THE UNIVERSITY OF MELBOURNE

A comparison of the discourse produced at different score levels of the OET writing sub-test

Final report

Sally O'Hagan, Ute Knoch, Ali Rastgou, Woranon Sitajalabhorn, Catriona Fraser, Cathleen Benevento

November, 2013

Table of contents

Executive summary

The report describes a small-scale study designed to investigate the discourse produced by OET test takers in response to the writing task. In particular, the aim was to compare the discourse produced across different OET score bands and across two writing tasks from the medicine-specific writing task. We have focussed on this profession only because it was not feasible to include data from more than one profession into the study. The study was designed to provide validity evidence for the OET writing task and to feed into future rater training and further refinements to the task specifications and the rating scale.

The report begins by providing a detailed literature review on validity, related research using discourse analyses on writing samples written in response to other tests and a review of relevant discourse-analytic measures.

The writing samples used in the analysis were drawn from two operational administrations of the OET – March 2011 and November 2012. A total of 166 test performances were analysed, where available 25 performances at each score level for each task. The data was analysed using a range of discourse-analytic measures focussing on the following categories: accuracy, fluency, complexity (both lexical and syntactic), content, structure, coherence, and cohesion. All variables apart from structure were analysed quantitatively and were subjected to a MANOVA analysis and follow-up ANOVAs.

The results showed that the only measures successful in distinguishing between score levels were related to accuracy (i.e. percentage of error-free clauses and percentage error-free t-units). For these two variables, the biggest difference was between the Level B and C score levels. Most other variables did not result in a consistent trend across score levels. If they did (e.g. clauses per t-units, proportion of irrelevant idea units), they were not found to be statistically significant. Based on the findings, it can be argued that the OET writing raters

mostly base their decision on the perceived accuracy of the writing sample rather than focussing on broader features when assigning their scores.

The results also showed striking differences between the discourse produced in response to the two tasks. While any conclusions need to be tentative as we did not have a fully matched design with each test taker writing in response to both tasks, the analysis showed that the discourse produced differed markedly assuming that the test takers were of approximately equal proficiency. These differences can be explained by a review of the two writing prompts used in the study. The two tasks differed substantially in terms of the amount of information provided in the input material.

Based on the findings of the study, we make the following recommendations to the OET Centre:

1. We recommend a careful review of the rater training procedures to ensure that raters apply the analytic rating scale in line with expectations and understand that a focus merely on the accuracy of the writing sample is not capturing all expected features of the writing.
2. We recommend that raters who are regularly found to be overfitting as part of the analysis of the operational assessment data are reminded to rate 'analytically' to avoid the exhibition of a halo effect.
3. We recommend that as part of the rater training procedures, raters are shown some of the findings in relation to the structure of referral/discharge letters identified as part of this study.
4. We recommend that the specifications for the writing sub-test are carefully reviewed to ensure that the tasks are approximately equal in terms of the amount of case notes provided and the level of cognitive difficulty.

We have provided annotated samples of writing scripts at the end of the report. These samples might be useful for future rater training or standardisation sessions.

Introduction

While previous studies on the Occupational English Test (OET) have investigated rater behaviour and decision-making on both the speaking and writing tasks (e.g. Knoch, 2011; Iwashita & Grove, 2003; Lumley, 1995), no studies to date have investigated the ways in which test taker discourse rated at different OET score levels varies in terms of quality of the discourse. This type of study is important since it forms part of a validation argument for the OET and the results can be used to support or suggest refinements to the rating scale descriptors, and may also play a role in rater training procedures.

The present study analyses writing samples produced by OET test takers in live administrations, performing detailed discourse analyses of the samples. A total of 166 sample performances are being analysed at a range of score levels including 50 performances at each of OET Grades B and C, which encompass the pass/fail boundary (cut-score) for most professional accreditation bodies that recognise the OET. A range of discourse-analytic measures are being employed which represent constructs of interest including, but not limited to, accuracy, fluency, and complexity measures, as well as measures specifically relevant to medical discourse. For the purposes of the study, the sample performances have been grouped into proficiency levels (as determined by the grades based on the OET raters' scores) and statistical tests comparing the various levels on each of the discourse analysis measures will be performed to establish the significance of differences between the grade levels. One of the outcomes of this project will be to identify benchmark performances at different grade levels which can be used for rater training or be displayed as sample performances, for example on the OET website.

Literature review

The review of the literature will be divided into three main sections. The first section will introduce the concept of validity argumentation as an approach to test validation and will explain how this study contributes validation evidence for the OET within such a validity framework. The next section will review some key discourse studies of test taker writing. The final section is dedicated to the discourse-analytic variables that can be used to analyse writing samples such as those used in this study, and will consider the suitability of specific measures for the present study.

Validity

Current approaches to validity in language testing are not limited to concerns about qualities of the test, but rather, extend to the context of test use including interpretation and uses of test scores. Such approaches build on Messick's (1989; 1995) (re)conceptualisation of validity as a unified concept. But the conception of validity as an *argument*, as set out by Kane (2006) and others (e.g. Bachman, 2005) extends the notion of test validation beyond that inherent in the unitary concept. Within an interpretive argument framework (e.g. Chapelle, Enright and Jamieson, 2008), an argument-based approach to validity requires that "the reasoning inherent in the proposed interpretations and uses [be made] *explicit* [emphasis added]" (Kane, 2012: 35). A number of processes contribute to making this 'reasoning' plain, including the provision of empirical evidence to provide backing for the various inferences in a validity argument, which can be conceived as a 'chain' of inferences in which each successive link in the chain must in turn be backed by evidence for any dependent interpretations to be considered valid (Chapelle, Enright and Jamieson, 2008).

The construction of validation arguments based on principals as set out in Kane (2006; 2012) has become integral to current validation practices and is particularly useful in the context of high stakes tests, such as the OET. By providing discourse-based evidence of

qualitative differences in test taker discourse across different score levels, this study provides empirical backing to support some of the inferences in the validity argument for the OET. The first inference in the validity argument chain most relevant to this study is the 'evaluation' inference. According to Chapelle et al., evaluation is "based on the warrant that observations of performance on test tasks are evaluated to provide observed scores reflective of targeted language abilities" (2008, p. 15). This study offers support to the warrant underlying this inference in that an analysis of test discourse can be used to validate or develop scoring rubrics that reflect the relevant knowledge and skills used by the test takers (Enright, et al., 2008; Xi, 2008).

A further relevant inference in the validity argument is 'explanation' (Chapelle et al., 2008). This inference rests on the assumption that test tasks engage language abilities similar to those underlying real world tasks in the relevant domain (Xi, 2008), and that scores are therefore based on the writing construct that the test task has been designed to measure. In relation to the OET, by analysing test taker discourse for evidence of variation across score levels in domain-relevant and domain-specific discourse-analytic measures, we aim to provide backing for the warrant underlying this assumption.

Research comparing writing ability levels

Research comparing test takers discourse at different writing score levels has become increasingly popular in recent years. These studies have often been conducted as part of the ongoing validity investigations of major testing companies (see for example, Cumming et al., 2006; Banerjee et al., 2007; Knoch, Macqueen & O'Hagan, forthcoming) or to use as a basis for the development of empirically-based rating scales (e.g. Knoch, 2009). While the results of the studies as well as the exact discourse measures employed have varied, they have generally found that students at higher writing ability levels write longer essays (although

there is often a ceiling effect), use more complex vocabulary and grammar (although the latter has had very mixed findings), and display a higher level of coherence and cohesion.

Research on features of essays written based on source material is still scarce, although a number of studies have been published in recent years. Gebril and Plakans (2013), for example compared the responses to an integrated argumentative essay and found that as the writing quality of the essays increased, so did the number of words. Measures of grammatical and lexical complexity were less successful in differentiating between score levels. Lower level learners were less able to transform the source material (e.g. by using paraphrasing strategies) than higher level learners.

Discourse measures

Accuracy

Accuracy is defined by Wolfe-Quintero, Inagaki, and Kim (1998) as "the ability to be free from errors while using language to communicate in either writing or speech" (p. 33). Attempts have been made to measure accuracy in writing, and a range of accuracy measures have been invented. These measures can be divided into four broad types: holistic rating scales, error-free measures, error-frequency measures, and error severity measures.

The first approach to measuring accuracy in writing, ***holistic rating scales***, have been employed by Cumming, Kantor, Baba, Eouanzoui, Erdosy, and James (2006), Cumming, Kantor, Baba, Erdosy, Eouanzoui, and James (2005), and Gebril and Plakans (2009) to investigate accuracy in compositions at different levels of writing proficiency. Their research results yielded significant differences across the proficiency levels; that is, compositions produced by more proficient writers were grammatically more accurate than those written by less proficient writers. Polio (1997) also used a holistic rating scale to gauge accuracy of compositions and reported that this method was much more time-saving than using error-free and error-frequency measures. However, she found that the inter-rater reliability was low

because the scale consisted of several vague descriptors, resulting in raters being forced to make subjective decisions.

Researchers desiring to evaluate accuracy more objectively have opted for the other three types of accuracy measures. These measures involve segmenting compositions into structural units (e.g., clauses and T-units). Error-free measures, such as error-free clauses, error-free T-units, percentage of error-free clauses, and percentage of error-free T-units, are basically attempts to count the number or to calculate the proportion of error-free units in compositions. ***The number of error-free clauses*** and ***the number of error-free T-units*** are problematic in that they can be affected by essay length. As a result, the percentage of error-free clauses (the total number of error-free clauses divided by the total number of clauses), and the percentage of error-free T-units (the total number of error-free T-units divided by the total number of T-units) have been widely utilised by researchers.

Using ***the percentage of error-free clauses***, Ishikawa (1995) found that the essays composed by one of the groups of novice learners were significantly more accurate after three months of instruction. As for ***the percentage of error-free T-units***, this measure has been found to relate significantly to test scores (Arnaud, 1992; Hirano, 1991; Vann, 1979), program levels (Hirano, 1991; Larsen-Freeman, 1978; Larsen-Freeman & Strom, 1977), and grades (Kawata, 1992; Tomita, 1990). Likewise, Knoch (2009) has found a significant relationship between the percentage of error-free T-units and proficiency measured by test scores of a diagnostic writing test.

Defining the term ‘error’ is a thorny issue of error-free measures. For example, Wolfe-Quintero et al. (1998) maintain that different researchers might have different views on what constitutes an error; thus, some degree of subjectivity still exists. Error-free measures have also been criticised by Bardovi-Harlig and Bofman (1989) because they treat a unit with

one error and a unit with many errors similarly. Moreover, the measures treat all types of error the same way; they disregard the fact that some errors are more severe than others.

To avoid some of the aforementioned drawbacks of error-free measures, researchers have turned to error-frequency measures which involve identifying errors in written texts and counting them. Like the number of error-free clauses and the number of error-free T-units, ***the number of errors*** has not been popularly used due to its sensitivity to text length. To eliminate the interference of text length, researchers have instead opted for ***errors per clause*** (the total number of errors divided by the total number of clauses) and ***errors per T-unit*** (the total number of errors divided by the total number of T-units). Fisher (1984) using errors per clause and Flahive and Snow (1980), Perkins (1980), Perkins and Leahy (1980) all using errors per T-unit found that these measures related significantly with proficiency as measured by holistic ratings of essays. However, errors per T-unit have been found to have no significant relationship with grades in a course (Perkins & Leahy, 1980) and proficiency levels (Flahive & Snow, 1980). No recent studies have been conducted that employed errors per clause and errors per T-unit to explore compositions at different levels of writing proficiency. Errors-frequency measures are not entirely problem-free, though. Whilst they take the number of errors in each unit into account, the problems of defining an error and the different degrees of severity of each error are still present.

The last type of measure of accuracy has recently been proposed by Wigglesworth and Foster (2008). This accuracy measure takes the issue of ***error severity*** into consideration. The two researchers argue that errors have different degrees of wrongness from minor errors that hardly affect the comprehensibility to more severe errors that render a piece of writing nearly incomprehensible. Wigglesworth and Foster (2008) suggested using clauses as the unit of analysis and scoring each clause according to the severity of errors it contains. They trialed this measure with three writing scripts of different levels of writing proficiency and found

that the measure can effectively capture small developments in accuracy. However, this accuracy measure has not yet been employed in a study that investigates essays at different writing proficiency levels.

Fluency

Researchers have defined fluency in several different ways, and most of its existing definitions are based on oral rather than written production of language. One of its definitions that cover both forms of language production is given by Housen and Kuiken (2009) who define fluency as "a person's general language proficiency, particularly as characterized by perceptions of ease, eloquence, and 'smoothness' of speech or writing" (p. 463).

As for measures of fluency, Wolfe-Quintero et al. (1998) maintain that most fluency measures have been invented to gauge fluency in oral discourse. For example, Skehan (2003) argues that fluency is a multifaceted construct including four different aspects, i.e. breakdown fluency, repair fluency, speech rate (temporal aspects of fluency), and automatisation. He further proposes that each of these facets needs a unique measure; for instance, breakdown fluency can be measured by silence; repair fluency by replacement, repetition, reformation, and false starts; speech rate by words/syllables per minute; and automatisation by length of run.

With regard to measures of fluency in written discourse, Knoch (2009) argues that although it is possible to measure all the aspects of fluency suggested by Skehan (2003) in the context of writing, it will require several tools, such as videotapes and sophisticated computer programs (see Chenowith and Hayes's (2001) study for example). Knoch (2009) then further contends that on the basis of written products alone repair fluency and writing rate (or the temporal aspects of fluency) are the two components of fluency that seem practical to measure. To measure repair fluency, she counted ***the number of self-corrections*** (instances of insertions and deletions made by test takers) in diagnostic writing essays at different levels

of writing proficiency. Although she found that the number of self-corrections decreased as the proficiency increased, the differences between the pairs of adjacent levels were not statistically significant.

To measure the temporal aspects of fluency, Wolfe-Quintero et al. (1998) suggested counting the number of words and structural units produced by a writer within a limited time. ***The number of words*** has been widely used by several researchers to index fluency, and these studies have produced mixed results. Investigating Test of Written English (TWE) compositions at different levels of writing ability, Grant and Ginther (2000) found an increase in the number of words as the proficiency level became higher. However, whether the differences in the number of words were statistically significant was not reported. In a similar study investigating International English Language Testing System (IELTS) compositions by Kennedy and Thorp (2002), it was found that compositions at higher level contained more words than those at lower level. Nevertheless, the minimum and maximum number of words of the essays at each level considerably overlapped although the levels of the essays investigated in this study were not adjacent. According to Knoch (2009), this might indicate that the differences in the number of words of essays at higher levels might not be significant. Cumming et al. (2006) compared the number of words of essays at levels 3, 4, and 5 written for the field test of the new Test of English as a Foreign Language (TOEFL) writing tasks and found an increase in the number of words when the proficiency level increases. Nonetheless, the test of significance showed that the differences occurred only between levels 3 and 4 and levels 3 and 5 essays but not between levels 4 and 5 essays. Similarly, Knoch (2009) analysed diagnostic essays at levels 4, 5, 6, 7, and 8 and found an overall increase in the number of words as the proficiency level became higher. Even so, there was much overlap between the minimum and maximum number of words of compositions at each level. Additionally, the essays at levels 6, 7, and 8 contained very similar numbers of words,

showing that there might be a ceiling effect. On the contrary, Gebril and Plakan's (2009) study yielded the results that showed that the number of words was an effective measure of the writing rate. They found significant differences in the number of words in integrated reading-writing compositions at three writing proficiency levels. What is more, the differences in the mean number of words of all level pairs (e.g., 1 and 2, 1 and 3, and 2 and 3) were statistically significant.

