammar
2. sociolinguistic features
3. content
4. cohesion and coherence
5. orthographic control.

In what follows, we present the findings on each of the five categories that were extracted from the focus group data.

1) *Vocabulary and grammar*

The participants commented that compared with C1 and C2, email writing samples at B2.2 featured noticeably more errors, though most of them did not impede comprehension. The comment in Excerpt 1 from an ESL expert on an email writing sample at B2.2 helps to illustrate this point:

Excerpt 1

> *I think the first one (B2.2) probably is more striking for me in terms of grammatical errors, sort of combining more than one idea in a sentence and you can see that someone is at lower level straight away. That's probably sort of a striking aspect for me.* (E2, EG1)

In addition, it was also mentioned that the email writing samples at high levels, in particular C2, featured the use of a wider range of, and more, sophisticated vocabulary and grammar, with improved accuracy. An Aptis rater, for instance, made the following comment on a writing sample at C2:

Excerpt 2

> *I mean they both (C1 and C2) seemed to be quite well-organised. But in the second email (C2), it has much more use of complex grammar, complex sentence forms without errors.* (R6, RG2)

2) *Sociolinguistic features*

The participants commented on three sociolinguistic features that distinguished email writing samples at the three levels: a) the use of email conventions, particularly opening and closing; b) polite style; and c) audience awareness. Email writing samples at high levels seemed to include opening and closing, and demonstrate more appropriate use of these features in their writing. In Excerpt 3 below, an Aptis rater comments on an email writing sample at C1 in comparison with another one at B2.2:

Excerpt 3

> Researcher: *In comparison with the first one (B2.2), what kind of features can you find about this one (C1)?*
>
> R1: *Well, it's better.*
>
> Researcher: *In what ways? Can you elaborate a bit?*
>
> R1: *It's like a letter with opening and closing, not an essay.*

The participants also considered the polite style as another feature that made a difference for the email writing samples at the three levels, arguing that email writing samples at high levels exhibited a better use of the polite style. Speaking of the differences between two samples at C1 and C2 respectively, an ESL expert made the following comment:

Excerpt 4

> *There are three points to go back and "I'm looking forward to your response". Again, very polite. I'm expecting you to respond to this. So, that, to me, seems to be different from the other one (C1).* (E2, EG1)

Some participants mentioned that the email writing samples at high levels also seemed to engage readers better by showing better audience awareness. In Excerpt 5 below, two ESL experts (E3 and E4) discuss an email writing sample at C2, which was supposed to be about a complaint. As observed by E4, the writer provided the context needed for the reader, and by doing so, the email did not look *just like a complaint*.

Excerpt 5

> E3: *I liked the way it begins… I'm quite interested in the function, I suppose, at this level (C2).*
>
> E4: *She gave us contexts.*
>
> E3: *Contexts. That's what it is. So then it's like we do kind of…*
>
> E4: *It's not just like a complaint.*
>
> (E3 and E4, EG2)

3) Content

The participants mentioned two aspects pertaining to the content which tended to distinguish email writing samples across the three levels in focus: a) use of source information; and b) delivery of arguments. With regard to the first aspect, those at high levels seemed to expand or elaborate on the notes that were provided in the task prompts; in comparison, those at a lower level, particularly B2.2, were inclined to copy the notes verbatim in their writing, with limited or no elaboration. An Aptis rater (R6, RG2) observed that the difference between an email writing sample at C1 and B2.2 was that '*they (the writers) have really developed from the notes*. The participants also mentioned that those at high levels delivered their viewpoints more effectively than those at a lower level, whose ideas in some cases appeared somewhat unconvincing. The following remark made by an ESL expert illustrates this point.

Excerpt 6

> *What else did I notice about level 1 (B2.2)? Some unconvincing ideas brought up to support what they were saying, like I got a headache which is dangerous on the road. Fair enough. But it just seems sort of unconvincing.* (E3, EG2)

4) Cohesion and coherence

Cohesion and coherence were yet another category that featured quite prominently in the focus group data. In most cases, the participants’ comments focused on the use of linking words or cohesive devices in the email writing samples. In Excerpt 7, for instance, an ESL expert was commenting on the difference between two writing samples at C1 and C2 in terms of the use of linking devices:

Excerpt 7

> *Very little structuring in the level 2 (C1) one. We had ‘firstly, lastly’, and that was about it. So all of the transitions were quite abrupt, compared with level 3 (C2), where there are a lot of smooth transitions with ‘however’ and ‘besides’.* (E3, EG2)

In a similar vein, an Aptis rater (R4, RG2) remarked that in comparison with a writing sample at B2.2, the one at C1 featured *better use of cohesive devices…ideas are arranged a bit more coherently. You can understand it a bit better.*’

Other participants also mentioned that the email writing samples at higher levels appeared to be better organised, as indicated by the comments made by an ESL expert in Excerpt 8 below:

Excerpt 8

> *That’s good for level 2 (C1). That’s better than level 1 (B2.2). There are reasons why this not a level 1 as well, because it’s got those…introducing the topics, and it’s very clearly structured. It’s not messy. Bit out of order but at least I know this is what I’m reading about.* (E4, EG2)

5) Orthographic control

Some participants mentioned that orthographic control features, or specifically, spelling, punctuation, and paragraphing, distinguished writing samples at different levels, especially between C1 and B2.2. An Aptis rater (R4, RG2), for instance, commented on an email writing sample at B2.2, in comparison with C1, that *the language is more correct; the punctuation is much better.*

4.1.2 Written comments

Overall, five features were identified as the most prominent in the written comments made by the participants on the email writing samples across the three levels (B2.2, C1 and C2, see also Appendix 5):

1. vocabulary and grammar
2. sociolinguistic features
3. content
4. cohesion and coherence
5. orthographic control.

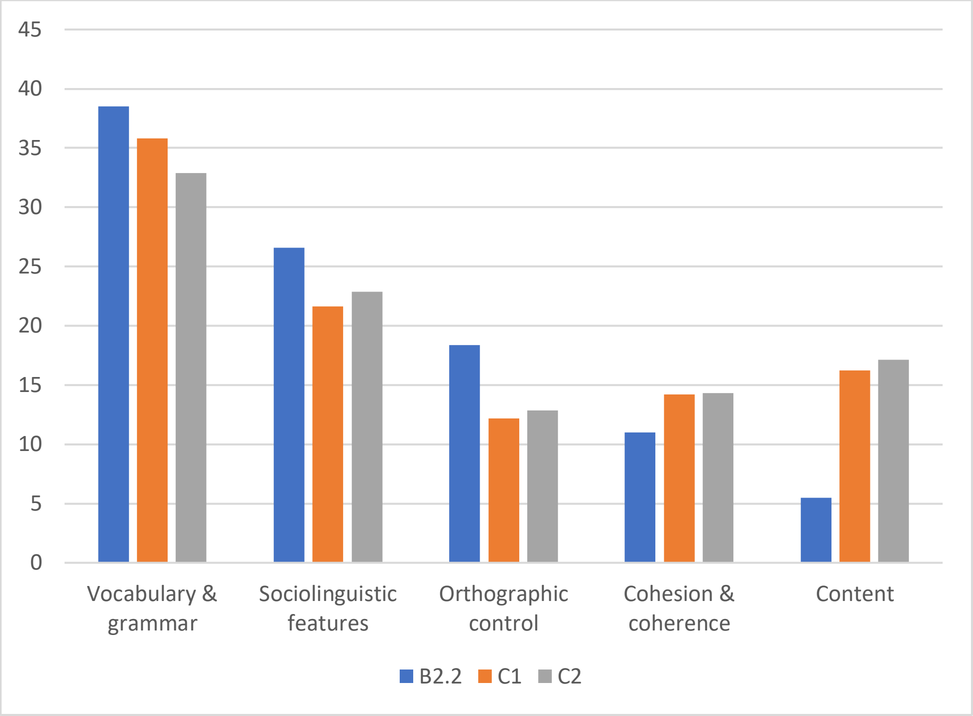
In Table 7, we present the frequency statistics of the five features that were extracted from the written comments. Based on their nature, we coded the comments into two categories: 'positive' and 'negative' (see Appendix 4 for examples in the coding scheme). To make the frequency statistics comparable across the three levels, we also calculated the percentage statistics of each feature as well as the positive and negative categories across the three levels (see Table 7).

Table 7: Frequency statistics of the features of email writing (written comments)

Feature	B2.2		C1		C2	
	k	%	*k*	%	*k*	%
Vocabulary and grammar	42	38.53	53	35.81	46	32.86
• Positive	14	12.84	26	17.57	33	23.57
• Negative	28	25.69	27	18.24	13	9.29
Sociolinguistic features	29	26.61	32	21.62	32	22.86
• Positive	12	11.01	24	16.22	24	17.14
• Negative	17	15.60	8	5.41	8	5.71
Content	6	5.50	24	16.22	24	17.14
▪ Positive	1	0.92	17	11.49	23	16.43
▪ Negative	5	4.59	7	4.73	1	0.71
Cohesion and coherence	12	11.01	21	14.19	20	14.29
• Positive	4	3.67	12	8.11	18	12.86
• Negative	8	7.34	9	6.08	2	1.43
Orthographic control	20	18.35	18	12.16	18	12.86
• Positive	6	5.50	8	5.41	10	7.14
• Negative	14	12.84	10	6.76	8	5.71
Total	**109**	**100.00**	**148**	**100.00**	**140**	**100.00**

Based on the statistics in Table 7, we plotted the percentages of the written comments on the five features across the three levels in focus. As displayed in Figure 3, considerably more comments seem to focus on two features, namely, a) vocabulary and grammar and b) sociolinguistic features, with small variations across the three levels. Another pattern which appears quite explicit is that with the increase of score levels, the participants commented less frequently on the features of 'vocabulary and grammar' and 'sociolinguistic features' but more frequently on 'cohesion and coherence' and 'content'. This trend is particularly noticeable for the feature of 'content', with more frequent comments at C1 and C2 than B2.2.

Figure 3: Comments on the five aspects of email writing across the three levels



We also plotted the percentage statistics of positive and negative comments across the three levels (see Figures 4 and 5). As illustrated in Figure 4, not surprisingly, email writing samples at high levels (e.g., C2) have a higher percentage of positive comments than those at a lower level (e.g., B2.2). This pattern is reversed when it comes to negative comments (see Figure 5).

Figure 4: Positive written comments on email writing samples across the three levels

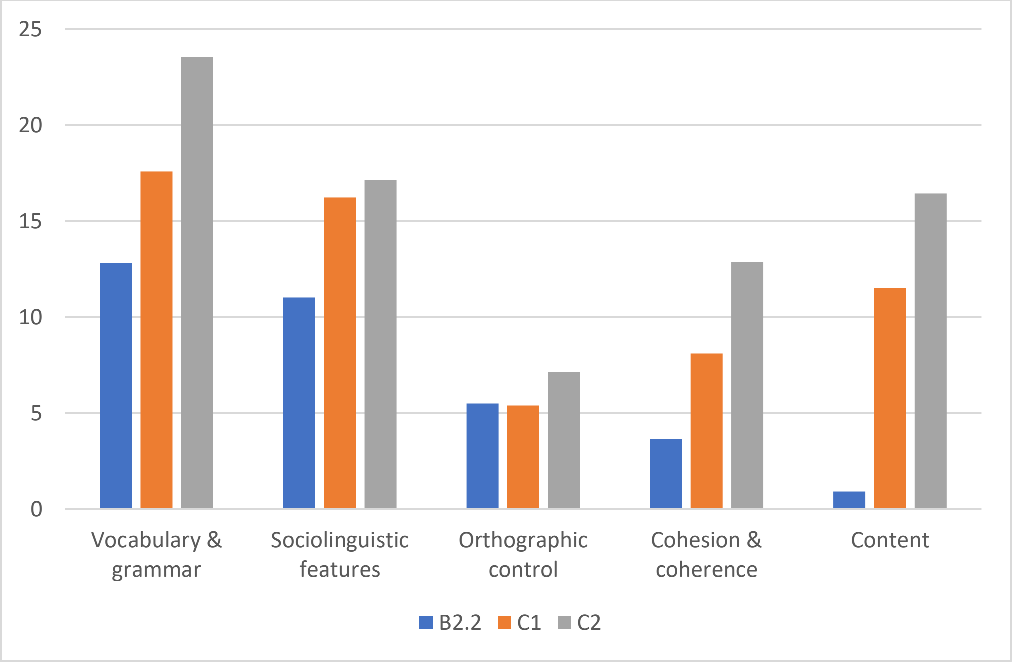
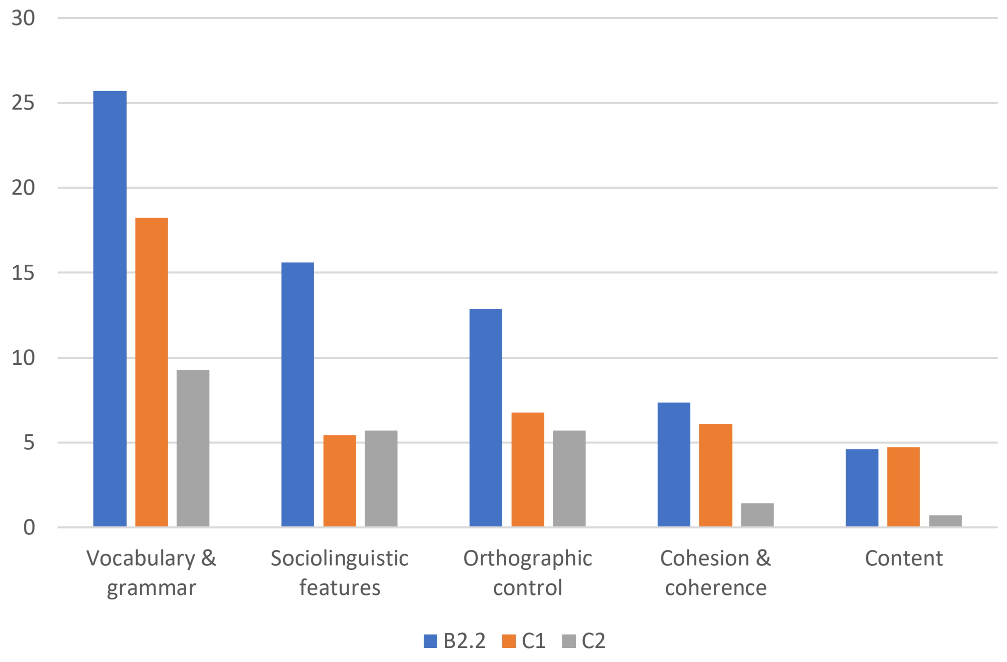


Figure 5: Negative written comments on email writing samples across the three levels



In the following section, we briefly present the qualitative findings from the written comments relating to each feature.

1) Vocabulary and grammar

Positive comments regarding vocabulary and grammar usually focused on their accuracy and range, and at higher levels (i.e. C1 and C2), sophistication; negative comments, on the other hand, primarily concerned the errors in the use of vocabulary and grammar, particularly for the writing samples at B2.2, though it was also mentioned that these errors did not usually impede comprehension or expression of meaning. Below are two examples of comments on vocabulary and grammar:

- *Grammatical structures/vocabulary higher B1, with errors with complex structures and poor choice of lexis but meaning generally clear* (B2.2, R6, RG2)
- *There are complex structures; some are erroneous* (C1, R4, RG2)
- *A wide range of vocabulary accurately used* (C2, E2, EG1)

2) Sociolinguistic features

Written comments on sociolinguistic features mostly concerned: a) email conventions, particularly whether opening and closing were included; and b) whether a polite tone was used. Email writing samples at high levels were found to exhibit improved use of sociolinguistic features as compared with those at a lower level. Some examples of comments are provided below:

- *No proper opening and closure* (B2.2, E3, EG2)
- *Incorrect register; essay format; quite rude* (B2.2, R5, RG2)
- *Appropriate tone that interacts with the recipient* (C1, R3, RG1)
- *Good greeting and salutation with appropriate register* (C2, R1, RG1)

3) Content

Email writing samples at higher levels seemed to exhibit the following merits as compared with those at a lower level: a) including all the important points; b) expanding the notes in the task prompts; c) effective introduction and conclusion; and d) the writing was on task which addressed the communicative goals effectively. Email writing samples at a lower level, however, tended to be characterised by: a) direct or verbatim use of the notes in the task prompts; and b) arguments not well supported with relevant ideas. Below are a few illustrative comments on content:

- *Three points – follow prompts too closely (word-for-word) and explanations are vague – no specific examples given* (B2.2, E4, EG2)
- *The ideas are organised relatively well, and each question is covered* (C1, E3, EG2)
- *Arguments reinforced with supporting ideas and specific example* (C2, E2, EG1)
- *On task. Points are developed in an original manner* (C2, R5, RG2)

4) Cohesion and coherence

When it comes to cohesion and coherence, the use of linking words or cohesive devices constituted the focus of the written comments. The email writing samples at high levels were found to be more coherent, characterised by better use of linking words or cohesive devices. Below are some illustrative comments:

- *Poor use of cohesion (linking devices)* (B2.2, E3, EG2)
- *Cohesive devices more sophisticated than sample 6 – e.g. Firstly, When I finally…but not always successful e.g. Well, in my opinion to introduce point 3* (C1, E4, EG2)
- *Effective and accurate use of discourse markers: first of all*, *in addition*, *moreover* (C2, E2, EG1)
- *Good use of cohesive devices and linking words* (C2, R1, RG1)

5) Orthographic control

Comments on orthographic control mostly focus on three aspects: a) spelling, b) punctuation, and c) paragraphing. Overall, email writing samples at high levels demonstrated better orthographical control. A few illustrative comments are provided below:

- *Spelling and punctuation show some errors* (B2.2, R2, RG1)
- *Excellent use of punctuation* (C1, R4, RG2)
- *Helpful paragraphing makes the response easy for the reader to process and refer back to* (C2, R3, RG1)

4.1.3 Discourse analyses of email writing samples

As Table 6 (see Section 3.3.2) indicates, the following discourse features were included in our analysis of email writing samples (see also Appendix 4):

- email conventions
- fluency
- syntactic complexity
- lexical complexity
- cohesion and coherence
- content
- orthographic control.

As mentioned previously, the selection of these features was determined primarily based on the findings from Phase 1 of this study. Findings concerning each discourse feature are depicted below.

1) Email conventions

This discourse feature was manually coded according to two criteria:

- whether opening and closing were included
- whether opening and closing were appropriately used.

Table 8 presents the frequency statistics of this feature across the three score levels.

Table 8: Frequency statistics of email conventions across the three levels (email response, n = 60)

Feature	Level	Inclusion		Appropriateness	
		Yes	No	Yes	No
Opening	B2.2	18	2	16	4
	C1	19	1	18	2
	C2	20	0	18	2
	Combined	57	3	52	8
Closing	B2.2	18	2	16	4
	C1	18	2	16	4
	C2	20	0	19	1
	Combined	56	4	51	9

As indicated in Table 8, most test-takers included opening (n = 57) and closing (n = 56) in their email writing, though slight differences could be observed across the three levels. A similar pattern could be identified for the appropriateness of opening and closing. Overall, most of the openings (n = 52) and closings (n = 51) were considered as appropriate. Results of the Kruskal-Wallis test did not reveal any significant differences among the three score levels regarding this discourse feature (inclusion of opening: x^2 = 2.07, df = 2, p = 0.365; inclusion of closing: x^2 = 2.11, df = 2, p = 0.349; appropriateness of opening: x^2 = 3.78, df = 2, p = 0.151; appropriateness of closing: x^2 = 2.31, df = 2, p = 0.314). In addition, we also performed Chi-square tests to compare the feature of 'email conventions' on the two test forms (i.e. Forms A and B). The results were all non-significant (inclusion of opening: x^2 = 3.16, df = 1, p = 0.076; inclusion of closing: x^2 = 0.00, df = 1, p = 1.000; appropriateness of opening: x^2 = 0.00, df = 1, p = 1.000; appropriateness of closing: x^2 = 0.13, df = 1, p = 0.718), suggesting that email samples in the two forms exhibited similar uses of email conventions.

2) *Fluency*

Three measures were manually coded to capture fluency of email writing:

- number of words
- number of T-units
- number of clauses.

Table 9 presents the descriptive statistics of the three fluency measures across the three score levels in focus. As shown in this table, the number of words increases with the increase of score levels. It is worth noting, however, that the standard deviation of this measure at C2 is considerably larger than that at C1, which is in turn larger than B2.2, suggesting the greater variability in the number of words in the email writing samples written by those at higher score levels. When it comes to the number of T-units and clauses, a reverse pattern was identified. As shown in Table 9, as the level ascends, the number of both T-units and clauses decreases, albeit by only a small margin.

Table 9: Descriptive statistics of fluency measures across the three score levels (email response, n = 60)

Proficiency level	Number of words		Number of T-units		Number of clauses	
	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 20)	148.20	15.62	10.55	2.26	15.45	3.18
C1 (n = 20)	153.95	22.71	10.45	2.86	14.95	3.65
C2 (n = 20)	163.50	33.67	10.20	2.69	14.75	3.85
Combined (n = 60)	155.22	25.50	10.40	2.57	15.05	3.52

ANOVA results indicated a non-significant effect of proficiency level on the three fluency measures (number of words: F = 1.89, df = 2, p = 0.160; number of T-units: F = 0.09, df = 2, p = 0.909; number of clauses: F = 0.20, df = 2, p = 0.816). Independent-samples t-tests did not reveal a significant difference between the two test forms in terms of the three fluency features (number of words: t = -1.04, df = 58, p = 0.303; number of T-units: t = 0.30, df = 58, p = 0.766; number of clauses: t = -0.04, df = 58, p = 0.971), suggesting that writing samples in the two forms were similar in terms of fluency.

3) Accuracy

The proportion of error-free T-units (EFTs) was used as an indicator of accuracy. To obtain the proportion of EFTs, the following three procedures were followed:

- coding the number of T-units
- coding the number of error-free T-units
- dividing the number of T-units by the number of error-free T-units.

We did not code the types and gravity of errors that writers committed in their email writing because it was difficult to achieve high inter-coder reliability, as suggested by previous discourse analysis studies in L2 writing (e.g., Knoch, et al., 2014).

Table 10 presents the descriptive statistics of the proportion of EFTs across the three score levels. As shown in Table 10, the proportion of EFTs increases by the increase of score levels, suggesting that as writers' proficiency increases, their accuracy of email writing also improves.

Table 10: Descriptive statistics of the proportion of EFTs across the three score levels (email response, n = 60)

Level	Proportion of EFTs	
	Mean	SD
B2.2 (n = 20)	0.63	0.18
C1 (n = 20)	0.66	0.22
C2 (n = 20)	0.78	0.19
Combined (n = 60)	0.69	0.20

ANOVA results indicated that the difference was nearing significant, with a small effect size (F = 3.10, df = 2, p = 0.053, η^2 = 0.098). An independent-samples t-test was performed to compare the two test forms, indicating that Form B featured significantly better accuracy than Form A, with a medium effect size (t = 2.32, df = 58, p = 0.024, d = 0.598).

4) Syntactic complexity

Three measures were used to capture syntactic complexity:

- Number of words per T-unit
- Number of clauses per T-unit
- Number of words per clause.

As indicated in Table 11, the number of words per T-unit increases with the increase of the proficiency level. A similar pattern could be identified for number of words per clause. In the case of number of clauses per T-unit, the mean values are almost the same at the three score levels.

Table 11: Descriptive statistics of syntactic complexity measures across the three score levels (email response, n = 60)

Proficiency level	Number of words per T-unit		Number of clauses per T-unit		Number of words per clause	
	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 20)	14.48	2.71	1.49	0.30	9.86	1.64
C1 (n = 20)	15.69	4.53	1.46	0.24	10.78	2.70
C2 (n = 20)	16.40	2.38	1.47	0.24	11.36	1.74
Combined (n = 60)	15.53	3.38	1.47	0.26	10.67	2.14

ANOVA results indicated non-significant differences in the three syntactic complexity measures across the three levels (number of words per T-unit: F = 1.68, df = 2, *p* = 0.196; number of clauses per T-unit: F = 0.10, df = 2, *p* = 0.910; number of words per clause: F = 2.63, df = 2, *p* = 0.081). Results of independent-samples t-tests showed that the differences between the two forms were non-significant in light of the three syntactic complexity measures (number of words per T-unit: t = -1.11, df = 58, *p* = 0.271; number of clauses per T-unit: t = -0.645, df = 58, *p* = 0.522; number of words per clause: t = -0.516, df = 58, *p* = 0.608).

5) Lexical complexity

Three measures were used to represent lexical complexity of the email writing samples:

- average word length (AWL)
- lexical density
- lexical sophistication.

The first measure, AWL, was generated by Coh-Metrix (Graesser, et al., 2004), an automatic tool for text complexity and coherence analysis; the other two measures were both analysed in VocabProfile (Cobb, 2002), a computer program which analyses lexical complexity based on a corpus-based frequency list. Lexical density was defined as the number of content words (nouns, verbs, adjectives, and adverbs) divided by the total number of words. Lexical sophistication was not directly calculated by VocabProfile. Rather, a number of different aspects of the VocabProfile output were used to calculate this lexical complexity measure. In this study, we obtained the lexical sophistication values through dividing the sum of the AWL tokens and the off-list word tokens by the total number of content words in the text. As indicated in Table 12, as the proficiency level increases, so does AWL.

Table 12: Descriptive statistics of lexical complexity measures across the three score levels (email response, n = 60)

Proficiency level	Average word length		Lexical density		Lexical sophistication	
	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 18)	4.25	0.20	0.46	0.04	0.11	0.05
C1 (n = 16)	4.37	0.21	0.47	0.03	0.14	0.06
C2 (n = 18)	4.53	0.28	0.49	0.04	0.16	0.06
Combined (n = 52)	4.39	0.26	0.48	0.04	0.14	0.06

ANOVA results identified a significant difference in AWL among the three score levels, with a medium effect size (F = 7.12, df = 2, *p* = 0.002 < 0.017, η^2 = 0.20). Post hoc results revealed that while the difference in the AWL mean values between B2.2 and C1 was non-significant, but that between C1 and C2 (*p* = 0.034) as well as B2.2 and C2 (p < 0.001) were both significant.