Another way to tap into the temporal aspects of fluency is to rely on the number of structural units produced in written texts (Wolfe-Quinter et al., 1998). Only two researchers, Ishikawa (1995) and Kameen (1979), used ***the number of clauses*** to measure writing rate, but neither of them achieved significant results. ***The number of T-units*** has been utilised by various researchers (Hirano, 1991; Homburg, 1984; Ishikawa, 1995; Kameen; 1979; Kawata, 1992; Perkins, 1980; Tedick, 1990; Tomita, 1990), yet the measure produced a significant result only in one of the two analyses by Ishikawa (1995).

Complexity

Ellis and Barkhuizen (2005) classified complexity into several sub-categories. However, only two sub-categories, syntactic complexity and lexical complexity, are common in writing research (Wolfe-Quintero et al., 1998). Skehan (2009) also maintains that these two sub-categories of complexity are distinct areas of performance that should be discussed and measured separately.

Syntactic Complexity

Ortega (2003) defines syntactic complexity as "the range of forms that surface in language production and the degree of sophistication of such forms" (p. 492). As can be seen from its definition, syntactic complexity does not concern the number of structural units produced in a piece of writing but the grammatical/structural variation and sophistication of those units.

Based on their extensive review of the literature on syntactic complexity, Norris and Ortega (2009) argue that syntactic complexity is a multidimensional construct including complexification at global, phrasal, and clausal levels, and in order to reveal a complete picture of syntactic complexity in a piece of writing every level of complexification should be measured. They further assert that although there are several measures for global and clausal complexity, using only one measure is sufficient. Following their recommendation, this study will use the number of words per T-unit and the number of clauses per T-unit to measure complexification at global and causal levels respectively. As for phrasal complexity, Norris and Ortega (2009) argue that the number of words per clause is the only measure that can index complexification at phrasal level.

The number of words per T-unit, according to Ortega (2003) and Wolfe-Quintero et al. (1998), is one of the common measures of syntactic complexity utilised in second language (L2) writing research. It has been found to relate significantly to program placement (Hirano, 1991; Ho-Peng, 1983; Larsen-Freeman, 1978, 1983; Tedick, 1990), school levels (Cooper, 1976; Gipps & Ewen, 1974; Henry, 1996; Monroe, 1975; Tomita, 1990; Yau, 1991), and standardised test scores (Hirano, 1991). In two more recent studies by Cumming et al. (2006) and Gebril and Plakans (2009), this measure produced contrasting results in that Cumming et al. (2006) found significant differences in the number of words per T-unit in the TOEFL field test essays at different levels of writing ability whereas Gebril and Plakans (2009) did not found the same significant results from analysing reading-to-write compositions.

According to Wolfe-Quintero et al. (1998), ***the number of clauses per T-unit*** is one of the most promising measures despite the mixed findings of previous research. This measure has been found to have a significant relationship with program levels (Hirano, 1991), and school levels (Cooper, 1976; Monroe, 1975). However, this measure failed to

discriminate between writing scripts at different levels of writing proficiency in Cumming et al.'s (2006), Banerjee and Franceschina's (2006) and Knoch's (2009) studies.

The number of words per clause has not been employed much compared to the other two measures above, and the studies employing this measure yielded inconsistent results. A significant relationship has been found between this measure and non-adjacent program levels (Hirano, 1991), holistic ratings (Kameen, 1979), and non-adjacent school levels (Cooper, 1976; Monroe, 1975; Yau, 1991). On the other hand, it has failed to exhibit a significant relationship with test scores (Hirano, 1991; Kameen, 1979). Unfortunately, this measure has never been employed in a study that investigates compositions at different writing proficiency levels.

Lexical Complexity

Lexical complexity is defined by Wolfe-Quintero et al. (1998) as "[t]he richness of a writer's lexicon" or more specifically "the range (lexical variation) and size (lexical sophistication) of a … writer's productive vocabulary" (p. 101). Similar to syntactic complexity, lexical complexity focuses more on how varied and sophisticated the words or word types in a piece of writing are rather than on how many they are included in a text.

Type-token ratio (the total number of different word types divided by the total number of words) is probably the most well-known measure of lexical complexity which has long been used to index lexical density. However, it has been criticised for being heavily dependent on the length of writing scripts. Often time, according to Ellis and Barkhuizen (2005), research employing this measure has found that a shorter text displayed higher type-token ratio than a longer one does.

In response to the problem of type-token ratio, Malvern and Richards (1997, 2002) and Malvern, Richards, Chipere, and Durán (2004) have proposed the ***d-value*** as a new

measure of lexical richness that is no longer sensitive to text length. The calculation of the *d*-value is based on the assumption that text of different length has different type/token ratios; consequently, each text should be represented by a set of type/token ratios rather than just one ratio value. Text samples of different sizes are analysed to produce a curve of a set of type/token ratios, and these values are then be computed into a single value, the *d*-value.

The percentage of sophisticated lexical words (the total number of sophisticated lexical words divided by the total number of lexical words) is another measure of lexical complexity that has been used to tap into lexical sophistication. This measure involves identifying sophisticated lexical words, such as words on Coxhead's (2000) Academic Word List (AWL) and words that are not on Cobb's (2002) list of basic words. Employing this measure in her study, Knoch (2009) found that higher-scored essays contained more sophisticated lexical words than their lower-scored counterparts, and the difference in the number of sophisticated lexical words was also statistically significant.

A further measure of lexical complexity is ***the percentage of words from the Academic Word List***. Knoch (2009) analysed compositions written in the context of diagnostic writing assessment utilising this measure and found an increase in the percentage of words from the AWL as the proficiency level increased. This result, moreover, was statistically significant.

Another common measure of lexical complexity is ***average word length*** (the total number of characters divided by the total number of words). It has been used successfully by several researchers (e.g., Engber, 1995; Frase, Faletti, Ginther, & Grant, 1999; Grant & Ginther, 2000) to show that compositions at higher level of proficiency contained words that are overall longer than did those at lower level of proficiency. However, Cumming et al. (2006) and Gebril and Plakans (2009) did not obtain such results when employing this measure. In Knoch's (2009) study, average word length was found to successfully

differentiate between writing scripts at different levels of writing proficiency, and the differences in average word length was statistically significant.

Coherence

Measuring coherence objectively in texts has challenged researchers in both L1 and L2 writing research. One such coding scheme was proposed by Lautamatti (1987) in the form of topical structure analysis (TSA). Lautamatti proposed three types of progression which advance a discourse topic by developing a sequence of sentence topics, namely parallel progression, sequential progression and extended parallel progression. These are described in more detail below (see also Hoenisch, 1996):

- Parallel progression, in which the topics of successive sentences are the same, producing a repetition of topic that reinforces the idea for the reader (<a, b>, <a, c>, <a, d>);
- Sequential progression, in which the topics of successive sentences are always different, as the comment of one sentence becomes, or is used to create, the topic of the next (<a, b>, <b, c>, <c, d>); and
- Extended parallel progression, in which the first and the last topics of a piece of text are the same but are interrupted with some sequential progression (<a, b>, <b, c>, <a, d>).

Since this original work on topical structure analysis, which was conducted in the area of L1 writing, researchers have used this method in research on second/foreign language writing (see e.g. Schneider and Connor, 1990; Wu, 1997) and added additional categories. Knoch (2007a; 2007b) further expanded the method (to seven categories) and applied it to essays at several different writing ability levels and transferred the findings of her study into a rating scale.

Unfortunately, however, using seven categories makes the measure unwieldy and difficult to apply to large numbers of essays. For that reason Knoch et al. (forthcoming)

simplified the measure to only include two categories, coherent t-units and coherence breaks. In Knoch et al.'s forthcoming study on TOEFL iBT essays, this measure was able to differentiate between writers at different proficiency levels (albeit with a small effect size).

Cohesion

Halliday and Hasan's (1976) categories of cohesion have been applied in a number of research projects with varying results. Witte and Faigley (1981) in the context of L1 English, for example, compared the cohesion of high and low level essays. They found a higher density of cohesive ties in high-level essays. Almost a third of all words in the higher-level essays contributed to cohesion and the cohesive ties spanned shorter distances than in lower-level essays. They also found that the majority of lexical ties in low-level essays involved repetition, whilst high-level essays relied more on lexical collocation. In contrast, Neuner (1987) found that none of the ties were used more in good essays than in poor quality freshman essays. He did, however, find a difference between cohesive chains (three or more cohesive ties that refer to each other), in the cohesive distance and in the variety of word types and maturity of word choice. For example, in good quality essays, cohesive chains are sustained over greater distances and involve greater proportions of the whole text. Good writers also used a greater variety of words in their cohesive chains as well as less frequent words than the poor writers. A very similar result was found by Crowhurst (1987), who compared cohesion at different grade levels in two different genres (arguments and narratives). He also found that the overall frequency of cohesive ties did not increase with grade level, but that synonyms and collocations (a sign of more mature vocabulary) did.

Jafarpur (1991) applied Halliday and Hasan's categories to ESL writing. He found in essays that the number of cohesive ties and the number of different types of cohesion successfully discriminated between different proficiency levels. Reid (1992), investigating ESL and NS writing, focussed on the percentages of coordinate conjunctions, subordinate

conjunctions, prepositions and pronouns and found that ESL writers used more pronouns and coordinating conjunctions than NS, but fewer prepositions and subordinating conjunctions. Two other studies also compared native and non-native speaking writers in terms of their use of connectors. Field and Yip (1992) were able to show that Cantonese writers significantly overuse such devices. However, Granger and Tyson (1996) in a large-scale investigation of the International Corpus of Learner English, were not able to confirm these findings. They emphasised that a qualitative analysis of the connectors is important. They documented the underuse of some connectors and overuse of others.

Some more recent studies compared the performances of test takers over different proficiency levels. Firstly, Kennedy and Thorp (2002), in the context of IELTS, were able to show that writers at levels 4 and 6 used markers like 'however, firstly, secondly' and subordinators more than writers at level 8. They concluded that writers at level 8 seemed to have other means at their disposal to mark these connections, whilst lower-level writers needed to rely on these overt lexico-grammatical markers to structure their argument. Also in the context of IELTS, Banerjee and Franceschina (2006) looked at the use of demonstrative reference over five different IELTS levels. They found that the use of 'this' and 'these' increased with proficiency level whilst the use of 'that' and 'those' stayed relatively level or decreased. Knoch's (2007a) analysis also showed that test takers at higher proficiency levels made use of 'this' and 'these' more frequently than their counterparts at lower proficiency levels. Knoch et al.'s recent study (forthcoming) investigated anaphoric reference in TOEFL iBT essays and found little variation in the density of anaphoric pronouns across different score levels.

Content

Quantifying content in essays is difficult and few objective methods have been proposed for independent tasks (i.e. writing tasks not directly drawing on input material). Those studies

that tried to operationalize content counted the number of details (Friedlander, 1990), the number of higher level propositions (Kepner, 1991), the number of main and supporting ideas (Knoch et al., forthcoming) and the number of idea units (Polio, 2001). Most other studies used rating scales to rate the content.

More work has been done on quantifying the content used in integrated tasks, (i.e. tasks that draw on input material in the form of listening or reading material or both). Much of the early work in this area was done in the area of L1 writing focusing on summarization tasks and we will now turn to reviewing some of that research.

To judge the quality of summaries, it is important to first evaluate the importance of the information in the source text. To achieve this, most researchers turn to the concept of idea units (sometimes referred to as propositions, content units, content idea units, linguistic subunits, pause acceptability units, or pasual units). The methods employed to determine idea units differ from study to study. In Kintsch and van Dijk's (1978) work, for example, an idea unit is the unit consisting of one predicate and its argument(s). Similarly, Coffman's (1994) idea units are "propositions that contained a subject and predicate combination plus restrictive clauses. Compound predicates were divided into separate propositions" (p. 26).

Once idea units are determined, their importance to the theme of the original text is further rated to distinguish the gist from trivia. Here the methods have also differed. Some researchers e.g. Brown et al. (1983), Brown and Smiley (1977), and Johnson (1970) had each idea unit retyped on a separate line and instructed a group of educated native speakers to remove a certain percentage of idea units that they considered least important to the theme of the passage. The remaining units were deemed the most important to the theme of the text.

Another alternative is to have native speaker experts write a model summary of the source text (e.g. Corbeil, 2000; Rivard, 2001; Yu, 2007, 2008, 2010). Although this approach

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

We were not able to find any studies that conducted similar kinds of research on summaries based on notes, or medical discharge or referral letters based on patient case notes. However, some studies reported in the literature on medical communication have audited authentic referral letters for quality of content in terms of domain expert judgments of accuracy, clarity and coverage of expected information (e.g. Newton et al., 1994; Burbach & Harding, 1997; Linné & Rössner, 1998; Moselhy & Salem, 2009), or for adherence to clinical guidelines (Shaw et al., 2005) or quality assurance standards (Al-Alfi et al., 2007). Tattersall et al. (1995) surveyed *referring* doctors on their preferences for content in the letters they receive from consultant physicians.

Structure

Different approaches have been taken to the study of *schemata*, or patterns of macro-structural elements in written discourse. Some of this has been undertaken in the context of specialized academic and professional domains and includes research on the rhetorical structures of scientific discourse (e.g. Widdowson, 1979), the schematic structures of news reportage (van Dijk 1988), or the generic 'moves' in research article introductions (Swales, 1990) and legal and business discourse (e.g. Bhatia, 1993). The various analytic frameworks that have been employed to study levels of organisation, or macro-level structures of discourse, in part reflects theoretical shifts in conceptualizations of 'discourse' (see Bhatia, 2004). Further, the specificity of discourse in its given context means that tailored coding schemes are required.

In health communication contexts, there has been some work devoted to the analysis of written medical discourse (see Harvey & Koteyko, 2013). This work has focussed on the written patient record (or case, or progress notes) and has examined its lexico-grammatical features (e.g. Anspach, 1988; Hobbs, 2003) and conventional format of SOAP (Subjective, Objective, Assessment, Plan) (see Hobbs, 2003). We are not aware of any similar studies of the macro-structure of medical discharge or referral letters based on patient case notes.

Methodology

The OET writing sub-test

The OET writing tasks is designed to assess a test takers' ability to produce written English relevant to their respective health profession. The task typically consists of a set of case notes, with test takers required to produce a letter of referral to another health professional. The candidate is required to synthesize the information in the case notes in approximately 180-200 words. The letter needs to be set out in an appropriate format.

Each of the scripts is routinely marked by two raters using an analytic scale with five criteria (overall task fulfilment, appropriateness of language, comprehension of stimulus, linguistic features, presentation features) and band descriptors ranging from 1 to 6.

The data is then analysed using the multi-faceted Rasch software Facets. For writing scripts identified to be misfitting by the analysis, a third rater is called in to provide an additional rating. Following the statistical analysis in Facets, each writing scripts is awarded a 'fair score' which is a score taking into account rater severity. These fair scores are then converted into a grade ranging from A to E.

The study

The data for the study was supplied by the OET Centre. The writing scripts were selected from two test administrations, March 2011 and November 2012. Only writing samples provided by doctors were used for this study. 85 test takers completed Prompt 1 of the test and 81, Prompt 2. In total, 166 scripts were analysed for the study. The distribution across the OET grade levels was as set out in Table 1 below. Unfortunately, not many scripts were available at grade level D.

Grade	Prompt 1 – March 2011	Prompt 2 – November 2012
A	25	25
B	25	25
C	25	25
D	10	6
Total	*85*	*81*

Table 1: Distribution of writing samples across OET grade levels

Instruments

The two writing tasks used in this study can be seen in full in Appendix A. Prompt 1 is a letter from a general practitioner to a plastic surgeon at a hospital. Prompt 2 is a letter of referral from a general practitioner to an admitting officer at a hospital. As can be seen in Appendix A, the amount of information in the case notes varies.

Participants

As mentioned above, all participants were doctors. They were all candidates taking the OET to apply for registration to practice in Australia. Of the 166 participants, 62 had taken the OET once only, while 104 had taken the OET twice or more. 28 different first language backgrounds were represented amongst the participants, with the largest first language groups being Arabic (22.3% of participants), Bengali (17.5%), Farsi (13.9%) and Burmese (7.0%). Participants came from 30 different countries, the most common countries of origin being Bangladesh (15.7% of participants), Iran (12.7%) and Iraq (9.0%).

Analysis – data preparation

The scripts were provided to the LTRC in scanned format. To prepare the files for analysis, all scripts were typed up and saved in text files. For the purposes of the lexical analysis, the data files underwent one further step of preparation. For this, all spelling mistakes were corrected under the assumption that if students knew a word but did not know how to spell it, the word should be included in the lexical analysis. Words that could not be recognised, were not corrected. Spacing after words and before and after punctuation marks was also corrected to prepare for automated analyses.

Analysis – coding

Table 2 below sets out the discourse-analytic measures used in this study. The exact coding procedures are set out in more detail below.

Category	Measure
Accuracy	Percentage error-free t-units Percentage error-free clauses
Fluency	Number of words Number of t-units Number of clauses
Syntactic complexity	Number of words per t-unit Number of clauses per t-unit Number of words per clause
Lexical complexity	D-value Average word length Lexical density Lexical sophistication Percentage words from the AWL
Coherence	Proportion of coherent t-units
Cohesion	Referential cohesion Number of connectives
Content	Proportion of required idea units Proportion of irrelevant idea units
Structure	Macro-structural elements

Table 2: Discourse-analytic measures used in the study

The coding of the ***proportion of error-free t-units and clauses*** involved manual coding of the clause and t-unit boundaries. For this, we used the coding scheme described in Cumming et

al. (2006). Following the coding of the boundaries, a researcher marked each error-free unit. We defined 'error-free t-units/clauses' as those which a highly proficient user of English would consider correct. Finally, the error-free units were divided by the total number of units to arrive at the proportion, which was expressed as a percentage. To ensure inter-coder reliability, a subset of 10% of scripts was double coded.

The ***number of words*** in each script were analysed using Microsoft Word. Before the analysis, abbreviations in the scripts (e.g.) were spelled out so that they could be captured by the system as words. The ***number of clauses*** and the ***number of t-units*** were coded by human coders following the guidelines in Cumming et al. (2006). Inter-coder reliability was calculated on a subset of 10% of scripts.

The measures of syntactic complexity could all be computed using the coding described above. For example, the ***number of words per t-unit*** were calculated by using the number of words and the number of t-units. Similar calculations were used for the ***number of clauses per t-unit*** and the ***number of words per clause***. No inter-coder reliability analysis was necessary as these measures were obtained from calculations as described.

The ***D-value*** for each script was calculated using the program d-tools (www.lognostics.co.uk). The details of the exact calculation of the d-value can be found in Meara and Miralpeix (n.d.). No inter-coder reliability analysis was necessary as this measure was calculated automatically.

Average word length was calculated using the computer tool, Coh-Metrix (McNamara et al. 2005), which automatically analyses written texts and computes selected discourse measures including average word length. Before the analysis, all writing scripts were screened for medical abbreviations and symbols which were spelled out for the sake of the analysis because the automated program is not able to recognise these abbreviations and symbols.

Coh-Metrix calculates the number of letters (excluding any punctuation marks) and the number of words automatically and based on these figures calculates the average word length.

Lexical density was calculated using the computer program Vocab Profile (Cobb, 2002). Lexical density is calculated by dividing the content words in each script by the total number of words.

The measure of lexical sophistication was also calculated with help of the program VocabProfile (Cobb, 2002). Tokens from the Academic Word List and real off-list tokens were combined to make up the sophisticated lexical words. These were then divided by the number of content words to arrive at an index of lexical sophistication.

The ***percentage words from the academic wordlist*** was also calculated using VocabProfile (Cobb, 2002).

The ***proportion of coherent t-units*** was calculated using Knoch et al.'s (forthcoming) coding scheme. For this measure, each t-unit is coded as either being coherent, or not. The coding is based on topical structure analysis (see e.g. Schneider & Connor, 1990) as well as previous work by Knoch (2007a, 2007b). This is based on strict coding guidelines created for this purpose and examples of coding can be found in Knoch et al. (forthcoming). To ensure inter-coder reliability, a subset of 10% of scripts was double-coded.

Cohesion was established by measuring the density of tokens of two key aspects of discourse cohesion: co-referentiality and connectivity (Halliday & Hasan, 1976; Graesser & McNamara, 2011). Measures of these were obtained using the computer tool, Coh-Metrix (McNamara et al. 2005), which automatically analyses written texts and computes selected measures of cohesion. Referential cohesion was established by the incidence of repetition, or overlap, of the same nouns and pronouns between all sentences. Overlap was identified

regardless of whether the matching noun was in singular or plural form so that, for instance, *visit* and *visits* occurring across sentences would be identified as an incidence of overlap. Cohesion through connection words was established by the incidence of all connectives. This measure encompassed the following types of connectives: causal, logical, contrastive, temporal and additive.

Content of the scripts was measured by the proportion of information (idea units) from the case notes (Prompt 1 and Prompt 2) that is included. Three types of idea units in the case notes were identified by five clinician informants: ideas which were *required* to be included (essential) in the referral letter, *irrelevant* ideas which should not be included, and ideas which were *optional* for inclusion. Content was measured by counting the number of *required idea units* and number of *irrelevant idea units* included in each script. The scores for each type of idea unit were divided by the total possible number of idea units of the same type i.e. the total number identified in the case notes. To ensure inter-coder reliability, a subset of 10% of scripts was double-coded.

Structure was established by coding the scripts for the presence of expected macro-structural elements. The coded scripts were then analysed qualitatively to establish their adherence to expected macro-structural schemata, or patterns of structural elements. To identify the elements and the expected patterns of these in the discourse, we sought the views of five clinicians on their preferred macro-structure for referral letters appropriate to the trigger case notes, Prompt 1 and Prompt 2. The clinicians agreed on which macro-structural elements should be included in the referral letter, and these elements are widely represented in the medical communication literature (e.g. Burbach & Harding, 1997; Tattersall et al., 2002). Clinicians varied in their preferences, however, for the order in which some of these elements should appear in the letters. To accommodate these differences, the scripts were analysed for

adherence to more than one acceptable schematic variant. To ensure inter-coder reliability, a subset of 10% of scripts was double-coded.

Statistical analysis

The analysis was undertaken in two parts. First, all discourse-analytic variables, with the exception of the measure for structure (which was analysed qualitatively), were entered into a MANOVA analysis. Because of the significant MANOVA, these variables were subsequently analyzed separately using ANOVAs. If a main effect for grade level was found, post-hoc analyses based on the least significant difference were conducted to explicate where the exact differences could be found. Because the data did not fulfil the assumption of normal distribution for an ANOVA, we analysed the data using (distribution-free) permutation tests with 4999 random permutations (Efron & Tibshirani, 1998), and it is these p-values that are reported throughout the report. In addition, partial eta-squared values were computed to estimate effect sizes. Following Cohen (1988), the effect sizes were interpreted as follows:

Small = .02

Medium = .13

Large = .26

Results

The results section is divided into two parts: first, the inter-coder reliability statistics will be reported and after that, descriptive statistics and the results from the MANOVA and ANOVA analyses for each of the discourse-analytic measures will be reported.

Inter-coder reliability

Table 3 below presents the results from the inter-coder reliability analysis.

Measure	Type	Result
T-units	Spearman *rho*	.925
Clauses	Spearman *rho*	.931
Error-free t-units	Spearman *rho*	.907
Error-free clauses	Spearman *rho*	.919
Number of coherent t-units	Percentage agreement	85.05%
Number of required idea units	Spearman *rho*	.822
Number of irrelevant idea units	Spearman *rho*	.816
Macro-structural elements	Percentage agreement	83.4%

Table 3: Inter-coder reliability statistics

The reliability statistics for the measures of accuracy, fluency and content are above .90 and therefore the reliability of the coding can be considered excellent. For the measures of coherence and structure, percentage agreements of around 85% and 83%, respectively, were achieved indicating very good to excellent levels of coding reliability.

Statistical analyses

MANOVA

A MANOVA was performed to investigate discourse-analytic features across the two sets of scripts (Prompt 1 and Prompt 2) and the four different score levels. The two (prompts) by four (score level) repeated measures multivariate analysis of variance with Pillai's trace and the random permutation tests revealed a main effect for prompt, $V = 0.86$, $F(1, 166) = 48.20$, $p < .001$; a main effect for proficiency level, $V = 0.94$, $F(3, 166) = 3.60$, $p < .001$; and an interaction effect, $V = 0.74$, $F(4, 166) = 2.58$, $p < .001$. Because all main effects and the interaction effect were found to be significant, further analyses were necessary to ascertain where these differences are located. For this reason, a series of ANOVAs, again with p-values determined by permutation tests, were conducted, and are reported in detail in the following section.

ANOVAs

Accuracy

Percentage of error-free t-units

Table 4 below presents the descriptive statistics for the percentage of error-free t-units. This is followed by graphical representations in Figure 1.

	Score level	N	Mean	SD
Percentage of error-free t-units – Prompt 1				
	A	25	57.78	16.23
	B	25	58.79	14.07
	C	25	40.64	16.60
	D	10	37.92	14.60
Percentage of error-free t-units – Prompt 2				
	A	25	60.90	13.09
	B	25	58.31	15.57
	C	25	39.42	19.53
	D	6	24.97	12.31

Table 4: Descriptive statistics – Percentage of error-free t-units

It can be seen from the table above and Figure 1 below that test takers at the two higher score levels (A & B) produced a higher proportion of error-free t-units compared with the lower scoring test takers (C & D). There is some difference in this result across the two prompts, although the difference is not great: the discourse of higher scoring test takers (at grade level A) responding to Prompt 2 was slightly more accurate, while the reverse is true for the discourse of the lowest scoring group of test takers (grade level D) responding to Prompt 2.

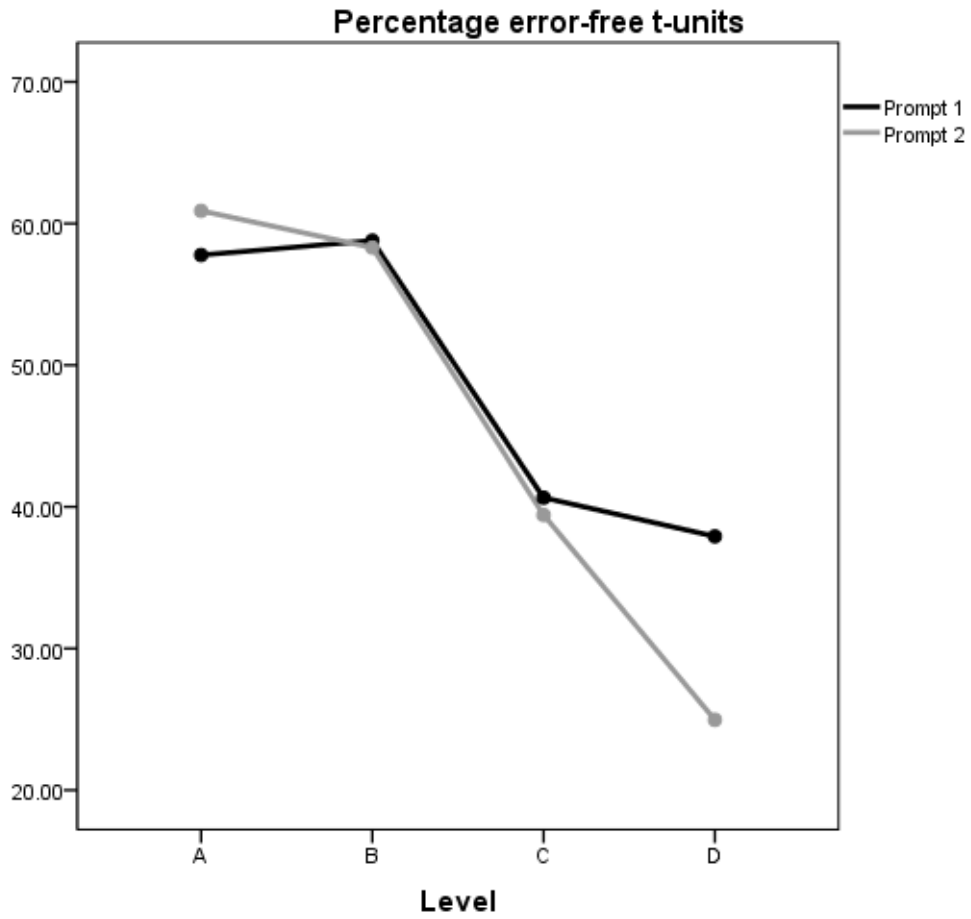

Figure 1: Percentage of error-free t-units

For the percentage of error-free t-units (EFTUs), the ANOVA showed a significant main effect and a large effect size for grade level [$F(3,166)=24.11$, $p<.001$, $\eta^2=.31$]. The main effect for prompt [$F(1,166)=1.052$, $p=.307$] and the interaction effect [$F(3,166)=1.00$, $p=.394$] were not significant. Post-hoc LSD tests were conducted to compare the different grade levels. These tests showed that for both prompts, discourse accuracy at levels A and B differed significantly from that at levels C and D. For both prompts these tests also showed there was no significant difference between the accuracy of discourse at levels A and B, and levels C and D.

Percentage of error-free clauses

Table 5 below presents the descriptive statistics for the percentage of error-free clauses. This is followed by graphical representations in Figure 2.