Lexical density and lexical sophistication, as shown in Table 12, both increased by a small margin with the increase of proficiency level. ANOVA results, however, indicated that neither differed significantly as a function of the proficiency level, though the result for lexical sophistication was nearing significant (lexical density: F = 2.34, df = 2, *p* = 0.105; lexical sophistication: F = 4.27, df = 2, *p* = 0.019)[1]. Results of independent-samples t-tests indicated that neither AWL (t = -0.49, df = 58, *p* = 0.628) nor lexical density (t = -1.00, df = 58, *p* = 0.320) differed significantly between the two test forms; however, when it comes to 'lexical sophistication', the writing samples in Form B were significantly more lexically dense than Form A, with a large effect size (t = -2.56, df = 58, *p* = 0.013 < 0.017, *d* = 0.724).

6) Cohesion and coherence

The following four measures were used to capture cohesion and coherence of the email writing samples:

- referential cohesion (the overlap of content words and ideas across sentences and the entire text)
- argument overlap (local – overlap between sentences in terms of nouns and pronouns)
- argument overlap (global – overlap of nouns and pronouns across the entire text)
- all connectives (the incidence of all cohesive links between ideas and clauses).

These four measures were all generated by Coh-Metrix (Graesser, et al., 2004).

Table 13 shows that with the increase of proficiency level, 'referential cohesion' and 'argument overlap (global)' improve. When it comes to 'argument overlap (local)', it increases as the proficiency level increases from B2.2 to C1 but levels off from C1 to C2; for 'all connectives', it increases from B2.2 to C1 but declines from C1 to C2.

[1] Given that there were three lexical complexity measures, the critical value was adjusted to 0.05/3 = 0.017 accordingly.

ANOVA results revealed that the four cohesion and coherence indices did not differ significantly as a function of the proficiency level (*referential cohesion*: F = 0.99, df = 2, *p* = 0.380; *argument overlap global*: F = 1.59, df = 2, *p* = 0.346; *argument overlap local*: F = 1.08, df = 2, *p* = 0.213; *all connectives*: F = 1.34, df = 2, *p* = 0.270). These findings suggest that the selected four cohesion and coherence measures could not distinguish email writing samples at the three levels in focus. Results of independent-samples t-tests, however, revealed that email writing samples in Form A exhibited a significantly better 'argument overlap global' than Form B, with a large effect size (t = 4.59, df = 58, p < 0.001, d = 1.20)[2]; the other three features, however, were not significantly different (*referential cohesion*: t = 1.85, df = 58, *p* = 0.069; *argument overlap local*: t = 2.45, df = 58, *p* = 0.017; *all connectives*: t = -0.63, df = 58, *p* = 0.530).

Table 13: Descriptive statistics of coherence measures across the three score levels (email response, n = 60)

Proficiency level	Referential cohesion		Argument overlap (local)		Argument overlap (global)		All connectives	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 20)	29.79	22.08	0.51	0.20	0.42	0.21	79.03	14.78
C1 (n = 20)	36.39	26.44	0.61	0.19	0.49	0.21	87.84	20.77
C2 (n = 20)	40.44	23.93	0.60	0.19	0.51	0.19	85.14	16.24
Combined (n = 60)	35.54	24.21	0.57	0.19	0.47	0.20	84.00	17.55

7) Content

The coding of content was based on the research notes that were provided to test-takers. For the email writing tasks in both Form 1 and 2, test-takers were required to include three research notes in their email writing. In the coding process, we first examined whether each note was included. Next, we examined whether each note was either simply included or elaborated on. A three-point scale was used to code each research note:

- 0 – if a note was not included
- 1 – was included but not elaborated on
- 2 – was both included and elaborated on.

By using this coding method, the maximum number of points that could be achieved for each writing sample was six (that is, all three research notes in the task prompt were included and elaborated on, see also Appendix 4).

Table 14 presents the descriptive statistics of the use of research notes in email writing samples, indicating that its mean values increase with the increase of proficiency level. ANOVA results, however, did not identify a significant difference among the three levels (F = 0.43, df = 2, *p* = 0.654).

[2] Since there are four measures of cohesion and coherence, the critical value was adjusted to 0.05/4 = 0.0125.

Results of independent-samples t-test revealed a non-significant difference between the two forms with regard to this feature (t = 1.45, df = 58, *p* = 0.151).

Table 14: Descriptive statistics of use of research notes across the three score levels (email response, n = 60)

Level	Content	
	Mean	SD
B2.2 (n = 20)	4.95	0.89
C1 (n = 20)	5.15	0.99
C2 (n = 20)	5.20	0.83
Combined (n = 60)	5.10	0.90

8) Orthographic control

Two measures were used to capture the orthographic control of email writing:

- spelling errors
- punctuation errors.

Paragraphing was not included in our coding as we noticed that email writing samples were mostly quite short and that it was difficult to reach consistent coding results. To make the two measures comparable, both were standardised through dividing the number of spelling and punctuation errors by the number of words in a writing sample.

As shown in Table 15, writing samples at B2.2 appeared to exhibit more spelling and punctuation errors than those at C1 and C2. To investigate whether spelling and punctuation errors differed as a function of proficiency level, the Kruskal-Wallis test was adopted because: a) the data of ‘spelling errors’ violated the normal distribution assumption; and b) the error variances were not equal across the samples for the ‘punctuation errors’ data. The results indicated that the numbers of both spelling and punctuation errors differed significantly across the three levels (spelling: x^2 = 6.74, df = 2, *p* = 0.034; punctuation: x^2 = 11.65, df = 2, *p* = 0.003)[3]. Post hoc tests with the Dunn-Bonferroni method revealed that for both variables, significant differences existed between the B2.2 and C2 levels (spelling: *p* = 0.028; punctuation: *p* = 0.002), but not between other levels. A Mann-Whitney U test revealed an insignificant result with spelling errors (U = 478, n = 60, *p* = 0.660); independent-sample t-test results indicated that punctuation errors did not differ significantly between the writing samples in the two forms, either (t = 0.68, df = 58, *p* = 0.502).

[3] The critical value was adjusted to 0.05/2 = 0.025 in this analysis.

Table 15: Descriptive statistics of punctuation errors across the three levels (email response, n = 60)

Proficiency level	Spelling errors		Punctuation errors	
	Mean	SD	Mean	SD
B2.2 (n = 20)	0.014	0.018	0.044	0.026
C1 (n = 20)	0.005	0.006	0.029	0.017
C2 (n = 20)	0.004	0.009	0.018	0.011
Combined (n = 60)	0.008	0.013	0.031	0.022

4.1.4 Summary of findings – email response

Five discourse features were identified in the verbal and written comments made by ESL experts and Aptis raters, including:

1. vocabulary and grammar
2. sociolinguistic features
3. content
4. cohesion and coherence
5. orthographic control.

Findings concerning each feature are summarised below.

1) Vocabulary and grammar

Compared with the writing samples at B2.2, those at C1 and C2 seemed to feature the use of more sophisticated vocabulary and grammar with improved accuracy; writing samples at B2.2 tended to demonstrate more errors though they usually did not impede comprehension. This finding, however, was not supported by the discourse analysis results, although the measure for accuracy neared significance.

2) Sociolinguistic features

The participants mentioned email conventions, particularly the inclusion of appropriate openings and closings, as a feature of distinguishing performances at the levels in focus; however, discourse analysis results did not support this observation. The level of politeness did not seem to differentiate the writing samples quantitatively, either, though writing samples at higher levels (that is, C1 and C2) were found to better engage readers.

3) Content

The participants believed that writers at higher levels expanded or elaborated on the notes in the task prompts, as opposed to using them verbatim in their writing, which tended to be a feature with the writing samples at B2.2. This observation, however, was not supported by the discourse analysis of email writing samples. The arguments of writers scored at higher levels were found to be more convincing.

4) Cohesion and coherence

Though the participants commented that email writers at higher levels demonstrated more sophisticated use of cohesive or linking devices and were more skilled in organising their email writing, discourse analysis results did not reveal any significant differences in terms of four selected cohesion and coherence measures.

5) Orthographic control

The participants observed that writers at higher levels exhibited better proficiency in spelling, punctuation, and paragraphing. Discourse analysis of the email writing samples partly lent support to this observation, indicating that writing samples at C2 featured significantly less spelling and punctuation errors than those at B2.2.

On top of the findings listed above, the participants also commented that it was difficult to tell the differences between adjacent levels or across the three levels. This observation was largely consistent with the discourse analysis results which indicated that writing samples at the three levels in focus in most cases did not demonstrate significantly different discourse features. Another finding which is worth mentioning concerns the equivalence between the two forms. Three features, including accuracy, lexical complexity, and argument overlap (global), were statistically different for the writing samples in the two test forms, raising concerns over the equivalence of the email response task in parallel test forms. These findings will be further discussed in Section 5.

4.2 Website article

4.2.1 Focus groups

Overall, the two groups of participants commented on the following features which they considered as distinguishing website writing samples either at two adjacent levels (e.g., B2.2 versus C1, C1 versus C2) or across the three score levels in focus (i.e. B2.2, C1, and C2, see also Appendix 3):

- vocabulary and grammar
- sociolinguistic features
- content
- cohesion and coherence
- orthographic control.

Note that these five broad categories which emerged from the coding process were the same as those extracted about email writing (see Section 4.1); however, as will be detailed in this section, the themes under each category were somewhat different. We present the findings relating to each of these five categories in the following section.

1) *Vocabulary and grammar*

The participants commented that compared with the website writing samples at a lower level, those at high levels seemed to feature more sophisticated and a wider range of vocabulary and grammar, with improved accuracy; writing samples at the lower levels, in particular B2.2, on the contrary, demonstrated more errors. This is evidenced by the following excerpt where an Aptis rater was commenting on vocabulary and grammar of a website writing sample at C1, in comparison with B2.2.

Excerpt 9

> R4: *The use of mixed simple and complex structures. Good range of vocabulary here. So much better in terms of progression of level.*
>
> Researcher*: You mean compared with the previous one (B2.2)?*
>
> R4*: Yes, that's correct.* (R1, RG1)

2) *Sociolinguistic features*

The participants mentioned that the writers at high levels appeared to be more proficient in engaging readers through several methods such as the use of questions, and the second person pronoun 'you'. In the following excerpt, an ESL expert was commenting on a website writing sample at C2, in comparison with C1:

Excerpt 10

> *But that sort of…lots of good use of rhetorical moves to keep the reader with…you know. "Did you know that London has the most CCT camera as per capita?" Yeah, probably a bit unsophisticated to use a question at the start. But it says the reader there and you there. And it raises the question and kind of make sure you want to read more.* (E1, EG1)

In a similar vein, an Aptis rater remarked that *like use of questions, not necessarily rhetorical questions but questions to start get people thinking* (R6, RG2).

3) *Content*

Three aspects of content were identified through coding the focus group data:

- use of source information
- delivery of arguments
- use of cultural or literacy references.

In the website writing task, test-takers are provided with several research notes and some statistics that they can use in their writing. The participants reported that website writing samples at different levels tended to demonstrate somewhat different use of the source information. They argued that those at higher levels seemed to make better use of the source information in the sense that they elaborated on the notes and used them to build an argument. An ESL expert's comment in the following excerpt helps to illustrate this point.

Excerpt 11

Yeah they clearly stated their own position in relation to the issue as well. And even the way they fleshed out each point using the steps given in the input. So compared with those at level 2 (C1) who kind of listed the steps for the three cities as examples of the same point, those at level 3 (C2) used that information to make comparisons…and built an argument. (E1, EG1)

In addition to the use of source information, the participants also believed that those at higher levels were more effective in delivering their arguments. Compared with those at a lower level who tended to provide a list of the points that they intended to make, writers at high levels were more successful in integrating the data that was provided to them. This is evidenced by the conversations between two ESL experts in Excerpt 12:

Excerpt 12

E4: *So they've linked that all. And they've integrated the data very well with their overall argument.*

Researcher: *Argument. OK.*

E4: *So it's not just thrown in there as details.*

E3: *Digested. Haven't they?*

E4: *Yes. So this… to use the academic English terminology, they've synthesised…*

E3: *Yup!*

E4: *And paraphrase very well.* (E3 and E4, EG2)

The participants also mentioned the use of cultural or literary references in website writing samples as a feature of the email writing samples at high levels. It is worth noting, however, that this feature tends to be related to the nature of a website writing task. For example, the topic of one of the website writing tasks in focus (Form A) is the use of cameras and surveillance tools in society. Some test-takers used literacy references such as George Orwell and the characters in his book *1984* (e.g., Big Brother). The participants commented favourably on the literacy references which, they believed, also helped to make the writing more interesting and engaging for readers. In Excerpt 13 below, an ESL expert comments on a website writing sample at C2. She reasons that the title with the literary reference of 'Big Brother', immediately captured her attention.