	Score level	N	Mean	SD
Percentage of error-free clauses – Prompt 1				
	A	25	66.58	14.45
	B	25	68.11	10.73
	C	25	51.32	14.44
	D	10	49.19	14.99
Percentage of error-free clauses – Prompt 2				
	A	25	69.11	10.81
	B	25	65.00	12.64
	C	25	46.14	18.20
	D	6	27.28	14.39

Table 5: Descriptive statistics – Percentage of error-free clauses

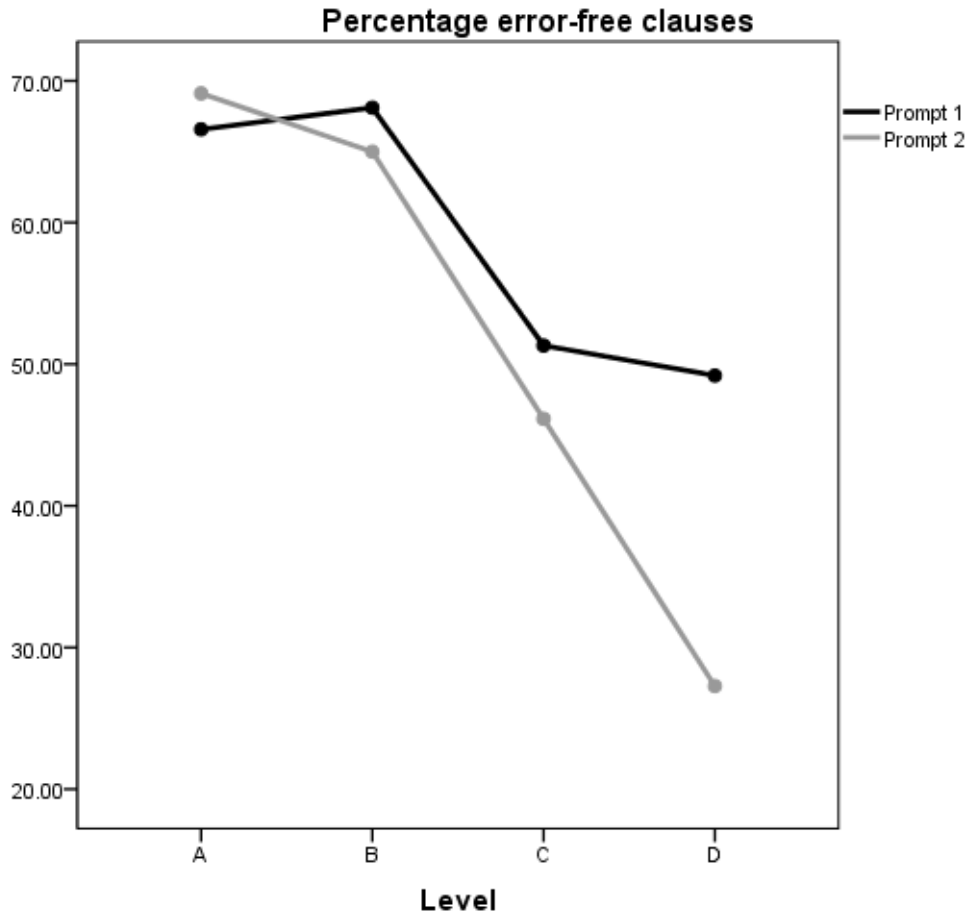

Figure 2: Percentage of error-free clauses

From the Table 5 and Figure 2 above, it can be seen that the two higher levels (A & B) consistently had a higher percentage of error-free clauses (EFCs) than the lower levels (C &

D). There is some difference in this result across the two prompts, the most pronounced difference being for discourse produced at the lowest level (D) which was less accurate at clause level for Prompt 2.

For the proportion of error-free clauses (EFCs), the ANOVA showed a significant main effect and a large effect size for grade level [$F(3,166)=32.021$, $p<.001$, $\eta^2=.31$]. It also showed a significant main effect for prompt with a small effect size [$F(1,166)=7.843$, $p=.006$, $\eta^2=.047$] and a significant interaction effect with a small to medium effect size [$F(3,166)=1.00$, $p=.394$, $\eta^2=.055$]. The post-hoc analysis based on the LSD showed that for Prompt 1, levels A and B each differ significantly from C and D. There was no significant difference found between A and B, or between C and D. For Prompt 2 the levels did not differ significantly.

Fluency

Number of words

Table 6 below presents the descriptive statistics for the number of words per script. This is followed by graphical representations in Figure 3.

	Score level	N	Mean	SD
Number of words – Prompt 1				
	A	25	205.32	25.07
	B	25	208.76	30.75
	C	25	199.56	25.89
	D	10	191.80	32.03
Number of words – Prompt 2				
	A	25	258.60	42.69
	B	25	232.96	31.68
	C	25	251.76	43.43
	D	6	206.50	80.60

Table 6: Descriptive statistics – Number of words

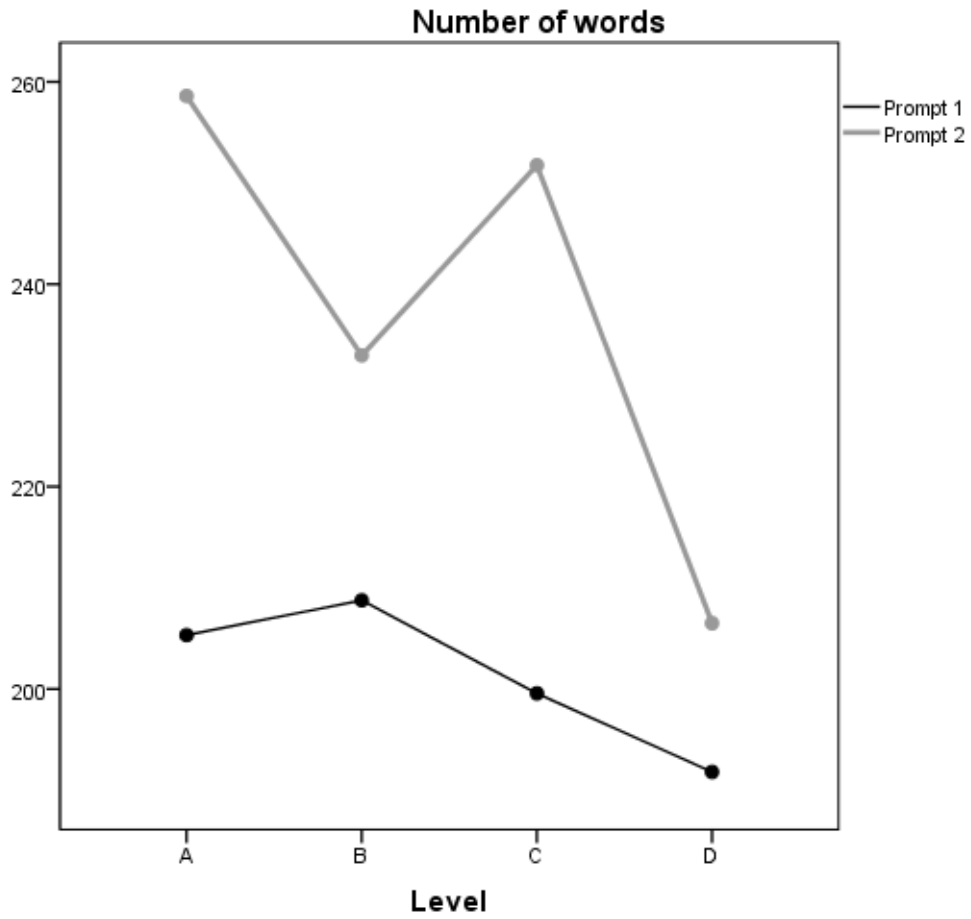

Figure 3: Number of words

From Table 6 and Figure 3 above, it can be seen that there were no consistent differences between the levels in terms of the number of words produced. Across the two prompts, a greater number of words were produced at levels A, B and C in response to Prompt 2.

The ANOVA showed a significant main effect for grade level, with a small/medium effect size [$F(3,166)=3.293$, $p=.022$, $\eta^2=.059$]. It also showed a significant main effect for prompt with a medium effect size [$F(1,166)=31.154$, $p<.001$, $\eta^2=.165$]. The interaction effect was not significant [$F(3,166)=2.415$, $p=.069$]. The post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1. For prompt 2 it showed that A differed significantly from B and D, and C differed significantly from D. B did not differ significantly from C or D.

Number of t-units

Table 7 below summarizes the descriptive statistics for the number of t-units per script. This is followed by graphical representations in Figure 4.

	Score level	N	Mean	SD
Number of t-units – Prompt 1				
	A	25	14.24	2.44
	B	25	15.52	2.43
	C	25	15.44	3.98
	D	10	14.10	3.90
Number of t-units – Prompt 2				
	A	25	19.72	4.12
	B	25	19.12	3.98
	C	25	19.60	4.15
	D	6	16.83	6.77

Table 7: Descriptive statistics – Number of t-units

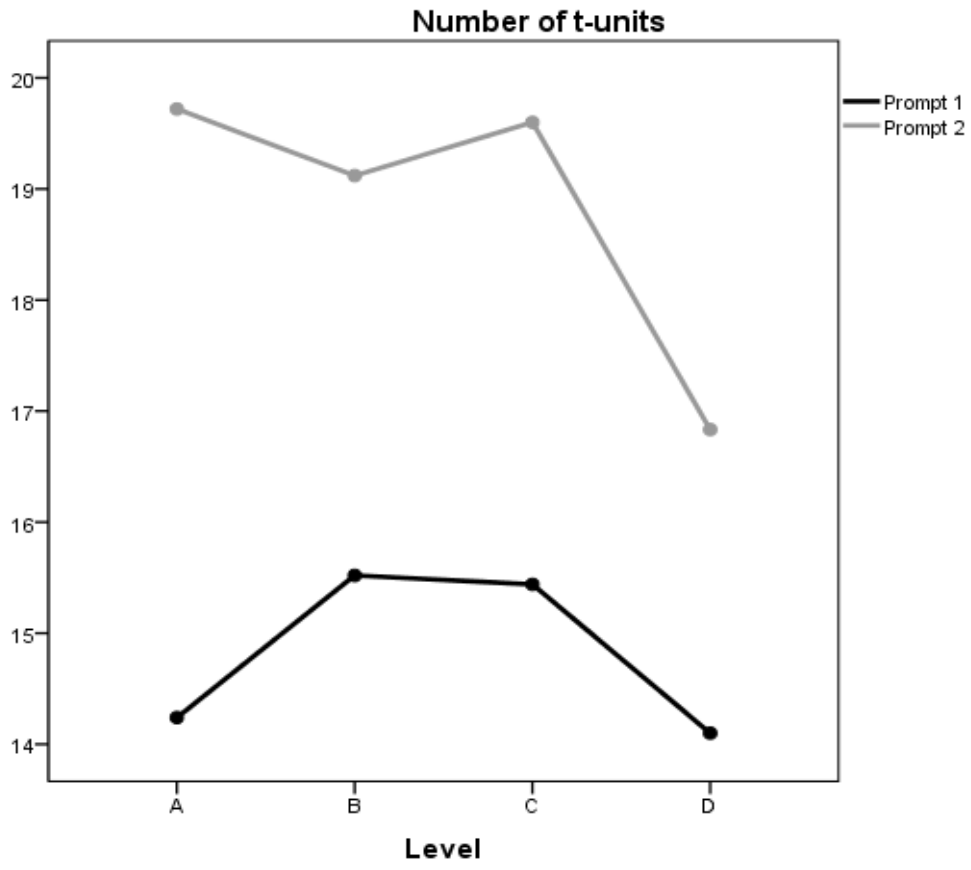

Figure 4: Number of t-units

From Table 8 and Figure 5 above, it can be seen that there were no consistent differences between the levels in terms of the number of t-units produced. Across the two prompts, a greater number of t-units were produced at all levels in response to Prompt 2.

The ANOVA showed no significant main effect for grade level [$F(3,166)=1.226$, $p=.302$] and no significant interaction effect [$F(3,166)=.777$, $p=.509$]. It showed a significant main effect for prompt with a medium to large effect size [$F(1,166)=35.642$, $p<.001$, $\eta^2=.184$]. The post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1. It also showed no significant differences between levels for Prompt 2.

Number of clauses

Table 8 below shows the descriptive statistics for the number of clauses per script. This is followed by graphical representations in Figure 5.

	Score level	**N**	**Mean**	**SD**
Number of clauses – Prompt 1				
	A	25	20.36	3.79
	B	25	22.12	3.89
	C	25	21.08	4.16
	D	10	19.20	5.27
Number of clauses – Prompt 2				
	A	25	26.04	5.18
	B	25	24.52	5.19
	C	25	24.84	6.00
	D	6	19.33	7.09

Table 8: Descriptive statistics – Number of clauses

From the table above and Figure 4 below, it can be seen that there were no consistent differences between the levels in terms of the number of clauses produced, with the exception of level D scripts which had fewer clauses. Across the two prompts, a greater number of words were produced at levels A, B and C in response to Prompt 2.

The ANOVA showed a significant main effect for grade level, with a small/medium effect size [$F(3,166)=3.127$, $p=.027$, $\eta^2=.056$]. It also showed a significant main effect for prompt with a small to medium effect size [$F(1,166)=12.637$, $p<.001$, $\eta^2=.074$]. The interaction effect was not significant [$F(3,166)=1.749$, $p=.159$]. The post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1. For Prompt 2 it showed significant differences between A and D; B and D; and C and D. A, B and C did not differ significantly.

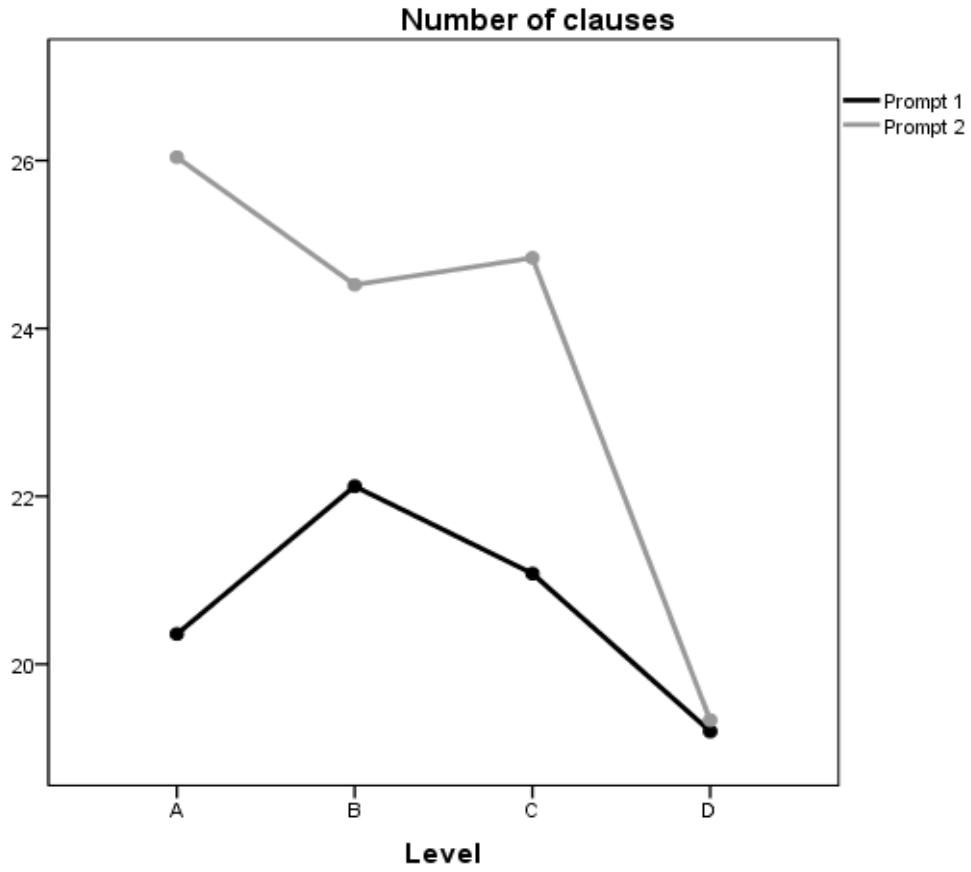

Figure 5: Number of clauses

Complexity

Syntactic complexity – Words per t-unit

Table 9 below shows the descriptive statistics for the measure, words per t-unit. This is followed by graphical representations in Figure 6.

	Score level	N	Mean	SD
Words per t-unit – Prompt 1				
	A	25	14.60	1.67
	B	25	13.62	2.02
	C	25	12.82	3.21
	D	10	14.31	3.68
Words per t-unit – Prompt 2				
	A	25	13.34	1.77
	B	25	12.62	2.83
	C	25	13.15	2.44
	D	6	12.60	2.34

Table 9: Descriptive statistics – Words per t-unit

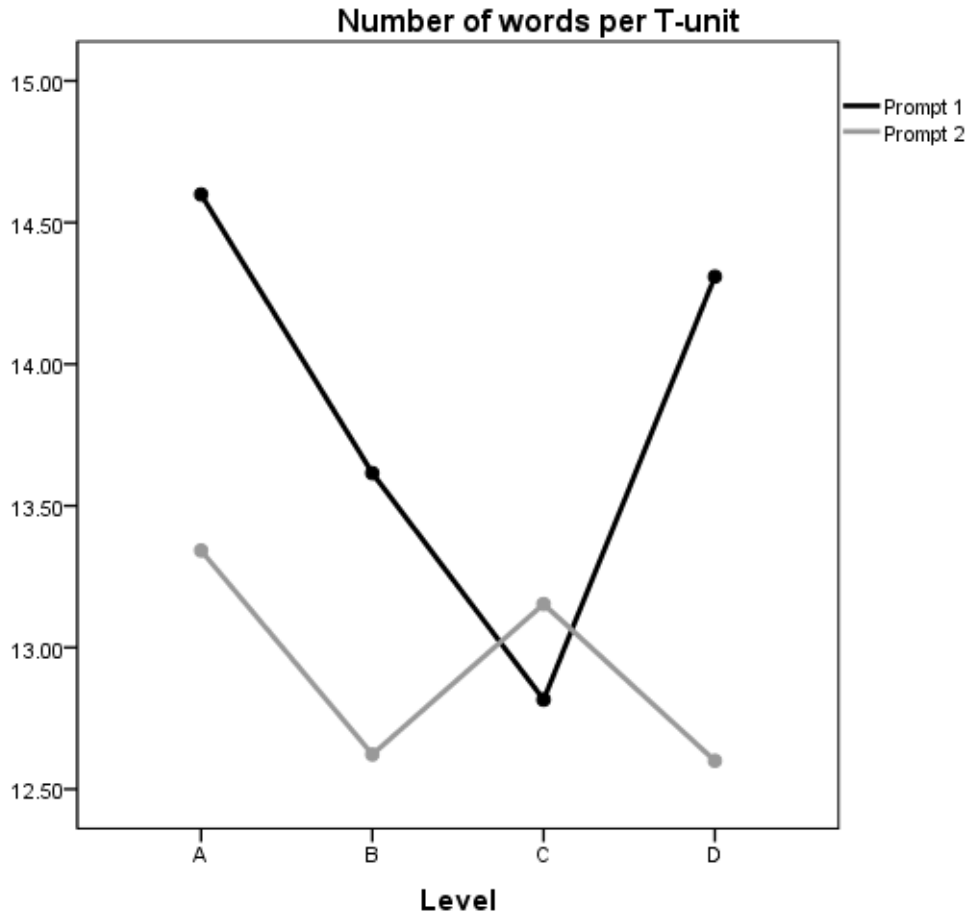

Figure 6: Words per t-unit

From the table above and Figure 6 below, it can be seen that there were no consistent differences between the levels in terms of the number of words per t-unit produced.

The ANOVA showed no significant main effect for grade level [F(3,166)=1.554, p=.203] and no significant interaction effect [F(3,166)=1.199, p=.312]. It showed a significant main effect for prompt with a small effect size [F(1,166)=4.201, p=.042, $\square^2$=.022]. The post-hoc analysis based on the LSD for Prompt 1 showed no significant differences between B and D, D and A, or B and A. A differed significantly from C. It showed no significant differences between levels for Prompt 2.

Syntactic complexity – Clauses per t-unit

The results for the second measure of syntactic complexity, clauses per t-unit, can be seen in Table 10 below. This is followed by graphical representations in Figure 7.

	Score level	**N**	**Mean**	**SD**
Clauses per t-unit – Prompt 1				
	A	25	1.43	.16
	B	25	1.43	.18
	C	25	1.40	.19
	D	10	1.37	.14
Clauses per t-unit – Prompt 2				
	A	25	1.33	.13
	B	25	1.29	.15
	C	25	1.28	.17
	D	6	1.16	.11

Table 10: Descriptive statistics – Clauses per t-unit

From the table above and Figure 7 below, it can be seen that there were no consistent differences between the levels in terms of the number of clauses per t-unit. Across prompts, a greater number of clauses per t-unit were found for Prompt 1 at all levels.

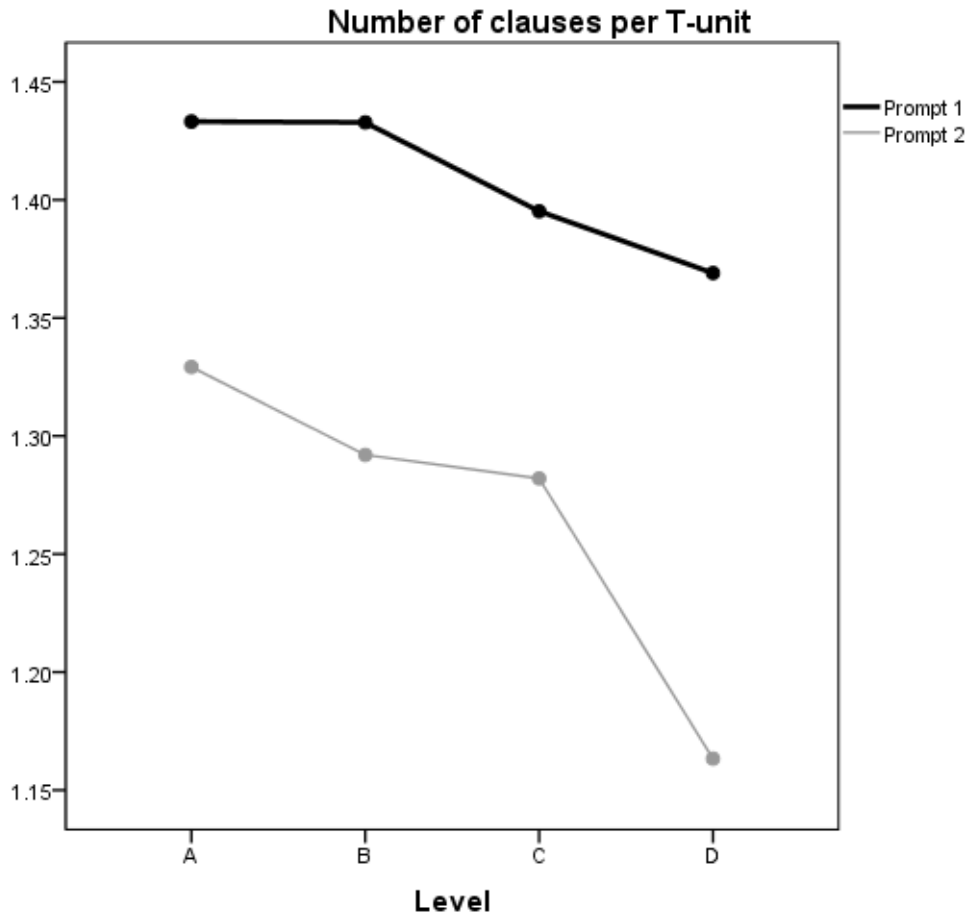

Figure 7: Clauses per t-unit

The ANOVA showed no significant main effect for grade level [F(3,166)=2.131, p=.098] and no significant interaction effect [F(3,166)=.440, p=.725]. It showed a significant main effect for prompt with a medium effect size [F(1,166)=23.930, p<.001, η^2=.132]. The post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1. For Prompt 2, A and D differed significantly. It did not show a significant difference between A and B or A and C; nor between B and C, B and D or C and D.

Syntactic complexity – Words per clause

Table 11 below shows the descriptive statistics for the measure, words per clause. This is followed by graphical representations in Figure 8.

	Score level	N	Mean	SD
Words per clause – Prompt 1				
	A	25	10.28	1.49
	B	25	9.55	1.13
	C	25	9.70	1.56
	D	10	10.48	2.63
Words per clause – Prompt 2				
	A	25	10.04	1.06
	B	25	10.04	2.19
	C	25	10.24	1.22
	D	6	10.83	1.74

Table 11: Descriptive statistics – Words per clause

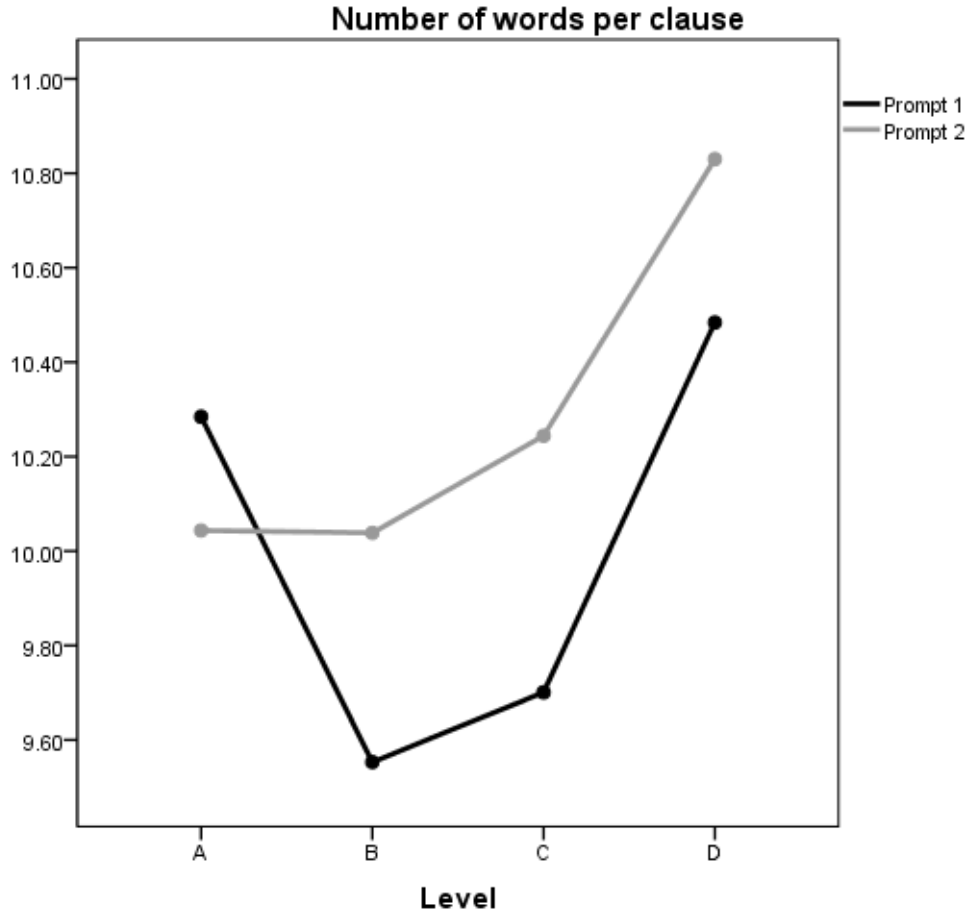

Figure 8: Words per clause

From Table 11 and Figure 8 above, it can be seen that there were no consistent differences between the levels in terms of the number of words produced per clause. There is some

difference across prompts, with discourse produced for Prompt 2 at levels B, C and D containing a greater number of words per clause. Conversely, at level A, discourse for Prompt 1 showed greater syntactic complexity in terms of the number of words per clause.

The ANOVA showed no significant main effect for grade level [F(3,166)=1.285, p=.281], for prompt [F(1,166)=1.009, p=.317] and no significant interaction effect [F(3,166)=.636, p=.593].

Lexical complexity – D-value

Table 12 below shows the descriptive statistics for the first measure of lexical complexity, d-value. This is followed by graphical representations in Figure 9.

	Score level	**N**	**Mean**	**SD**
d-value – Prompt 1				
	A	25	106.63	20.76
	B	25	107.00	13.69
	C	25	99.10	20.93
	D	10	114.80	20.14
d-value – Prompt 2				
	A	25	107.00	13.69
	B	25	102.76	14.96
	C	25	102.80	17.66
	D	6	101.56	20.57

Table 12: Descriptive statistics – d-value

From the table above and Figure 9 below, it can be seen that there were no consistent differences between the levels in terms of d-value.

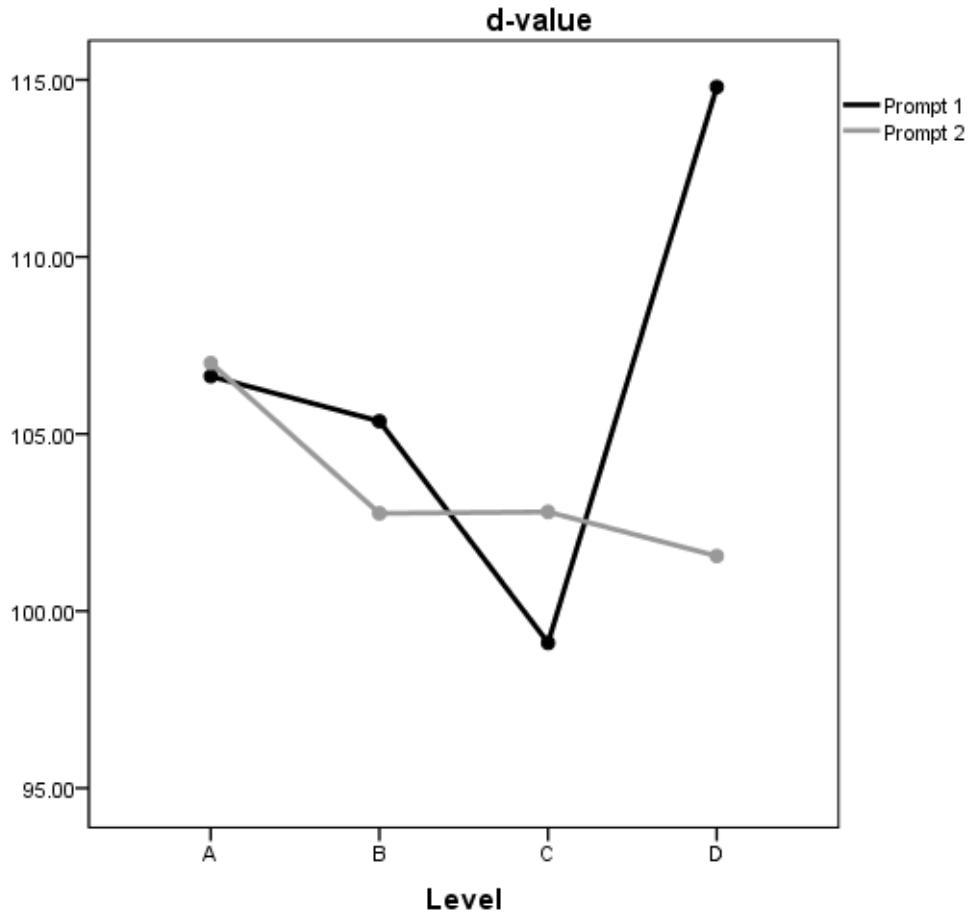

Figure 9: d-value

The ANOVA showed no significant main effect for grade level [F(3,166)=1.050, p=.372]. It also showed no significant main effect for prompt [F(1,166)=.779, p=.379] and no significant interaction effect [F(3,166)=.846, p=.471]. The post-hoc analysis showed that for Prompt 1 based on the LSD, A and B did not differ significantly from each other or from C and D, but C and D differed significantly. For Prompt 2 the post-hoc analysis showed that A, B, C and D did not differ significantly.

Lexical complexity – Average word length

Table 13 below shows the descriptive statistics for the measure, average word length. This is followed by graphical representations in Figure 10.