Excerpt 13

I thought it captured the reader's attention. I knew exactly I would wanna read this. Big Brother is watching it. It's kind of captured my attention right from the title. That sort of…lots of good use of rhetorical moves to keep the reader…you know. (E2, EG1)

4) *Cohesion and coherence*

The participants commented on this category primarily from the following two perspectives:

- Use of cohesive or linking devices
- Organisation.

Compared with website writing samples at C1 and C2, the participants observed that those at B2.2 failed to some extent to demonstrate the skilful use of cohesive or linking devices in their writing, hence negatively affecting their writing quality. In Excerpt 14 below, an Aptis rater was commenting on a website writing sample at B2.2, holding the view that limited use of cohesive or linking devices constituted a problem with this sample.

Excerpt 14

> *You saw the information…It's not well-written in the sense that…its limited use of discourse markers, and a very small number of discourse markers but also very simple ones.* (R3, RG1)

Organisation of ideas was another aspect that the participants commented on. The website writing samples at high levels (C1 and C2) appeared to be better organised or structured, with a good beginning and conclusion than those at a lower level. Speaking of a website writing sample at C2, an Aptis rater (R5, RG2) commented that it '*organised ideas and information logically, and I can see there's clear progression*'.

5) *Orthographic control*

The participants commented on the orthographic control of the website writing samples mainly from the following three aspects:

- spelling
- punctuation
- paragraphing.

The participants indicated that the writing samples at B2.2 demonstrated more issues with orthographic control as compared those at higher levels (i.e. C1 and C2). For instance, an Aptis rater made the following comment on the orthographic control of a website writing sample at B2.2.

Excerpt 15

> *First, a lot of issues in terms of paragraphing. There're quite a few paragraphs with only one sentence. A lot of punctuation issues as well. Many sentences start with lower case rather than capital letter. And again there're quite a few issues with spelling; even some of the words which were obviously in the question, they haven't correctly spelt them, even though they could just copy them.* (R4, RG2)

It was also indicated, however, that even the website writing samples at C2 had issues, albeit less noticeable, with orthographic control. An Aptis rater (R6, RG2) commented on a website writing sample at C2 that *I don't think there are many spelling and punctuation issues. The only issue maybe here is still with the, you know, one big paragraph.*

4.2.2 Written comments

Overall, five discourse features were identified as the most prominent in the written comments made by the participants on the website writing samples across the three levels (i.e. B2.2, C1 and C2, see also Appendix 4):

- vocabulary and grammar
- sociolinguistic features
- content
- cohesion and coherence
- orthographic control.

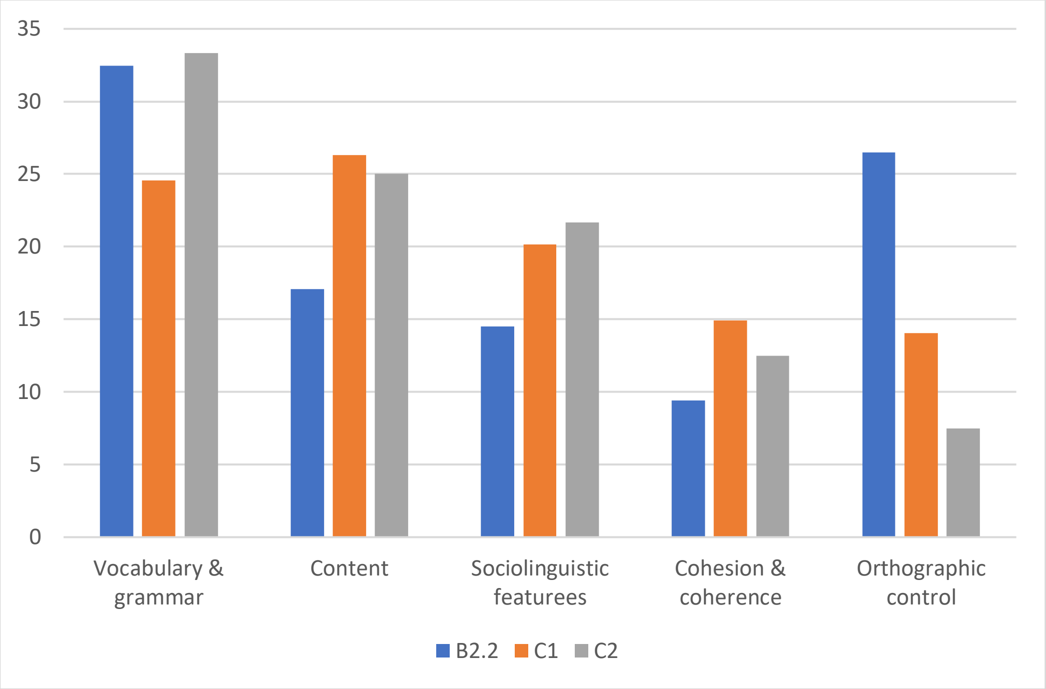
In Table 16, we present the frequency statistics of the five features that were extracted from the written comments. Based on their nature, we coded the comments into two categories: 'positive' and 'negative' (see also Appendix 5 for examples in the coding scheme). To make the frequency statistics comparable across the three levels, we also calculated the percentage statistics of each feature as well as of the positive and negative categories across the three levels (see Table 16).

Table 16: Frequency statistics of the features of website writing (written comments)

Feature	B2.2		C1		C2	
	k	**%**	***k***	**%**	***k***	**%**
Vocabulary and grammar	38	32.48	28	24.56	40	33.33
• Positive	8	6.84	15	13.16	31	25.83
• Negative	30	25.64	13	11.40	9	7.50
Sociolinguistic features	17	14.53	23	20.18	26	21.67
• Positive	0	0.00	20	17.54	25	20.83
• Negative	17	14.53	3	2.63	1	0.83
Content	20	17.09	30	26.32	30	25.00
• Positive	1	0.85	22	19.30	25	20.83
• Negative	19	16.24	8	7.02	5	4.17
Cohesion and coherence	11	9.40	17	14.91	15	12.50
• Positive	2	1.71	14	12.28	10	8.33
• Negative	9	7.69	3	2.63	5	4.17
Orthographic control	31	26.50	16	14.04	9	7.50
▪ Positive	6	5.13	2	1.75	5	4.17
▪ Negative	25	21.37	14	12.28	4	3.33
Total	**117**	**100.00**	**114**	**100.00**	**120**	**100.00**

Based on the statistics in Table 16, we plotted the percentages of the written comments on the five features across the three levels in focus. As displayed in Figure 6, the participants seemed to comment most frequently on three features of website writing across the three levels: vocabulary and grammar, content, and sociolinguistic features, though it should be noted that orthographic control received considerably more comments for website writing samples at B2.2.

Figure 6: Comments on the five aspects of website writing across the three levels



We also plotted the percentage statistics of positive and negative comments across the three levels (see Figures 7 and 8). As displayed in Figure 7, a quite explicit pattern is that website writing samples at high levels received more positive comments than those at a lower level. This trend is particularly noticeable when we compare the writing samples at B2.2 and C2; the differences between C1 and C2, however, are not very marked. With respect to the negative comments, a reverse pattern can be identified, as displayed in Figure 8.

Figure 7: Positive written comments on website writing samples across the three levels

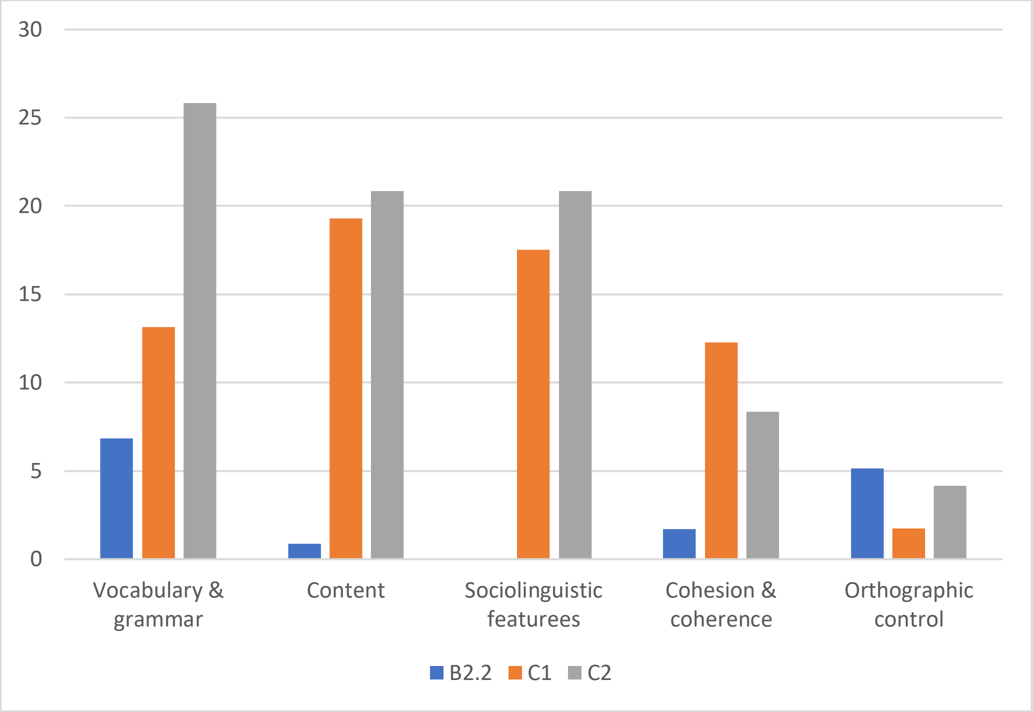
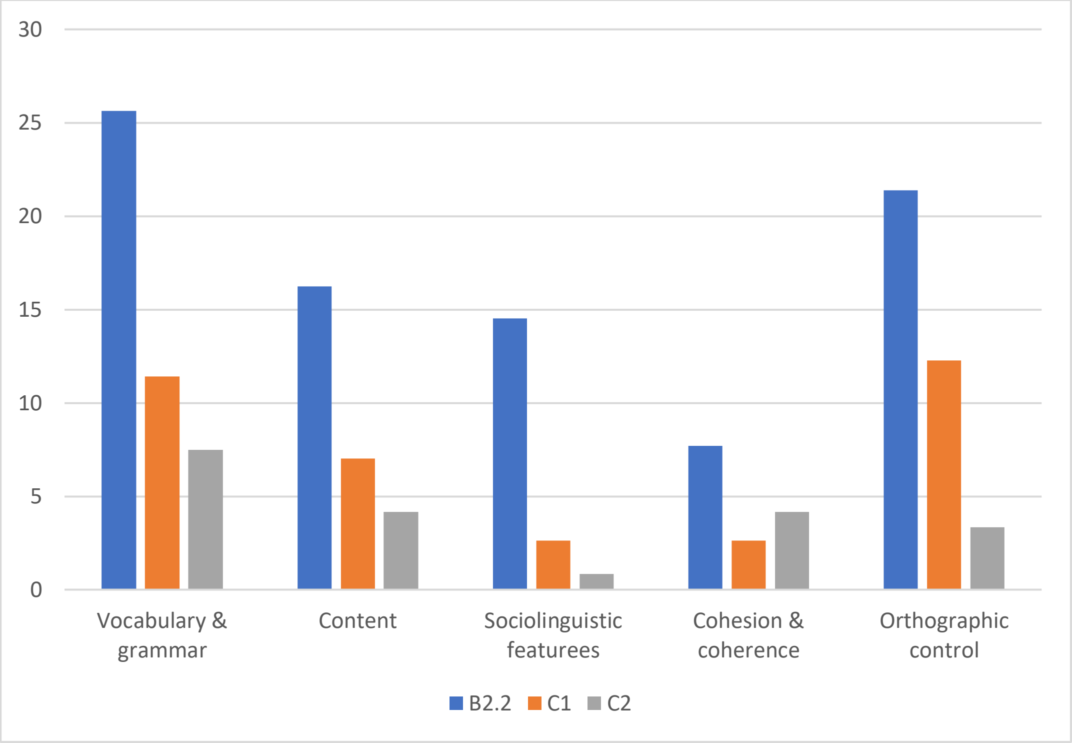


Figure 8: Negative written comments on website writing samples across the three levels



In the following section, we briefly summarise the findings relating to each feature.

1) Vocabulary and grammar

Website writing samples at high levels (i.e. C1 and C2) tended to feature the use of more sophisticated vocabulary and grammar, whereas those at B2.2 exhibited more lexical and grammatical errors. Some illustrative comments are provided below:

- *Tense and agreement errors throughout, very salient* (B2.2, E1, EG1)
- *There is a large sample of well-constructed complex sentences. There are, however, minor errors.* (C1, R4, RG2)
- *Excellent use of vocabulary allowing for precise accuracy and flexibility* (C2, R1, RG1)

2) Sociolinguistic features

Website writing samples at higher levels demonstrated better audience awareness through the use of the second person pronoun (‘you’), and rhetorical devices (e.g., questions). As a result, they appeared more engaging to readers, as indicated by the illustrative comment below:

- *Nice use of rhetorical question to address/engage audience* (C1, E4, EG2)

3) Content

Writers at high levels tended to elaborate on the research notes and synthesise them in building their arguments, as opposed to simply presenting a list of the points which seemed to characterise the website writing samples at B2.2. Moreover, writers at high levels were more skilful in developing and delivering their arguments as compared with those at a lower level and engaging the reader with literary or cultural references. Below are some illustrative comments:

- *Use of input not effective* (B2.2, E1, EG1)
- *As in the previous response, a lot is lifted from the prompt, but is used in a more sophisticated way that is less obvious to the reader* (C1, R2, RG1)
- *There is a stronger sense that the writer is giving an extended opinion on the issue. It seems that the material from the prompt is used to illustrate the writer’s ideas rather than meet the requirements of the test question* (C2, R3, RG1)

4) Cohesion and coherence

Website writing samples at high levels tended to feature more proficient use of linking or cohesive devices compared with those at a lower level. In addition, they were more clearly structured. Below are a few illustrative comments:

- *The structure is really ‘no structure’, with no sense of cohesion or organisation* (B2.2, E5, EG2)
- *Cohesion is better compared to the preceding example, but the overall organisation could have been better* (C1, E3, EG2)
- *Variety of discourse markers used accurately* (C2, R5, RG2)

5) Orthographic control

The written comments on orthographic control in most cases focused on spelling, punctuation, and paragraphing. Website writing samples at B2.2 were characterised by more spelling and punctuation errors or inappropriate paragraphing practices, as compared with those at C1 and C2. Below are a few illustrative comments:

- *Frequent errors – orthography, spelling, punctuation, capitalisation, full stops, grammar – word forms etc, vocabulary* (B2.2, E4, EG2)
- *Paragraphing is not always used appropriately* (C1, R2, RG1)
- *Helpful paragraphs guide the reader through the response* (C2, R3, RG1)

4.2.3 Discourse analysis of website writing samples

As Table 6 (see Section 3.3.2) indicates, the following discourse features were included in our discourse analysis of website writing samples (see also Appendix 4):

- style
- fluency
- accuracy
- syntactic complexity
- lexical complexity
- coherence
- content
- orthographic control

Findings concerning each discourse feature are detailed below.

1) Style

Style was operationalised as interestingness of website writing. A three-point scale was used to code the interestingness of website writing samples:

- ‘0’ – the writing was not interesting
- ‘1’ – the writing was somewhat interesting
- ‘2’ – the writing was very interesting.

Table 17 presents the frequency statistics of ‘interestingness’ across the three levels. We present the frequency statistics because the data was not normally distributed. As shown in this table, no website writing sample across the three levels was coded as ‘not interesting’; most website writing samples (n = 58) were coded as ‘somewhat interesting’ whereas only two ‘very interesting’. The Kruskal-Wallis test was performed to investigate whether ‘interestingness’ was significantly different as a function of the proficiency level. As expected, the results indicated that the difference was non-significant (x^2 = 1.07, df = 2, p = 0.601). We also performed a Mann-Whitney U test to compare whether the writing samples in the two test forms were different in terms of this feature, and the result was also non-significant (U = 420, n = 60, p = 0.154).

Table 17: Frequency statistics of interestingness across the three score levels (website article, n = 60)

Level	Not interesting	Somewhat interesting	Very interesting
B2.2 (n = 20)	0	19	1
C1 (n = 20)	0	19	1
C2 (n = 20)	0	20	0
Combined (n = 60)	0	58	2

2) Accuracy

As we did with the email writing samples, the proportion of EFTs was used as an indicator of accuracy. Table 18 presents the descriptive statistics of this measure. As shown in this table, the accuracy of website writing improves as the level of proficiency increases. ANOVA results, however, indicated that the difference was non-significant across the three levels (F = 0.89, df = 2, *p* = 0.418). Independent-samples t-test result indicated a non-significant difference (t = 0.10, df = 58, *p* = 0.921), suggesting that the website writing samples in the two test forms exhibited similar accuracy.

Table 18: Descriptive statistics of the proportion of EFTs across the three score levels (website article, n = 60)

Level	Proportion of EFTs	
	Mean	SD
B2.2 (n = 20)	0.56	0.20
C1 (n = 20)	0.62	0.25
C2 (n = 20)	0.65	0.25
Combined (n = 60)	0.61	0.23

3) Fluency

The following three measures were adopted to capture the fluency of the website writing samples:

- number of words
- number of T-units
- number of clauses.

Table 19 shows that the number of words increases with the increase of proficiency level. Note that this trend is more noticeable from B2.2 to C1. When it comes to the number of T-units and number of clauses, both of them increase as the proficiency level moves from B2.2 to C1; however, both of them decline, although by a small margin, as the proficiency level moves from C1 to C2 .

Table 19: Descriptive statistics of fluency measures across the three score levels (website article, n = 60)

Proficiency level	Number of words		Number of T-units		Number of clauses	
	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 20)	211.40	30.15	12.90	3.99	18.20	4.67
C1 (n = 20)	218.50	31.16	13.60	3.27	18.65	4.87
C2 (n = 20)	219.20	30.20	13.50	3.19	17.45	3.24
Combined (n = 60)	216.37	30.19	13.33	3.45	18.10	4.28

ANOVA results did not reveal significant differences at the three levels in focus for all three fluency measures that were included in our analysis (number of words: F = 0.40, df = 2, *p* = 0.672; number of T-units: F = 0.23, df = 2, *p* = 0.792; number of clauses: F = 0.39, df = 2, *p* = 0.677). Independent-samples t-tests did not reveal significant differences between the two forms in terms of the three fluency measures (number of words: t = -0.64, df = 58, *p* = 0.526; number of T-units: t = 0.30, df = 58, *p* = 0.768; number of clauses: t = 0.120, df = 58, *p* = 0.905).

4) Syntactic complexity

Three measures were used to represent syntactic complexity of the website writing samples:

- number of words per T-unit
- number of clauses per T-unit
- number of words per clause.

As indicated in Table 20, the number of words per T-unit declines as the proficiency level moves from B2.2 to C1, but then improves slightly from C1 to C2. Note that the standard deviation for this measure at the B2.2 level was considerably larger than that at the C1 and C2 levels, suggesting greater variability of this measure at this level. A different pattern could be observed of the number of clauses per T-unit which, as shown in Table 20, declines as the proficiency level increases. For the last syntactic complexity measure, namely, the number of words per clause, the mean value is the same for B2.2 and C1 but increases as the proficiency level moves from C1 to C2.

Table 20: Descriptive statistics of syntactic complexity measures across the three score levels (website article, n = 60)

Proficiency level	Number of words per T-unit		Number of clauses per T-unit		Number of words per clause	
	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 20)	18.05	6.86	1.50	0.56	12.28	3.11
C1 (n = 20)	16.73	3.43	1.38	0.21	12.28	2.81
C2 (n = 20)	16.82	3.32	1.32	0.21	12.79	1.83
Combined (n = 60)	17.20	4.78	1.40	0.37	12.45	2.61

ANOVA results indicated that the three syntactic complexity measures did not differ significantly as a function of proficiency level (number of words per T-unit: F = 0.47, df = 2, *p* = 0.630; number of clauses per T-unit: F = 1.28, df = 2, *p* = 0.286; number of words per clause: F = 0.25, df = 2, *p* = 0.779). Independent-samples t-tests did not reveal significant differences between the two forms in terms of the three syntactic complexity measures (number of words per T-unit: t = -1.04, df = 58, *p* = 0.305; number of clause per T-unit: t = -0.40, df = 58, *p* = 0.689; number of words per clause: t = -1.09, df = 58, *p* = 0.281).

5) Lexical complexity

Three measures were adopted to represent lexical complexity of the website writing samples:

- average word length (AWL)
- lexical density
- lexical sophistication.

Table 21 presents the descriptive statistics of the three measures across the three score levels. As shown in this table, AWL improves as proficiency level increases. With regards to both lexical density and lexical sophistication, their values are the same (after being rounded) for B2.2 and C1 but increase as the proficiency level moves from C1 to C2.

Table 21: Descriptive statistics of lexical complexity measures across the three score levels (website article, n = 60)

Proficiency level	AWL		Lexical density		Lexical sophistication	
	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 20)	4.66	0.21	0.53	0.04	0.25	0.06
C1 (n = 20)	4.69	0.26	0.53	0.04	0.25	0.06
C2 (n = 20)	4.82	0.17	0.54	0.02	0.30	0.04
Combined (n = 60)	4.73	0.22	0.53	0.04	0.26	0.06

ANOVA results indicated that neither AWL nor lexical density differed significantly across the three score levels (AWL: F = 0.25, df = 2, *p* = 0.779; lexical density: F = 0.31, df = 2, *p* = 0.055). The ANOVA showed, however, that lexical sophistication was significantly different across the three levels (F = 5.66, df = 2, *p* = 0.006)[4]. Post hoc tests revealed that a significant difference existed between B2.2 and C2 (*p* = 0.006) and C1 and C2 (*p* = 0.005). Independent-samples t-tests did not reveal significant differences between the two forms in terms of the three lexical complexity measures (AWL: t = -0.53, df = 58, *p* = 0.600; lexical density: t = 0.89, df = 58, *p* = 0.085; lexical sophistication: t = -1.95, df = 58, *p* = 0.056).

[4] In the analysis of lexical complexity measures, the critical value was adjusted to 0.05/3 = 0.017.

6) *Coherence*

The following four measures were used to capture the coherence of the website writing samples:

- referential cohesion (the overlap of content words and ideas across sentences and the entire text)
- argument overlap (local – overlap between sentences in terms of nouns and pronouns)
- argument overlap (global – overlap of nouns and pronouns across the entire text)
- all connectives (the incidence of all cohesive links between ideas and clauses).

Table 22 shows that the first three coherence measures, that is, referential cohesion, argument overlap (local), and argument overlap (global) all decline as the proficiency level increases from B2.2 to C2. A different pattern, however, is observed of the fourth coherence measure, that is, all connectives. As shown in Table 22, it increases as the proficiency level moves from B2.2 to C1 and then declines from C1 to C2.

ANOVA results did not reveal any significant difference in the four coherence measures across the three levels (referential cohesion: F = 0.72, df = 2, *p* = 0.491; argument overlap (local): F = 1.03, df = 2, *p* = 0.364; argument overlap (global): F = 0.68, df = 2, *p* = 0.509; all connectives: F = 0.87, df = 2, *p* = 0.425). Independent-samples t-tests did not reveal significant differences between the two forms in terms of the four selected measures of coherence and cohesion (referential cohesion: t = 0.85, df = 58, *p* = 0.400; argument overlap [local]: t = 1.98, df = 58, *p* = 0.053; argument overlap [global]: t = 1.65, df = 58, *p* = 0.105; all connectives: t = 1.17, df = 58, *p* = 0.248).

Table 22: Descriptive statistics of coherence measures across the three score levels (website article, n = 60)

Proficiency level	Referential cohesion		Argument overlap (local)		Argument overlap (global)		All connectives	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 20)	27.06	24.68	0.48	0.22	0.41	0.23	102.44	21.91
C1 (n = 20)	23.39	21.26	0.45	0.20	0.38	0.16	111.14	20.62
C2 (n = 20)	18.66	20.37	0.39	0.18	0.34	0.18	106.13	20.24
Combined (n = 60)	23.04	22.08	0.44	0.20	0.38	0.19	106.57	20.89

7) *Content*

In the Aptis Advanced writing test, test-takers are required to use the research notes and statistics provided to them in the website writing task. The content of website writing samples was therefore coded based on two criteria:

- use of research notes
- use of statistics.

A three-point scale was applied to code the use of research notes and statistics (see Appendix 4):

- 0 – ineffective use or no mention of the research notes/statistics to support or expand the writer's arguments
- 1 – somewhat effective use of the research notes/statistics to support or expand the writer's arguments
- 2 – very effective use of the research notes/statistics to support or expand the writer's arguments.

As shown in Table 23, 'use of research notes' improves from B2.2 to C1, and then levels off; use of statistics, however, improves consistently as the proficiency level increases.

Table 23: Descriptive statistics of content measures across the three score levels (website article, n = 60)

Proficiency level	Use of research notes		Use of statistics	
	Mean	SD	Mean	SD
B2.2 (n = 20)	1.20	0.41	0.95	0.51
C1 (n = 20)	1.55	0.51	1.00	0.56
C2 (n = 20)	1.55	0.51	1.10	0.45
Combined (n = 60)	1.43	0.50	1.02	0.50

ANOVA results indicated that neither 'use of notes' (F = 3.55, df = 2, *p* = 0.035)[5] nor 'use of statistics' (F = 0.45, df = 2, *p* = 0.639) differed significantly across the three levels. Independent-samples t-tests did not reveal significant differences between the two forms in terms of the two content measures (use of notes: t = -1.57, df = 58, *p* = 0.122; use of statistics: t = -0.25, df = 58, *p* = 0.800).

8) Orthographic control

Three measures were used to capture the orthographic control of website writing:

- spelling errors
- punctuation errors
- paragraphing.