	Score level	N	Mean	SD
Average word length – Prompt 1				
	A	25	4.93	0.18
	B	25	4.90	0.14
	C	25	4.94	0.17
	D	10	4.82	0.22
Average word length – Prompt 2				
	A	25	5.26	0.18
	B	25	5.15	0.21
	C	25	5.28	0.21
	D	6	5.05	0.30

Table 13: Descriptive statistics – Average word length

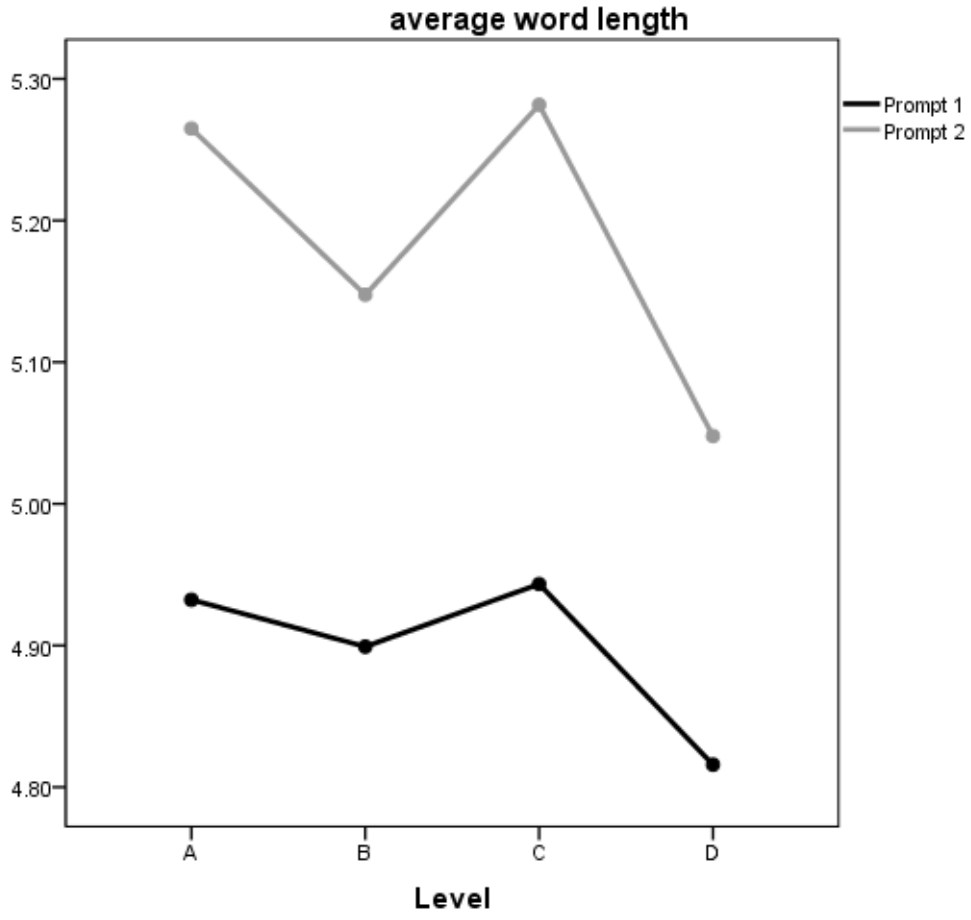

Figure 10: Average word length

As can be seen from Table 13 and Figure 10 above, there were no consistent differences between the levels in terms of average word length. There was some difference in this result

across prompts, with the discourse produced for Prompt 2 having higher average word lengths at all grade levels.

The ANOVA showed a significant main effect for grade level with a small to medium effect size [F(3,166)=4.760, p=.003, h^2=.083] and a significant main effect for prompt with a large effect size [F(1,166)=.71.575, p<.001, h^2=.312]. It showed no significant interaction effect [F(3,166)=.752, p=.523]. The post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1. For Prompt 2 the post-hoc analysis showed that A differed significantly from D; B differed significantly from C; and C differed significantly from D. A did not differ significantly from B or C, and B did not differ significantly from D.

Lexical complexity – Lexical density

Table 14 below shows the descriptive statistics for the measure, lexical density. This is followed by graphical representations in Figure 11.

	Score level	**N**	**Mean**	**SD**
Lexical density – Prompt 1				
	A	25	0.57	0.04
	B	25	0.59	0.04
	C	25	0.58	0.03
	D	10	0.58	0.03
Lexical density – Prompt 2				
	A	25	0.62	0.03
	B	25	0.60	0.03
	C	25	0.62	0.03
	D	6	0.64	0.03

Table 14: Descriptive statistics – Lexical density

As can be seen from the table above and Figure 11 below, there was no consistent differences between the levels in terms of lexical density. There were differences in this result across

prompts, with discourse produced for Prompt 2 showing greater lexical density at all grade levels.

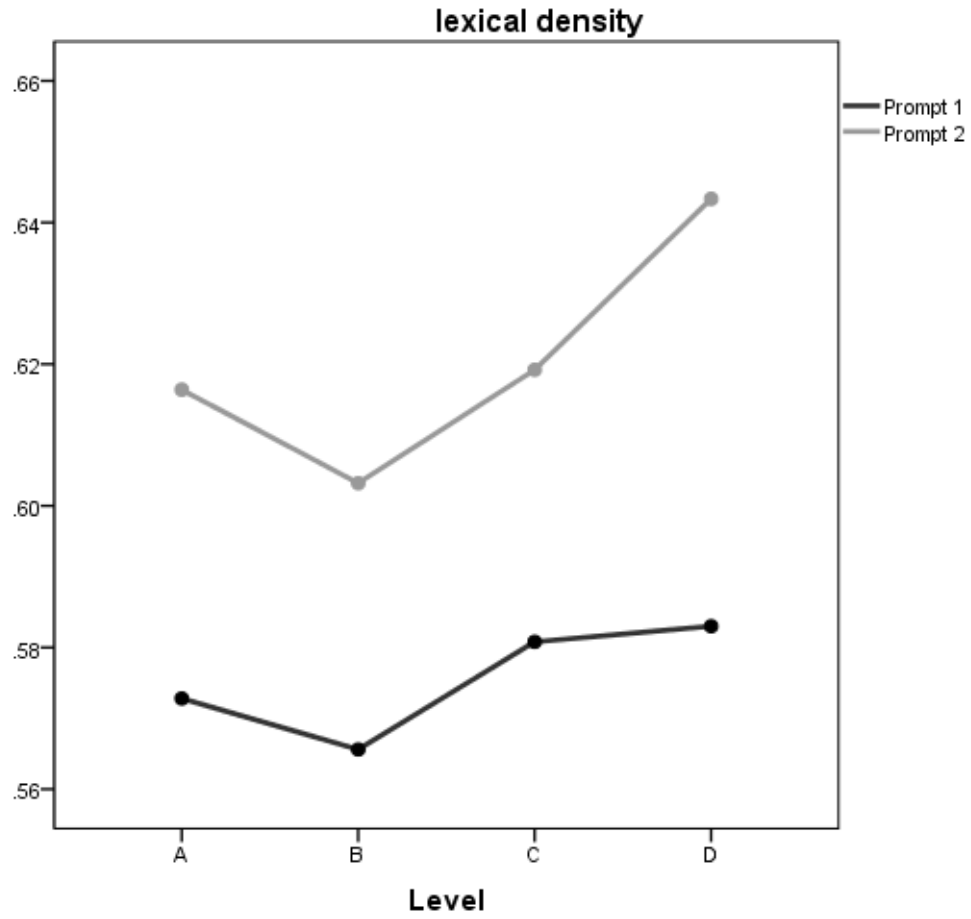

Figure 11: Lexical density

The ANOVA a significant main effect for grade level with a small to medium effect size [$F(3,166)=3.707$, $p=.013$, $h^2=.066$] and a significant main effect for prompt with a large effect size [$F(1,166)=60.114$, $p<.001$, $h^2=.276$]. It showed no significant interaction effect [$F(3,166)=.531$, $p=.662$]. The post-hoc analysis based on the LSD showed that for Prompt 1, there were no significant differences between A, B, C and D. The post-hoc analysis showed that for Prompt 2, B differed significantly from D but not from A and C. A did not differ significantly from C and D, and D did not differ significantly from C.

Lexical complexity – Lexical sophistication

Table 15 below shows the descriptive statistics for the measure, lexical sophistication. This is followed by graphical representations in Figure 12.

	Score level	N	Mean	SD
Lexical sophistication – Prompt 1				
	A	25	0.82	0.14
	B	25	0.86	0.17
	C	25	0.88	0.14
	D	10	0.77	0.11
Lexical sophistication – Prompt 2				
	A	25	0.92	0.18
	B	25	0.77	0.16
	C	25	0.82	0.11
	D	6	0.66	0.19

Table 15: Descriptive statistics – Lexical sophistication

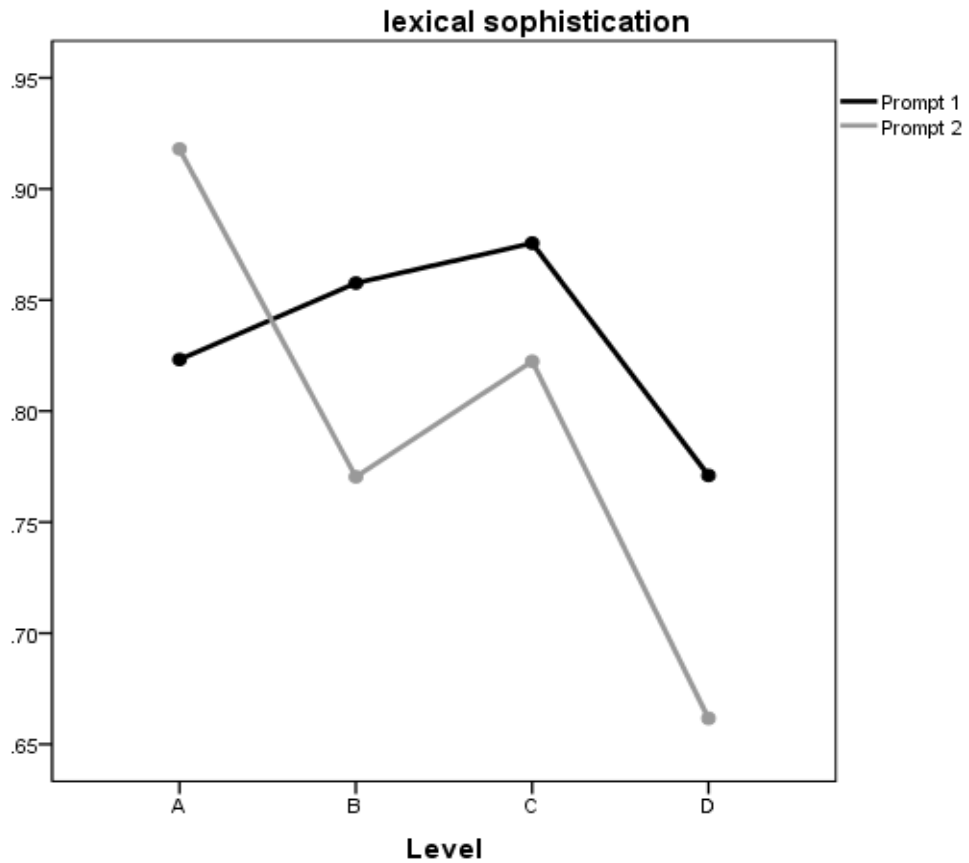

Figure 12: Lexical sophistication

As can be seen from Table 15 and Figure 12 above, there were no consistent differences between the levels in terms of lexical sophistication. Across the prompts, discourse produced for Prompt 1 at levels B, C and D show greater lexical sophistication compared with discourse at the corresponding levels for Prompt 2, whereas at level A, the reverse is true.

The ANOVA a significant main effect for grade level with a small to medium effect size [$F(3,166)=4.839$, $p=.003$, $h^2=.084$] and no significant main effect for prompt [$F(1,166)=2.250$, $p=.136$]. It showed a significant interaction effect with a small to medium effect size [$F(3,166)=4.184$, $p=.007$, $h^2=.074$]. The post-hoc analysis based on the LSD showed that for Prompt 1, A and B did not differ significantly from each other or from C and D, but that C and D differed significantly. For Prompt 2 it showed that A differed significantly from B, C and D. C and D also differed significantly, but B and C did not.

Lexical complexity – Percentage words from the AWL

Table 16 below shows the descriptive statistics for the measure, percentage of words from AWL. This is followed by graphical representations in Figure 13.

	Score level	**N**	**Mean**	**SD**
Percentage words from the AWL – Prompt 1				
	A	25	4.93	0.18
	B	25	4.90	0.14
	C	25	4.94	0.17
	D	10	4.81	0.22
Percentage words from the AWL – Prompt 2				
	A	25	5.26	0.18
	B	25	5.15	0.21
	C	25	5.28	0.21
	D	6	5.05	0.30

Table 16: Descriptive statistics – Percentage of words from the AWL

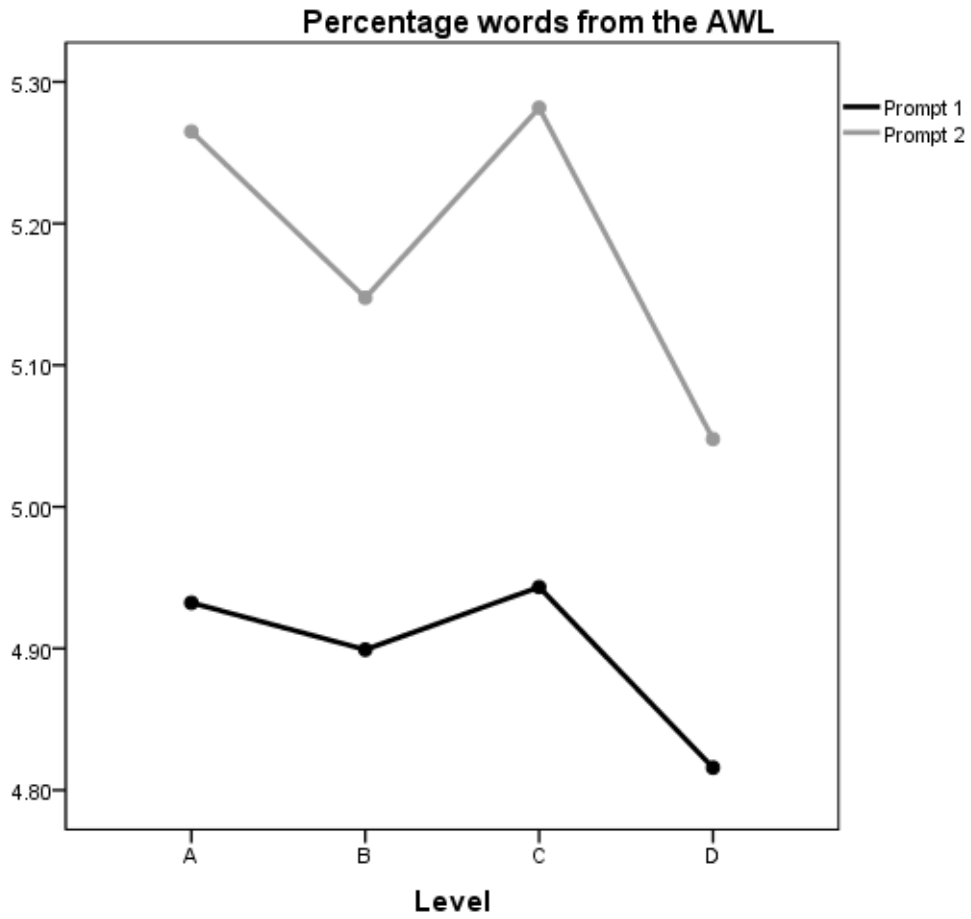

Figure 13: Percentage of words from the AWL

As can be seen from Table 16 and Figure 13 above, there were no consistent differences between the levels for the percentage of words from AWL. Across prompts, compared with Prompt 1, at all grade levels, the discourse produced for Prompt 2 contained a higher percentage of words from AWL.

The ANOVA a significant main effect for grade level with a small to medium effect size [$F(3,166)=4.760$, $p=.003$, $\eta^2=.083$] and a significant main effect for prompt with a large effect size [$F(1,166)=0.752$, $p<.001$, $\eta^2=.312$]. It showed no significant interaction effect [$F(3,166)=.752$, $p=523$]. The post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1. The post-hoc analysis showed that for Prompt 2 A differed significantly from D; B differed significantly from C; C also differed significantly

from D. A did not differ significantly from B and C, and B did not differ significantly from D.

Coherence

Proportion of coherent t-units

Table 17 below shows the descriptive statistics for the measure, proportion of coherent t-units. This is followed by graphical representations in Figure 14.

	Score level	N	Mean	SD
Proportion of coherent t-units – Prompt 1				
	A	25	0.72	0.09
	B	25	0.76	0.10
	C	25	0.78	0.10
	D	10	0.69	0.14
Proportion of coherent t-units – Prompt 2				
	A	25	0.73	0.11
	B	25	0.73	0.11
	C	25	0.76	0.10
	D	6	0.45	0.12

Table 17: Descriptive statistics – Proportion of coherent t-units

As can be seen from the table above and Figure 14 below, the proportion of coherent t-units in the discourse increased from level A through B and C, while the lowest proportion of coherent t-units was found in discourse at level D. There is some difference in this result across the two prompts, with the proportion of coherent t-units being lower for Prompt 2 at levels B, C and D (only slightly so at levels B & C), and the converse being true at level A.

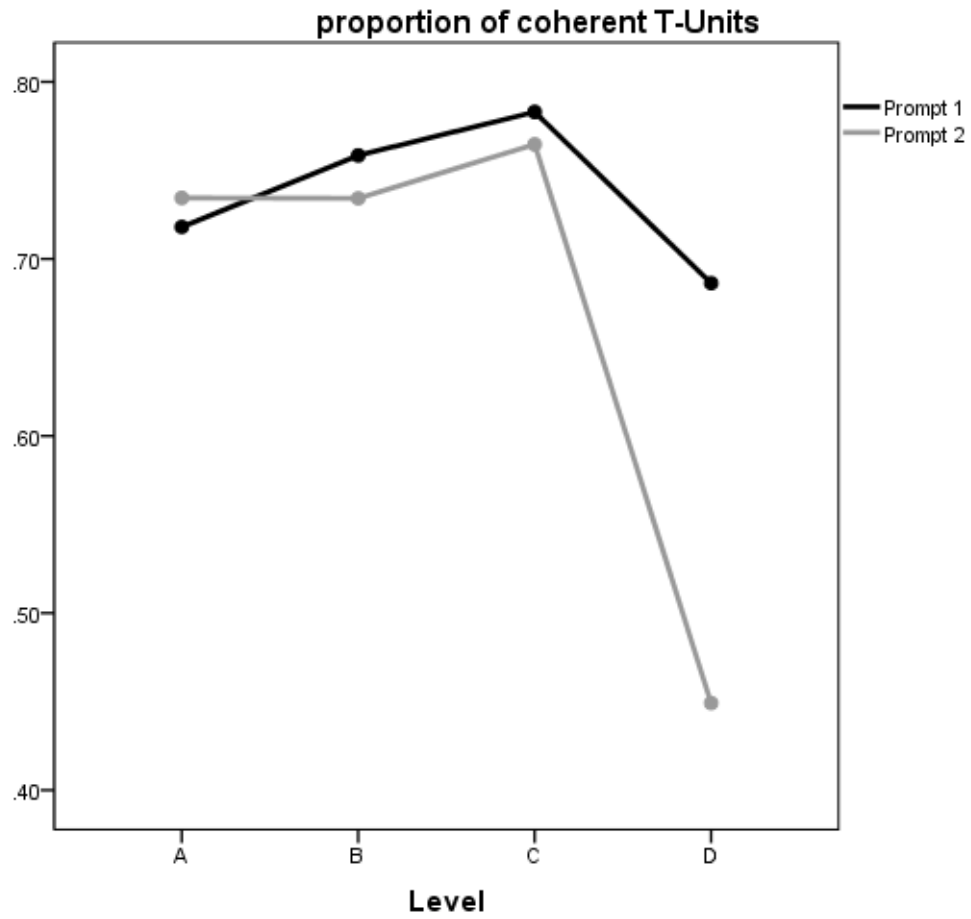

Figure 14: Proportion of coherent t-units

The ANOVA showed a significant main effect and a medium to large effect size for grade level [$F(3,166)=14.966$, $p<.001$, $h^2=.221$]. It showed a significant main effect for prompt with a small to medium effect size [$F(1,166)=12.184$, $p=.001$, $h^2=.072$] and a significant interaction effect with a small to medium effect size [$F(3,166)=5.637$, $p=.001$, $h^2=.097$]. The post-hoc analysis based on the LSD showed that for Prompt 1, A and C differed significantly. C differed significantly from D. There were no significant differences between B and A, D and A. Nor was there any significant difference between C and B, or D and C. For Prompt 2 it showed a significant difference between A and D; B and D; and C and D. A, B and C did not differ significantly.

Cohesion

Referential cohesion

Table 18 below shows the descriptive statistics for the measure, referential cohesion. This is followed by graphical representations in Figure 15.

	Score level	N	Mean	SD
Referential cohesion – Prompt 1				
	A	25	0.39	0.12
	B	25	0.35	0.14
	C	25	0.41	0.13
	D	10	0.38	0.23
Referential cohesion – Prompt 2				
	A	25	0.33	0.11
	B	25	0.36	0.12
	C	25	0.39	0.18
	D	6	0.34	0.14

Table 18: Descriptive statistics – Referential cohesion

As can be seen from the table above and Figure 15 below, there were no consistent differences between the levels in terms of the incidence of tokens of referential cohesion. There were some differences across prompts, with referential cohesion being higher in the discourse produced for Prompt 1 at levels A, C and D.

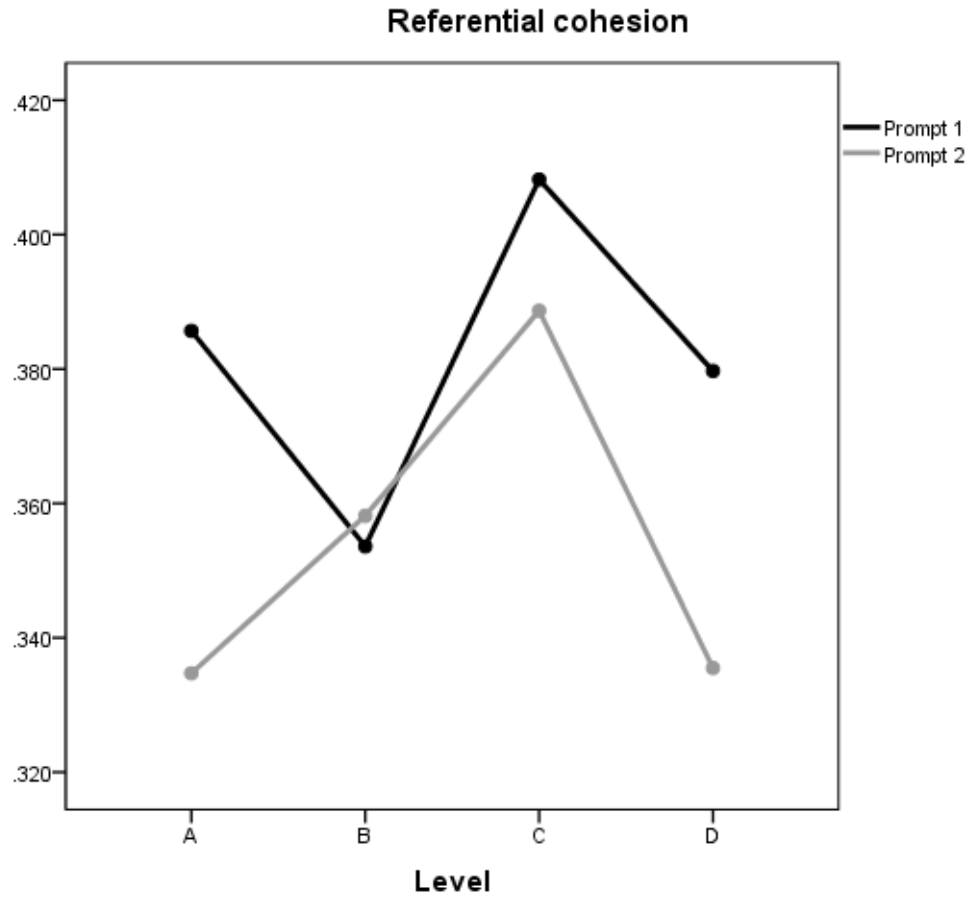

Figure 15: Referential cohesion

The ANOVA showed no significant main effect for grade level [F(3,166)=.912, p=.437]. It also showed neither a significant main effect for prompt [F(1,166)=1.138, p=0.288] nor a significant interaction effect [F(3,166)=.334, p=.801]. The post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1 or Prompt 2.

Number of connectives

Table 19 below shows the descriptive statistics for the measure, number of connectives. This is followed by graphical representations in Figure 16.

As can be seen from Table 19 and Figure 16 below, there were no consistent differences between the levels in terms of number of connectives. Across prompts, the incidence of connectives was higher in the discourse produced for Prompt 2 at all grade levels.

	Score level	N	Mean	SD
Number of connectives– Prompt 1				
	A	25	85.98	13.15
	B	25	84.65	16.90
	C	25	86.11	16.97
	D	10	83.00	16.97
Number of connectives – Prompt 2				
	A	25	96.44	12.33
	B	25	90.60	19.63
	C	25	94.65	21.17
	D	6	89.82	11.21

Table 19: Descriptive statistics – Number of connectives

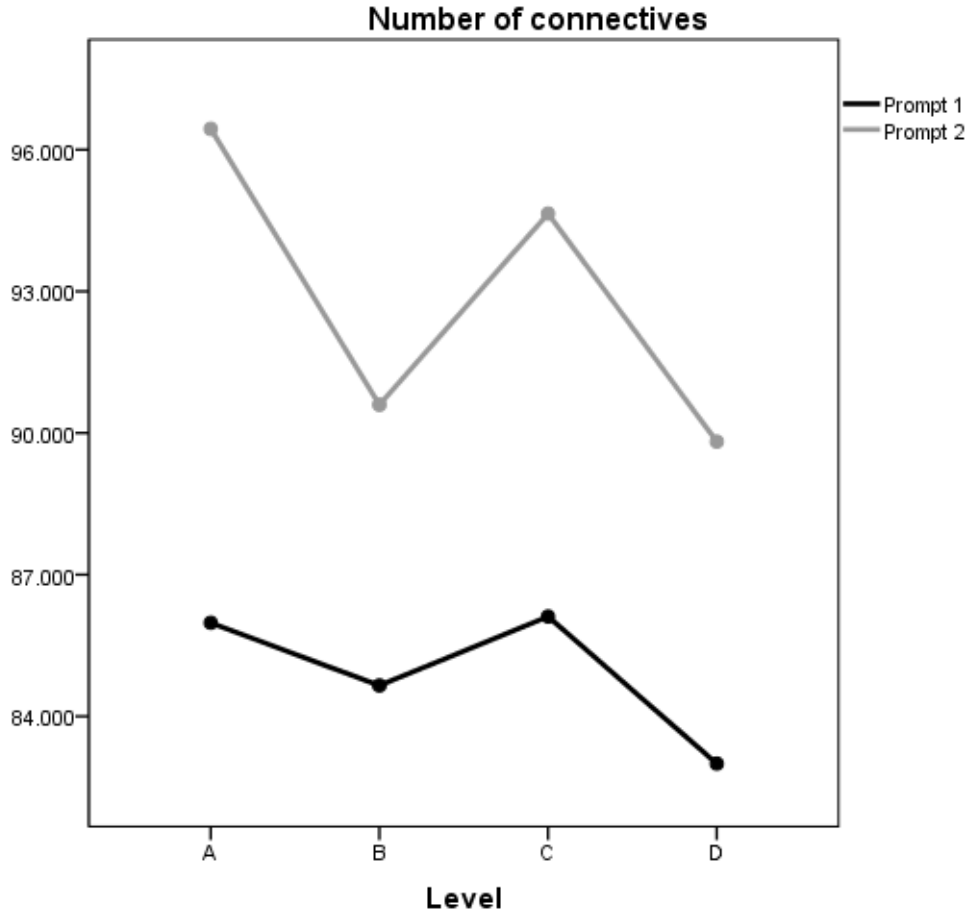

Figure 16: Number of connectives

The ANOVA showed no significant main for grade level [F(3,166)=.522, p=.668]. It showed a significant main effect for prompt with a small effect size [F(1,166)=6.194, p=.014,

h^2=.038] and no significant interaction effect [F(3,166)=.141, p=.935]. The post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1. For Prompt 2 A differed significantly from C; and B differed significantly from C and D. B did not differ significantly from A, and D did not differ significantly from A or C.

Content

Proportion of required idea units

Table 20 below shows the descriptive statistics for the measure, proportion of required idea units. This is followed by graphical representations in Figure 17.

	Score level	N	Mean	SD
Proportion of required idea units – Prompt 1				
	A	25	0.81	0.09
	B	25	0.84	0.08
	C	25	0.61	0.10
	D	10	0.54	0.20
Proportion of required idea units – Prompt 2				
	A	25	0.60	0.10
	B	25	0.57	0.10
	C	25	0.61	0.10
	D	6	0.54	0.20

Table 20: Descriptive statistics – Proportion of required idea units

As can be seen from the table above and Figure 17 below, there were no consistent differences between the levels in terms of proportion of required idea units. Across prompts, the discourse produced for Prompt 1 at all grade levels contained a higher proportion of required idea units compared with the discourse for Prompt 2.