To code the first two measures, the numbers of spelling and punctuation errors were counted in website writing samples. For paragraphing, a three-point scale was applied to code the website writing samples:

- 0 – ineffective paragraphing
- 1 – somewhat effective paragraphing
- 2 – effective paragraphing.

[5] The critical value was adjusted to 0.05/2 = 0.025.

A standardisation procedure was implemented to make the measures of spelling and punctuation comparable. We divided the number of spelling and punctuation errors by the number of words in a writing sample. The descriptive statistics for these three measures are presented in Table 24. As indicated in this table, 'spelling error' decreases with the increase of proficiency level. This trend is particularly noticeable when the proficiency level moves from B2.2 to C1; 'punctuation error' decreases from B2.2 to C1 but increases slightly from C1 to C2. Finally, paragraphing seems to improve as the proficiency level increases.

Table 24: Descriptive statistics of orthographic control measures across the three score levels (website writing, n = 60)

Proficiency level	Spelling errors		Punctuation errors		Paragraphing	
	Mean	SD	Mean	SD	Mean	SD
B2.2 (n = 20)	0.024	0.020	0.034	0.019	1.05	0.67
C1 (n = 20)	0.011	0.013	0.023	0.014	1.45	0.76
C2 (n = 20)	0.009	0.010	0.025	0.014	1.60	0.60
Combined (n = 60)	0.015	0.016	0.028	0.016	1.37	0.71

Prior to ANOVA, we checked the data distribution of these three variables and observed that only the data for 'spelling errors' violated the normal distribution assumption. Therefore, it was decided to apply the Kruskal-Wallis test to the spelling data whereas a regular ANOVA procedure was applied to that of punctuation and paragraphing. The Kruskal-Wallis test indicated that 'spelling errors' differed significantly across the three levels (x^2 = 10.97, df = 2, *p* = 0.004)[6]. Post hoc tests with the Dunn-Bonferroni method revealed that significant differences existed between B2.2 and C1 (*p* = 0.033) and between B2.2 and C2 (*p* = 0.006). The Mann-Whitney U test did not indicate a significant difference between the two forms in terms of 'spelling errors' (U = 422, n = 60, *p* = 0.677).

ANOVA results indicated that neither 'punctuation errors' nor 'paragraphing' differed significantly across the three levels (punctuation errors: F = 3.08, df = 2, *p* = 0.054; paragraphing: F = 3.45, df = 2, *p* = 0.038). Independent-samples t-tests revealed that Form A had significantly more punctuation errors than Form B, with a medium effect size (t = 2.45, df = 58, *p* = 0.017, d = 0.57); however, the two forms were similar with respect to 'paragraphing' (t = -0.22, df = 58, *p* = 0.830).

[6] Given that three measures were adopted to reflect orthographic control, the critical value was adjusted to 0.05/3 = 0.017 accordingly.

4.2.4 Summary of findings – website article

Five categories of writing were identified in the verbal and written comments made by ESL experts and Aptis raters, including:

- vocabulary and grammar
- sociolinguistic features
- content
- cohesion and coherence
- orthographic control.

Findings concerning each feature are summarised below.

1) *Vocabulary and grammar*

Website writing samples at high levels demonstrated more proficient use of vocabulary and grammar in the sense that they tended to use more sophisticated and a wider range of vocabulary and grammar. Website writing samples at B2.2, however, were characterised by noticeably more errors. Although the same trend appeared in the discourse measures, only lexical sophistication successfully differentiated between levels.

2) *Sociolinguistic features*

Website writers at high levels demonstrated better audience awareness through several approaches such as using the second person pronoun ‘you’, questions, and cultural or literary references in their writing; writing samples at high levels were also found to be more interesting to readers. Discourse analysis results, however, suggested similar level of interestingness across the three levels in focus.

3) *Content*

Compared with website writing samples at a lower level, particularly B2.2, those at high levels seemed to: a) make better use of the source information in the sense that they elaborated on the research notes and/or statistics and integrated them in building their arguments; and b) deliver their arguments more effectively. These findings, however, were not supported by discourse analysis results.

4) *Cohesion and coherence*

Website writing samples at high levels exhibited more skilful use of linking or cohesive devices; they were also characterised by a clearer organisation and more logical progression of ideas. Discourse analysis results, however, did not lend support to these findings, suggesting that the writing samples demonstrated similar cohesion and coherence across the three score levels.

5) *Orthographic control*

Website writing samples at B2.2 seemed to have more issues with orthographic control, including spelling and punctuation errors and inappropriate paragraphing practices. This finding was partly supported by discourse analysis results which indicated that writing samples at C1 and C2 had fewer spelling errors as compared with those at B2.2.

In addition to the findings listed above, another finding which is worth mentioning is that except 'punctuation errors', no significant differences were detected across the two test forms with regard to other discourse features that were included in our analysis, thus generally supporting the equivalence of the website article task in the parallel test forms. This finding will be further discussed in Section 5.

5. DISCUSSION AND IMPLICATIONS

It is challenging to differentiate L2 writers' performances at higher levels as they tend to be based on rhetorical skills and expression of complex ideas, which have been less researched and understood (e.g., Banerjee, et al., 2015; Hulstijn, 2015). In consequence, writing test developers face the daunting task to develop rating scales that can be used to capture and discriminate test-takers' performances at higher levels.

In this study, we set out to investigate the discourse features of advanced writing on two task types of the Aptis Advanced writing test (i.e. email and website) from multiple perspectives with the view of further explicating the task-specific rating criteria. Adopting an exploratory sequential mixed-methods design, the study collected data from both ESL experts and Aptis raters. According to Knoch et al., multiple sources can be identified and solicited for rating scale development and revision, including, for example, expert intuition, language proficiency frameworks such as the CEFR, the analysis of the target language use (TLU) domain, and theory/literature review. Knoch et al. categorised these sources into two broad categories of test-internal and test-external, with the former referring to review or analysis of test-takers' writing samples, input from raters, among others, whereas the latter encompasses the analysis of the TLU domain or the language proficiency frameworks, etc. The views from the two groups of participants in this study, that is, ESL experts and Aptis raters, represent the test-internal sources. Another important test-internal source is the analysis of test-takers' writing samples which was also included in this study. By incorporating the findings from both the qualitative (i.e. the comments by ESL experts and Aptis raters) and quantitative analyses (i.e. the discourse analysis of test-takers' writing samples), this study illuminates the writing constructs that are assessed in the two tasks of the Aptis Advanced writing test, that is, email response and website.

The findings from the qualitative component of this study are largely consistent with the characteristics of advanced writing identified through the literature review. For example, both ESL experts and Aptis raters commented that for both email and website writing samples, writers at higher levels (C1 and C2) used more sophisticated and a wider range of vocabulary and grammar with improved accuracy whereas those at a lower level (i.e. B2.2) tended to make more mistakes in their choice of words or grammatical structures though most of them were considered as not impeding comprehension. This finding is not surprising and could find resonance in a number of previous studies on L2 writing, indicating that advanced writers tend to use more sophisticated and varied lexicon and/or more complex grammatical structures to demonstrate their writing ability with fewer errors. For example, in their analysis of test-takers' writing samples on the Academic Writing Task 2 of IELTS, Mayor et al. (2007) found that writers at higher levels tended to use more sophisticated grammatical structures with fewer errors, as compared with those at lower levels.

In another comprehensive discourse analysis study of test-takers' writing samples on the TOEFL iBT, Knoch et al. (2014) revealed that lexical sophistication was one of the variables identified as best differentiating different writing levels. Thewissen's (2013) analysis of the types of errors made by L2 writers at four levels on the CEFR (i.e. B1, B2, C1, and C2) is yet another example, which clearly indicated that writing samples at higher levels featured improved accuracy with significantly fewer errors, though certain error types were found to persist with the increase of proficiency levels.

In addition, the two groups of participants also commented that test-takers at higher levels were more skilful in using the notes and/or statistics that were provided to them in the sense that they tended to either elaborate on or expand the source information in building their arguments, as opposed to directly copying the information verbatim in their writing. This is particularly the case in the website article task where test-takers are expected to write a more extended response of 180 to 220 words. This observation is again largely consistent with our expectations and also resonates with some previous research involving L2 writers at higher proficiency levels. In their analysis of test-takers' writing samples on the TOEFL iBT, for example, Staples et al. (2013) found that L2 writers at lower proficiency levels tended to use the lexical bundles that appeared in the task prompts; however, those at higher proficiency levels relied less on the unanalysed chunks in the task prompts and were more capable of drawing on formulaic language from outside the test prompts.

While the focus group discussions and written comments pointed to differences between these levels, the discourse analysis resulted in statistically significant differences with only a few discourse features. For the email response task, three features differentiated between levels, including average word length, spelling errors and punctuation errors. For the website article task, two features, that is, lexical sophistication and spelling errors, indicated statistically significant differences. As our aim was to confirm the findings from the focus group study with the discourse analyses on a larger sample in an exploratory sequential mixed methods design, the findings may seem disappointing at first glance. There are two possible avenues to interpret the results.

Firstly, one could argue that our data offers very little information for the revision of the existing rating scales. However, we feel that the findings can be explained based on the cyclical nature of the design of our study, as is the case with most existent research on L2 writing. To code the discourse analytic measures at the three highest proficiency levels, we drew on samples that had been rated by Aptis raters using the existing rating criteria. As we have mentioned in the literature review section (Section 2.2) of this report, these scales are holistic in nature, and lack detail at the higher levels. Arguably, raters may find it difficult to clearly differentiate between specific features at higher levels, and may as a result draw on their own internal criteria, which may all differ slightly. It is telling that accuracy was found to be nearing significance in the email response task, but did not reach significance. This finding is not surprising as the qualitative analysis of ESL experts' and Aptis raters' written comments showed that the samples at the two highest levels are generally accurate and might be distinguished by other features (or combination of features). Another shortcoming of the quantitative analysis is that features of writing at higher levels are often difficult to quantify, and are more nuanced than what existing discourse analytic measures are able to capture. We will take this point up in our discussion of the implications of this study.

An alternative avenue to interpret (and use) the data might be to disregard the findings from the discourse analysis, and focus on the qualitative findings, which revealed quite clear differences between the three levels of interest. This qualitative information gathered from the focus group study is sufficiently rich to apply to a revised rating scale, as we will show below. In the following section, we discuss the findings separately for the task of email response and website article.

5.1 Email response

In this section, we summarise the findings for email writing, and propose an alternative wording for the rating scales at the three levels in question (see Table 25 below). It should be noted that we have not ordered the points in any particular manner, and that further thinking about the possible presentation of these features is needed, ideally in consultation with raters.

As is typically found when comments are elicited from raters in studies similar to this one, the highest number of comments were about vocabulary and grammar. A summary of the findings showed that the positive comments about C1 and C2 were relatively similar, but that the negative comments were much higher at B2.2 level. C2 was marked by more sophisticated vocabulary and grammar. We have indicated some possible elaborations to the existing scales in Table 25.

In terms of the sociolinguistic features, writers at C2 level used openings and closings appropriately (similarly to C1), but the writing at C2 level was also marked as more polite. Writers at C2 and C1 expanded and/or elaborated on the source material, while writers at B2.2 tended to use the source material with little elaboration, or copied the source material verbatim. Arguments were more convincing with an effective introduction and conclusion at the two higher levels. In terms of cohesion and coherence, writing at higher levels was marked with better use of transition devices. At Level C2, these were described as being used sophisticatedly and accurately. At B2.2, the use of transition markers was described as being mechanical at times. Finally, in terms of orthographical control, this improved at higher levels with a consistent upward trend.

Table 25: Possible revisions to the descriptors for the email response task based on findings of this study

Level	Current descriptors	Possible revisions
C2	Response fully on-topic and appropriate register used. Response shows the following features: • Range of complex grammar constructions used accurately. No grammar errors occur. • Range of vocabulary used. No awkward or inappropriate lexical choices. • No more than two minor punctuation or spelling errors occur. • Range of cohesive devices used to clearly indicate the links between ideas.	• Openings and closings are used appropriately. • Response generally polite. • Sophisticated vocabulary and grammar are used. • Grammar and vocabulary generally accurate. • Effective introduction and conclusion; convincing arguments put forward. • Source material expanded/elaborated. • Ideas well organised. • Sophisticated transition devices used accurately, appropriately and 'smoothly'. • Spelling, punctuation and paragraphing mostly accurate.
C1	Response fully on-topic and appropriate register used. Response shows the following features: • Range of complex grammar constructions used accurately. Minor grammar errors occur. • Range of vocabulary used. Some awkward or slightly inappropriate lexical choices. • Minor errors in punctuation and spelling occur. • Range of cohesive devices used to clearly indicate the links between ideas.	• Openings and closings are used appropriately. • Mix of sophisticated and more simple grammar and vocabulary. • Grammar and vocabulary generally accurate. • Effective introduction and conclusion; convincing arguments put forward. • Source material expanded/elaborated. • Ideas well organised. • Transition devices used appropriately. • Spelling, punctuation and paragraphing mostly accurate.
B2.2	Response fully on-topic and appropriate register used. Response shows the following features: • Some complex grammar constructions used accurately. Errors do not impede understanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not impede understanding. • Punctuation and spelling errors do not impede understanding. • Limited number of cohesive devices are used to indicate the links between ideas.	• Openings and closings not always used appropriately. • While errors occur, they do not impede understanding. • Arguments put forward may not be convincing. • Source material not further elaborated, or copied from prompt. • Ideas not always well organised. • Transition devices may be overused or used mechanically. • Some errors or inconsistencies in spelling, punctuation and paragraphing.

5.2 Website article

As was the case for the email writing tasks, vocabulary and grammar attracted the most comments in the focus groups. The comments generally mirrored what was found for the email response task, with higher level samples being characterised by more sophisticated and accurate grammar and a wider range and more accurate choice of vocabulary.

In terms of sociolinguistic features, writers at higher levels were able to display audience awareness. They drew on a range of devices (e.g., rhetorical questions, personal pronoun ‘you’, cultural or literary references), which resulted in more engaging writing. When evaluating the content of the response, samples at the two higher levels were more successful in elaborating on the research notes and/or statistics provided and using them to build a successful argument. In terms of cohesion and coherence and orthographic control, the comments in the focus group were no different to the email response task. In Table 26 below, we make suggestions for possible revisions of the current scale descriptors for the website article task.

Table 26: Possible revisions to the descriptors for the website article task based on findings of this study

Level	Current descriptors	Possible revisions
C2	Response is informative and interesting. Clever use of the input with mostly original text. Response shows the following features: • Range of complex grammar constructions used accurately. Minor errors occur. • Range of vocabulary used accurately. No awkward or inappropriate lexical choices. • Minor errors in punctuation and spelling occur. • Range of cohesive devices used to clearly indicate the links between ideas.	• Sophisticated vocabulary and grammar are used. • Grammar and vocabulary generally accurate. • Response displays audience awareness by drawing on devices such as rhetorical questions, use of personal pronoun 'you' or cultural/literary reference (if suitable to prompt). • Source material expanded/elaborated; used to build a convincing argument. • Ideas well organised. • Sophisticated transition devices used accurately, appropriately and 'smoothly'. • Spelling, punctuation and paragraphing mostly accurate.
C1	Response is not interesting. Correct use of the input with mostly original text. Response shows the following features: • Range of complex grammar constructions used accurately. Minor errors occur. • Range of vocabulary used. Some awkward or slightly inappropriate lexical choices. • Minor errors in punctuation and spelling occur. • Range of cohesive devices used to clearly indicate the links between ideas.	• Mix of sophisticated and more simple grammar and vocabulary. • Grammar and vocabulary generally accurate. • Response displays audience awareness by drawing on devices such as rhetorical questions, use of personal pronoun 'you' or cultural/literary reference (if suitable to prompt). • Source material expanded/elaborated. • Ideas well organised. • Transition devices used appropriately. • Spelling, punctuation and paragraphing mostly accurate.
B2.2	Response is not interesting. Correct use of the input with mostly original text. Response shows the following features: • Some complex grammar constructions used accurately. Errors do not impede understanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not impede understanding. • Punctuation and spelling errors do not impede understanding. • Limited number of cohesive devices are used to indicate the links between ideas.	• While errors occur, they do not impede understanding. • Arguments put forward may not be convincing. • Source material not further elaborated, or copied from prompt. • Ideas not always well organised. • Transition devices may be overused or used mechanically. • Some errors or inconsistencies in spelling, punctuation and paragraphing.

To ensure that our findings are consistent across test forms, we included two forms of each task in our discourse analysis. For the email response task, three measures were found to significantly distinguish the two test forms, that is, accuracy (indicated by the proportion of error-free T-units), lexical complexity and argument overlap (global). The finding might be interpreted in relation to the two different topics in the email response task. Though only three measures (out of a total of 22, see Table 6) were significantly different across the two test forms, this raises concerns over the equivalence of the email response task in parallel test forms. A more detailed investigation is beyond the scope of this report. Future research is warranted to interrogate parallel test form equivalence in the email response task in the Aptis Advanced writing test, as it relates to score interpretation and fairness (AERA, APA, & NCME, 2014). In comparison, only one feature of the website article responses (out of a total of 20, see Table 6), that is, punctuation errors, was found to be significantly different across the two parallel forms. This finding lends support to the equivalence of the website article task, at least based on the two tasks that were used in this study.

This study has a number of implications. In terms of theoretical implications for rating scale development and revision, the study shows some of the possible drawbacks of coding discourse measures based on scripts scored using an existing holistic rating scale. Our analysis resulted in very few significant differences between the three levels. There may be two reasons for this. Firstly, due to the cyclical nature of using scripts rated using an existing rating scale, the results of such an analysis will always be influenced by features in the scale that raters used to score the scripts. Secondly, features of advanced writing may also not be easily quantifiable in the same manner as characteristics of writing at lower levels. This is because advanced writing is marked by finer nuances, and more diversity of features, as the features that L2 writers at advanced levels could access expand.
To counter the first issue, it may be more fruitful in future research to revise the rating scale based on rater comments and then, based on new ratings, confirm the scale descriptors with a discourse-based study. Alternatively, it may be worth circumventing existing scores and categorising writing samples based on a different sorting system. To counter the second problem, we suggest considering alternative avenues of analysing advanced writing. The comments collected from raters and L2 writing experts were fruitful and provided sufficient information to add further detail to the existing scale descriptors. We also found that our practice of counting the positive and negative comments in the analysis of participants' written comments provided excellent guidance on how to add information to the level descriptors.

A further implication comes out of a comment made by some raters and several writing experts, who mentioned that they did not feel the tasks sufficiently elicited high-level writing. The Aptis program may want to consider replacing one of the writing tasks with a more complex task that is able to 'push' advanced writers to display more complex features of writing.

REFERENCES

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

Banerjee, J., Franceschina, F., & Smith, A. M. (2007). Documenting features of written language production typical at different IELTS band score levels. *IELTS Research Reports 2007, Vol 7*, 1. IELTS Australia, Canberra and British Council, London.

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing writing, 26*, pp. 5–19.

British Council. (2020). Technical report: Aptis Advanced technical supplement. Retrieved 23 November 2020 from: https://www.britishcouncil.org/sites/default/files/aptis_advanced_technical_supplement_final.pdf.

Cobb, T. (2002). VocabProfile. Retrieved from: http://www.lextutor.ca/vp/.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum Associates.

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: Sage.

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*, pp. 1–16.

Cumming, A., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). Analysis of Discourse Features and Verification of Scoring Levels for Independent and Integrated Prototype Written Tasks for the New TOEFL®. TOEFL® Monograph Series. MS-30. ETS RM-05-13. *ETS Research Report Series*.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg, France: Council of Europe.

Field, A. (2009). *Discover statistics using SPSS*. London: Sage.

Fulcher, G. (2003). *Testing second language speaking*. London, UK: Pearson.

Fulcher, G. (2012). Scoring performance tests. In Fulcher, G. & Davidson, F. (Eds.) *Routledge Handbook of Language Testing.* New York, NY: Routledge.

Fulcher, G. (2014). *Testing Second Language Speaking*. New York: Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), pp. 5–29.

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

Hamilton, J., Lopes, M., McNamara, T., & Sheridan, E. (1993). Rating scales and native speaker performance on a communicatively oriented EAP test. *Language Testing, 10*(3), pp. 337–353.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research.* Amsterdam: John Benjamins Publishing Company.

IBM. (2012). IBM SPSS statistics version 21. Boston, Massachusetts.

Isbell, D. R. (2017). Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. *Assessing writing, 34*, pp. 37–49.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing writing, 16*(2), pp. 81–96.

Knoch, U., Deygers, B. & Khamboonruang, A. (under review). Revisiting rating scale development for rater-mediated language performance assessment: Modelling construct and contextual choices made by scale developers.

Knoch, U., Macqueen, S., & O'Hagan, S. (2014). An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT® writing test. *ETS Research Report Series, 2014*(2), pp. 1–74.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), pp. 647–672.

Mayor, B., Hewings, A., North, S., Swann, J., & Coffin, C. (2007). A linguistic analysis of Chinese and Greek L1 scripts for IELTS Academic Writing Task 2. In L. Taylor & P. Falvey (Eds.), *IELTS collected papers: Research in speaking and writing assessment* (Vol. 19). Cambridge: Cambridge University Press.

McNamara, T., Hill, K., & May, L. (2002). 12. Discourse and assessment. *Annual Review of Applied Linguistics, 22*, pp. 221.

Meisel, J. M., Clahsen, H., & Pienemann, M. (1981). On determining developmental stages in natural second language acquisition. *Studies in second language acquisition, 3*(2), pp. 109–135.

Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook*. 3rd ed: Thousand Oaks, CA: Sage.

Montee, M. & Malone, M. (2014). Writing scoring criteria and score reports (pp. 1–13). In A. Kunnan (ed.), *The companion to language assessment.* London, UK: John Wiley & Sons..

QSR. (2012). *NVivo qualitative data analysis software*. Melbourne: QSR International Pty Ltd.

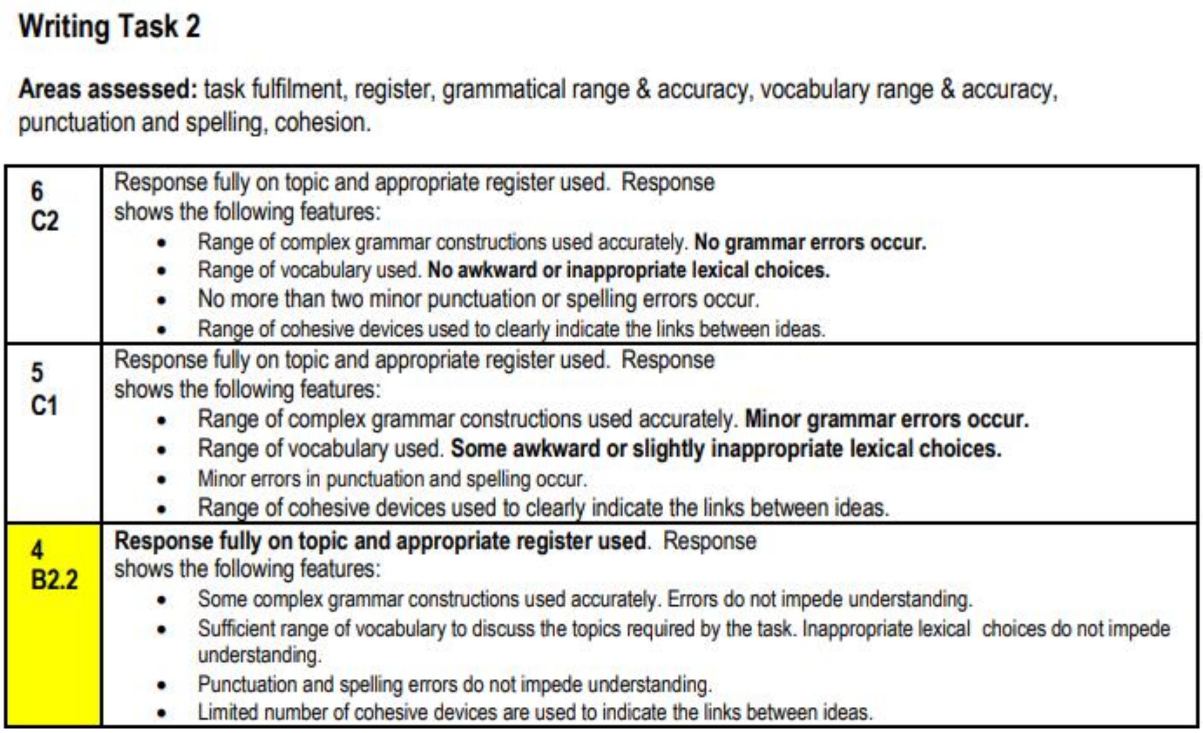
Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*(3), pp. 214–225.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal, 97*(S1), pp. 77–101.

Yang, W., & Sun, Y. (2012). The use of cohesive devices in argumentative writing by Chinese EFL learners at different score levels. *Linguistics and Education, 23*(1), pp. 31–48.

APPENDIX 1:

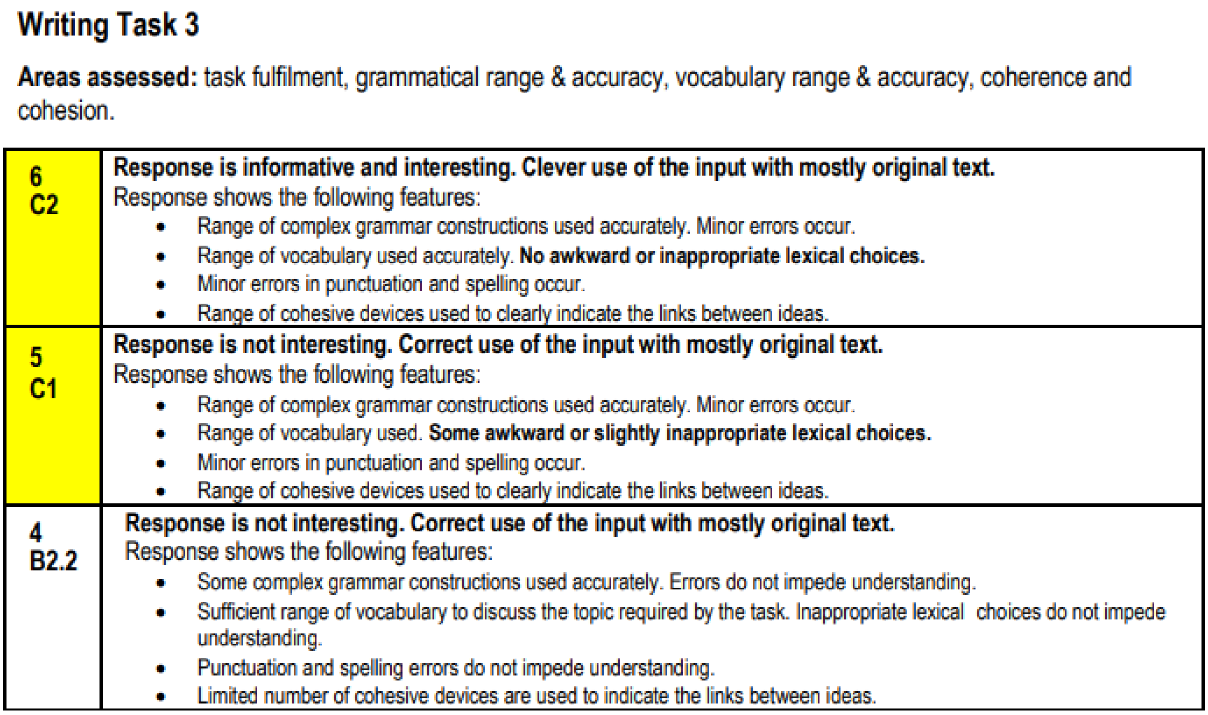
Rating criteria for Task B and C in the Aptis Advanced writing test (from British Council, 2020, pp. 40–41)



Writing Task 2

Areas assessed: task fulfilment, register, grammatical range & accuracy, vocabulary range & accuracy, punctuation and spelling, cohesion.

6 C2	Response fully on topic and appropriate register used. Response shows the following features: • Range of complex grammar constructions used accurately. **No grammar errors occur.** • Range of vocabulary used. **No awkward or inappropriate lexical choices.** • No more than two minor punctuation or spelling errors occur. • Range of cohesive devices used to clearly indicate the links between ideas.
5 C1	Response fully on topic and appropriate register used. Response shows the following features: • Range of complex grammar constructions used accurately. **Minor grammar errors occur.** • Range of vocabulary used. **Some awkward or slightly inappropriate lexical choices.** • Minor errors in punctuation and spelling occur. • Range of cohesive devices used to clearly indicate the links between ideas.
4 B2.2	**Response fully on topic and appropriate register used.** Response shows the following features: • Some complex grammar constructions used accurately. Errors do not impede understanding. • Sufficient range of vocabulary to discuss the topics required by the task. Inappropriate lexical choices do not impede understanding. • Punctuation and spelling errors do not impede understanding. • Limited number of cohesive devices are used to indicate the links between ideas.



Writing Task 3

Areas assessed: task fulfilment, grammatical range & accuracy, vocabulary range & accuracy, coherence and cohesion.

6 C2	**Response is informative and interesting. Clever use of the input with mostly original text.** Response shows the following features: • Range of complex grammar constructions used accurately. Minor errors occur. • Range of vocabulary used accurately. **No awkward or inappropriate lexical choices.** • Minor errors in punctuation and spelling occur. • Range of cohesive devices used to clearly indicate the links between ideas.
5 C1	**Response is not interesting. Correct use of the input with mostly original text.** Response shows the following features: • Range of complex grammar constructions used accurately. Minor errors occur. • Range of vocabulary used. **Some awkward or slightly inappropriate lexical choices.** • Minor errors in punctuation and spelling occur. • Range of cohesive devices used to clearly indicate the links between ideas.
4 B2.2	**Response is not interesting. Correct use of the input with mostly original text.** Response shows the following features: • Some complex grammar constructions used accurately. Errors do not impede understanding. • Sufficient range of vocabulary to discuss the topic required by the task. Inappropriate lexical choices do not impede understanding. • Punctuation and spelling errors do not impede understanding. • Limited number of cohesive devices are used to indicate the links between ideas.

APPENDIX 2:

Background questionnaires for raters and experts

Background questionnaire (ESL experts)

1. Your name (to be treated confidential): ....
2. Highest relevant qualification:..
3. Year when your highest qualification was completed: ....
4. Number of years' experience in teaching ESL: ....
5. Number of years' experience in teaching ESL writing: ....
6. Writing courses taught/currently teaching: ....

Background questionnaire (Aptis raters)

1. Your name (to be treated confidential): ....
2. Your language background: ....
3. Highest relevant qualification: ....
4. Year when your highest qualification was completed: ....
5. Number of years' experience in teaching ESL: ....
6. Number of years' experience in teaching ESL writing: ....
7. Number of years' experience in rating the Aptis writing: ....
8. Number of years' experience (if any) in rating other writing tests: ....
9. Briefly describe your experience in rating writing tests: ....

APPENDIX 3:

Focus groups coding scheme

Coding scheme (email writing – focus groups)

Category	Discourse feature
Vocabulary and grammar	• Accuracy • Sophistication
Sociolinguistic features	• Email conventions • Polite style • Audience awareness
Content	• Use of source information • Delivery of arguments
Cohesion and coherence	• Use of linking words or cohesive devices
Orthographic control	• Spelling, punctuation, and paragraphing

Coding scheme (website writing – focus groups)

Category	Discourse feature
Vocabulary and grammar	• Accuracy • Sophistication
Sociolinguistic features	• Audience awareness
Content	• Use of source information • Delivery of arguments • Use of cultural or literary references
Cohesion and coherence	• Use of linking words or cohesive devices • Organisation
Orthographic control	• Spelling, punctuation, and paragraphing

APPENDIX 4:

Coding scheme for written comments

Coding scheme (email writing, written comments)

Category	Examples	
	Positive	**Negative**
Vocabulary and grammar	• *Wide range of vocabulary accurately used*	• *Some grammatical/vocab choice errors*
	• *Only a few very minor grammatical errors*	• *Odd choice of expression at times*
Sociolinguistic features	• *Adheres to conventions of email opening and closing*	• *No opening or closing*
	• *Tone appropriate for email*	• *Quite a frustrated tone: complaint-focused*
Content	• *Arguments reinforced with supporting ideas and specific examples*	• *Lack of synthesising the information provided*
	• *Effective use of detail-clarifies*	• *Introduces unconvincing/ irrelevant ideas to support argument*
Cohesion and coherence	• *Effective and accurate use of discourse markers*	• *Relationships of ideas weak*
	• *Very good organisation and overall cohesion*	• *Cohesive devices limited*
Orthographic control	• *Mostly accurate spelling*	• *Many punctuation errors*
	• *Appropriate paragraph organisation*	• *Mechanical, punctuation errors*

Coding scheme (website writing, written comments)

Category	Examples	
	Positive	Negative
Vocabulary and grammar	- *Wide range of vocabulary used accurately to convey precise meaning*	- *Grammatical errors noticeable and frequent and make reading a bit challenging/slow reading*
	- *More sophisticated use of structures/grammar and expression overall*	- *Inappropriate use of expressions at times*
Sociolinguistic features	- *Good story title – attracts on attention*	- *No awareness of audience*
	- *Effective use of metaphor: 'big brother' – taps into literature, plus contemporary culture, so giving an instant context and sense to the reader*	- *Attempts at using engaging rhetorical devices (well, yes, and '?') are unsophisticated*
Content	- *The argument and language here are quite well developed*	- *No use of statistical info*
	- *Good use of the information provided*	- *Argumentation disjointed: thoughts are not expressed clearly or in a logical order*
Cohesion and coherence	- *Nice use of cohesive markers*	- *The structure is really 'no structure', with no sense of cohesion or organisation*
	- *Variety of discourse markers used accurately*	- *Sentences lack cohesion; they are mostly a list of points*
Orthographic control	- *Helpful paragraphs guide the reader through the response*	- *Many spelling and punctuation issues-often faulty*
	- *Very few spelling errors and punctuation is handled well throughout*	- *Spelling, word form issues*

APPENDIX 5:

Discourse analysis coding scheme

Coding scheme (email writing, discourse analysis)

Category	Discourse features	Codes/Values	Notes
Email conventions	Inclusion of opening	1, 0	Segments like 'dear …', 'To whom it may …', 'hi…', 'Hello…'
	Inclusion of closing	1, 0	Segments like 'Yours sincerely', 'thank you, … (name)'
	Appropriateness of opening	1, 0	Opening appropriate for the context of the situation in question; 1 = Yes; 0 = No
	Appropriateness of closing	1, 0	Closing appropriate for the context of the situation in question; 1 = Yes; 0 = No
Accuracy	Proportion of error-free T-units	No. of error-free T-units; No. of T-units	Number of error-free T-units/Total number of T-units
Fluency	Number of words	No. of words	Automated coding
	Number of clauses	No. of clauses	Number of clauses per text
	Number of T-units	No. of T-units	Number of T-units per text
Syntactic complexity	Words per T-unit	No. of words; no. of T-units	Number of words/Number of T-units
	Clauses per T-unit	No. of clauses; no. of T-units	Number of clauses/Number of T-units
	Words per clause	No. of words; no. of clauses	Number of words/Number of clauses
Lexical complexity	Average word length	Average word length	Automated
	Lexical density	Lexical density	Automated
	Lexical sophistication	Lexical sophistication	Automated
Coherence	Referential cohesion	Referential cohesion	Automated
	Argument overlap (local)	Argument overlap (local)	Automated
	Argument overlap (global)	Argument overlap (global)	Automated
	All connectives	All connectives	Automated
Content	Idea units	No. of idea units covered	Number of idea units based on the three notes in the task description, their mention and coverage. For each idea unit: 0 – not mentioned 1 – mentioned briefly 2 – sufficiently covered.
Orthographic control	Spelling	No. of spelling mistakes	Number of spelling mistakes/Number of words
	Punctuation	No. of punctuation mistakes	Number of punctuation mistakes/Number of words

Coding scheme (website writing, discourse analysis)

Category	Discourse features	Codes/Values	Notes
Style	Interestingness	0, 1, 2	Does the text have the following features? a. Use of questions. b. Listing facts in intriguing/engaging ways. c. Use of cultural references. Each script is coded as below: 2 – Very interesting 1 – Somewhat interesting 0 – Uninteresting.
Accuracy	Proportion of error-free T-units	No. of error-free T-units; no. of T-units	Number of error-free T-units/Total number of T-units
Fluency	Number of words	No. of words	Automated coding
	Number of clauses	No. of clauses	Number of clauses per text
	Number of T-units	No. of T-units	Number of T-units per text
Syntactic complexity	Words per T-unit	No. of words; no. of T-units	Number of words/Number of T-units
	Clauses per T-unit	No. of clauses; no. of T-units	Number of clauses/Number of T-units
	Words per clause	No. of words; no. of clauses	Number of words/Number of clauses
Lexical complexity	Average word length	Average word length	Automated
	Lexical density	Lexical density	Automated
	Lexical sophistication	Lexical sophistication	Automated
Coherence	Referential cohesion	Referential cohesion	Automated
	Argument overlap (local)	Argument overlap (local)	Automated
	Argument overlap (global)	Argument overlap (global)	Automated
	All connectives	All connectives	Automated
Content	Use of research notes	0, 1, 2	Each script is coded as below: 0 – Ineffective use/no mention of the research notes to support or expand the writer's arguments 1 – Somewhat effective use of the research notes to support or expand the writer's arguments 2 – Very effective use of the research notes to support or expand the writer's arguments.
	Use of statistics	0, 1, 2	Each script is coded as below: 0 – Ineffective use/no mention of the statistics to support or expand the writer's arguments 1 – Somewhat effective use of the statistics to support or expand the writer's arguments 2 – Very effective use of the statistics to support or expand the writer's arguments.
Orthographic control	Spelling	No. of spelling mistakes	Number of spelling mistakes/Number of words
	Punctuation	No. of punctuation mistakes	Number of punctuation mistakes/Number of words
	Paragraphing	No. of paragraphs	Number of paragraphs/Number of words

British Council Assessment Research Awards and Grants

If you're involved or work in research into assessment, then the British Council Assessment Research Awards and Grants might interest you.

These awards recognise achievement and innovation within the field of language assessment and form part of the British Council's extensive support of research activities across the world.

INVESTIGATING THE DISCOURSE PRODUCED AT SCORE LEVELS B2.2 TO C2 ON THE APTIS ADVANCED WRITING TEST

AR-G/2021/5

Ute Knoch
Jason Fan
Cathie Elder
Ksenia Zhao
Andrew Pitman
Language Testing Research Centre,
University of Melbourne

ARAGs RESEARCH REPORTS ONLINE

ISSN 2057-5203

The British Council is the
United Kingdom's international organisation
for cultural relations
and educational opportunities.

ISSN 2057-5203
9 772057 520008 >

www.britishcouncil.org/aptis/research