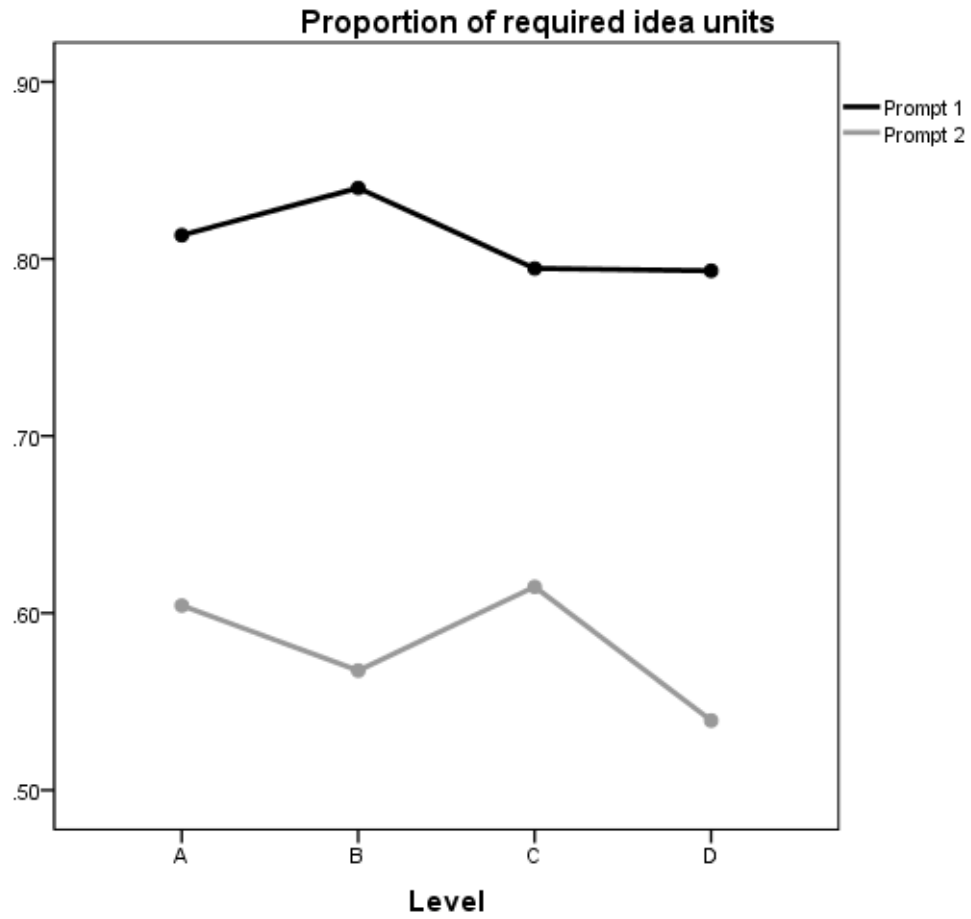

Figure 17: Proportion of required idea units

ANOVA showed no significant main for grade level [F(3,166)=.0.699, p=.554]. It showed a significant main effect for prompt with a large effect size [F(1,166)=157.968, p<.001, h^2=.500] and no significant interaction effect [F(3,166)=1.909, p=.130]. Post-hoc analysis based on the LSD showed no significant differences between levels for Prompt 1 or 2.

Proportion of irrelevant idea units

Table 21 below shows the descriptive statistics for the measure, proportion of irrelevant idea units. This is followed by graphical representations in Figure 18.

As can be seen from Table 21 and Figure 18 below, there were no consistent differences between levels in terms of proportion of irrelevant idea units. There were differences across prompts, with Prompt 1 discourse containing a higher proportion of irrelevant idea units at all grade levels compared with discourse produced for Prompt 2.

	Score level	N	Mean	SD
Proportion of irrelevant idea units – Prompt 1				
	A	25	0.36	0.17
	B	25	0.39	0.18
	C	25	0.48	0.20
	D	10	0.54	0.18
Proportion of irrelevant idea units – Prompt 2				
	A	25	0.19	0.10
	B	25	0.17	0.11
	C	25	0.25	0.18
	D	6	0.19	0.17

Table 21: Descriptive statistics – Proportion of irrelevant idea units

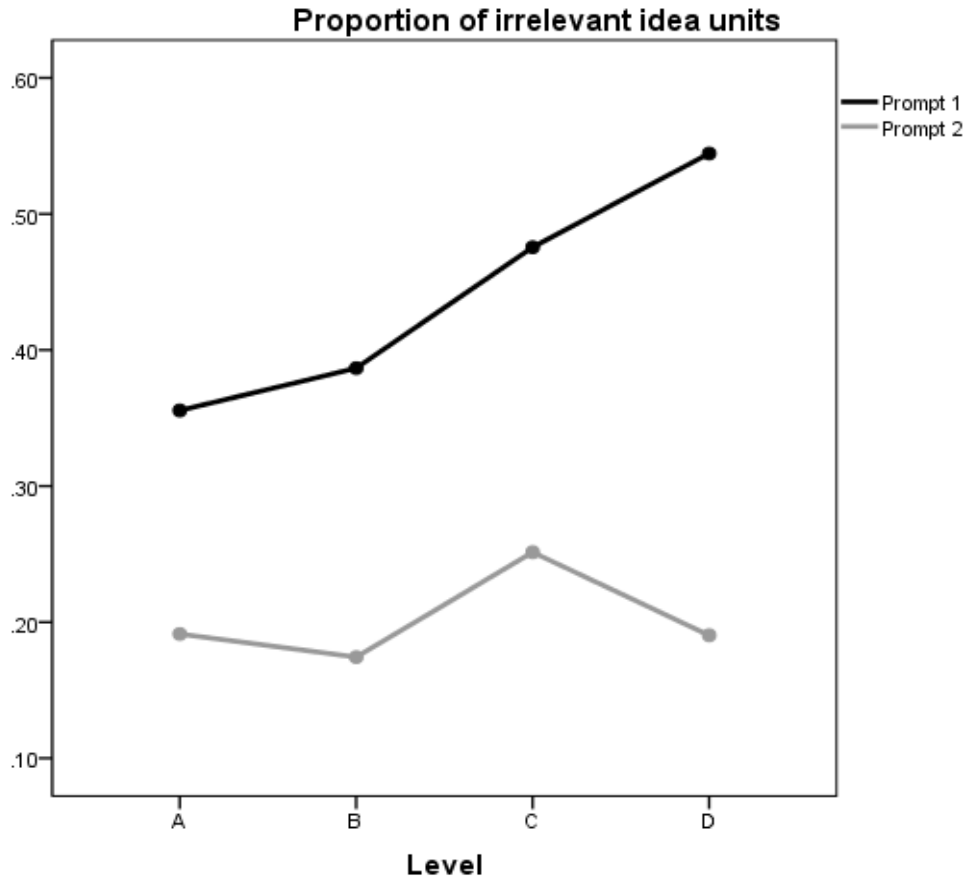

Figure 18: Proportion of irrelevant idea units

The ANOVA showed a significant main for grade level with a small to medium effect size [F(3,166)=.3.853, p=.011, h^2=.068]. It showed a significant main effect for prompt with a large effect size [F(1,166)=68.757, p<.001, h^2=.303] and no significant interaction effect

[F(3,166)=1.349, p=.025]. The post-hoc analysis based on the LSD showed that for Prompt 1, A differs significantly from C and D, and B from D. There was no significant difference between A and B, B and C or B and D. For Prompt 2 the levels did not differ significantly.

Structure

Expected macro-structural elements that appeared in the majority of scripts at all levels were: Purpose of referral, Restatement of purpose, Diagnosis, Presenting complaint, History, Findings/Results and Management. Pre-closing remarks, an optional element, appeared in approximately one third of the scripts. These elements are explicated in Table 22, below.

Element	Explication and example
Purpose (of referral)	Direct or indirect request/referral for assistance/advice/further assessment/care/admission e.g. *I am writing to refer Mrs Hong, a 43 year old woman*
Purpose-restatement	Repetition of initial statement of purpose of referral; more emphatic than initial statement and/or includes additional information/instructions e.g. *It would be greatly appreciated if you could assess Mrs Hong as soon as possible*
Diagnosis (medical diagnosis)	Statement of overall assessment of case, or overt/provisional/suggested diagnosis e.g. *[Mrs Hong] is suffering from right lower lobar pneumonia*
Presenting complaint *	Assessment of reason for patient visit/s e.g. *Initially…she presented to me with the complaints of productive cough and fever*
History	Patient medical history e.g. *Regarding her past medical history, she had been diagnosed with rheumatic carditis since childhood which resulted in mitral regurgitation and atrial fibrillation*
Findings/results *	Examination results i.e. physical findings, and investigation results i.e. test results (applicable to Prompt 2 only) e.g. *On examination, she looked tired with temperature 38 C, blood pressure 140/80*
Management *	Advice given, treatment commenced, tests ordered, plans to review e.g. *I prescribed Amoxycillin 500mg orally tds with an suggestion to reduce smoking*
Pre-closing	Formulaic comments before sign-off comprising references to attachments (e.g. test results) and/or offers/requests for follow up contact e.g. *Please do not hesitate to contact me if you require any further information*

* Where prompt includes case notes of initial and subsequent presentations (i.e. Prompt 2), these elements may recur in the referral letter, or may consist of a summary of presentations

Table 22: Macro-structural elements

Inclusion in the scripts of the optional element, Pre-closing remarks was not associated with level or prompt. Where present, Pre-closing remarks were always given immediately before final sign off (i.e. *Yours sincerely, Regards* etc.). The remaining elements listed above and which are expected in all scripts, appeared in a limited number of possible sequences as shown in Figure 19 below.

Purpose > Diagnosis > (History) > Presenting complaint > Findings/Results > Management > (History) > Purpose-restatement
Purpose > Diagnosis > History > Presenting complaint > Findings/Results > (Diagnosis) > Management > (Diagnosis) > Purpose-restatement
Prompt 2 only: Purpose > Diagnosis > Presenting complaint-summary > Findings/Results-summary > Management-summary > History > Purpose-restatement
Prompt 2 only: Purpose > Diagnosis > Presenting complaint-1 > Findings/Results-1 > Management-1 > Presenting complaint-2 > Findings/Results-2 > Management-2 > Presenting complaint-3, Findings/Results-3, Management-3 > History > Purpose-restatement

Figure 19: Prototypical sequencing of macro-structural elements

As shown in the figure above, four acceptable variations in the sequencing of elements were found: History can appear in two possible slots (shown in round brackets); Diagnosis can be repeated after Findings/Results and/or Management (shown underlined); and, for Prompt 2 only, Presenting complaint-Findings-Results/Management elements may be given *either* as a summary of the three presentations (patient visits) *or* in cyclical fashion for each of the three presentations (indicated with *summary*, or the numbers *1, 2, 3* in the figure above).

The sequences shown in the figure above were found to apply, respectively, to the majority of scripts, regardless of level or prompt. Exceptions were found in the discourse at the lowest level (grade D) where the following deviations from the expected macro-structures were evident: omission or unexpected sequencing of Diagnosis (embedded within the

Presenting complaint-Findings/Results-Management sequence) was found in 9 grade D scripts; omission or unexpected sequencing of Management (found in 5 grade D scripts). These findings are based on a sample that did not include many scripts at grade D, and therefore should taken with caution, they nevertheless suggest salient structural differences that could be further explored with a larger sample. For example, discourse at the highest level (grade A) was clearly distinguishable from all other levels by invariable positioning of Diagnosis in the opening sentence or paragraph.

Summary of results

Table 23, below, summarizes the results of the quantitative analyses. The size of the effect size is shown in brackets.

Measure	Level effect	Prompt effect	Interaction effect
Percentage error-free t-units	Sig (large)	Ø	Ø
Percentage error-free clauses	Sig (large)	Sig (small)	Sig (small/medium)
Number of words	Sig (small/medium)	Sig (medium)	Ø
Number of t-units	Ø	Sig (medium/large)	Ø
Number of clauses	Sig (small/medium)	Sig (small/medium)	Ø
Number of words per T-unit	Ø	Sig (small)	Ø
Number of clauses per T-unit	Ø	Sig (medium)	Ø
Number of words per clause	Ø	Ø	Ø
D-value	Ø	Ø	Ø
Average word length	Sig (small/medium)	Sig (large)	Ø
Lexical density	Sig (small/medium)	Sig (large)	Ø
Lexical sophistication	Sig (small/medium)	Ø	Sig (small/medium)
Percentage words from AWL	Sig (small/medium)	Sig (large)	Ø
Proportion of coherent t-units	Sig (medium/large)	Sig (small/medium)	Sig (small/medium)
Referential cohesion	Ø	Ø	Ø
Number of connectives	Ø	Sig (small)	Ø
Proportion required idea units	Ø	Sig (large)	Ø
Proportion irrelevant idea units	Sig (small/medium)	Sig (large)	Ø

Table 23: Summary of statistical results

Discussion

The discussion section will start by focussing on the 19 discourse-analytic variables employed in the study. Then, the broader issues of validity and rating scale validation will be discussed.

The aim of this study was to compare the quality of test taker discourse rated at different OET score levels. This was done by examining 19 discourse-analytic variables in a number of different areas of discourse: accuracy, fluency, syntactic and lexical complexity, coherence, cohesion, content and structure (see Table 2 for the complete list of discourse-analytic measures).

Two measures of ***accuracy*** were employed: percentage of error-free t-units and percentage of error-free clauses. These were chosen as they rely less on the intuitions of coders than other accuracy measures commonly employed (e.g. the number of errors per production unit). Both measures were significant for score level, showing that test takers at higher writing ability levels produced more accurate discourse. For error-free clauses only, there was only a small effect for prompt, with Prompt 2 discourse being less accurate at the lowest score level (grade D). Overall, the measures chosen for accuracy worked well in discriminating between test takers at different writing ability levels. The findings of this study are similar to findings by Cumming et al. (2006) and Knoch et al. (forthcoming) who also found a main effect for proficiency level.

To measure ***fluency***, this study employed three measures: the numbers of words, t-units and clauses. A limitation of using measures of temporal aspects of fluency such as these, the fact that researchers cannot be certain that writers spend all of the available time writing, has been pointed out in the literature review section. In the absence of any other means of measuring fluency however, these three measures were chosen. The measure, number of words has produced mixed results in previous studies. Although some researchers

have observed the number of words to increase with proficiency example, Cumming at al. (2006), Knoch (2009), and Knoch et al. (forthcoming), the results of these studies were not significant in all cases, include observations of a ceiling effect, and were task type dependent. In the present study, although the findings were significant for score level (with a small to medium effect), the number of words produced at each score level did was not increase unidirectionally in line with grade level. A medium effect was found for prompt, with the number of words in the discourse on Prompt 2 being greater than for Prompt 1. This result can be explained with reference to the input (case notes) for Prompt 2, which is somewhat longer than the input for Prompt 1. For the other two fluency measures, number of t-units and number of clauses, a level effect was found for number of t-units only. A prompt effect was found for both measures, with a greater number of t-units and clauses associated with the discourse for Prompt 2, also explainable with reference to the amount of the input for Prompt 2, compared with Prompt 1. In line with expectations based on previous studies of writing fluency including Cumming et al. (2006) (who employed number of words), Ishikawa (1995) and Kameen (1979) (number of clauses) and Hirano (1991), Ishikawa (1995) and others (using number of t-units) the results of the present study for these three measures are inconclusive, and therefore do not provide a useful means for discriminating between test takers at different writing ability levels.

Three measures were employed to investigate ***syntactic complexity***: number of words per t-unit, number of clauses per t-unit, number of words per clause. The findings of previous studies employing these measures have been mixed, with some finding a significant relationship between syntactic complexity and proficiency level as determined by scores and/or placement levels (for example, Cumming et al. 2006; 1991; Kameen, 1979) and others not (for example, Gebril & Plakans, 2009; Knoch, 2009). In the present study, the three syntactic complexity measures showed no consistent pattern of variation across levels, and

differences between levels were not statistically significant in nearly all cases (the exception was the number of clauses per t-unit for Prompt 2 discourse, which was greater at level A compared with level D). These measures are therefore not suitable for discriminating between test takers at different writing proficiency levels. A small prompt effect was found for number of words per t-unit (greater for Prompt 1) and a medium prompt effect for number of clauses per t-unit (greater for Prompt 1).

Five measures of ***lexical complexity*** were employed in this study: d-value, average word length, lexical density, lexical sophistication, and percentage of words from AWL. All of these measures were obtained using automated discourse analyses. The first measure, d-value, yielded no level effect and failed to discriminate amongst test takers at different writing ability levels. The other four measures, average word length, lexical density and sophistication, and percentage of AWL words, all showed a small to medium level effect. Lexical sophistication differed significantly between levels C and D (Prompt 1 only) and between level A and all other levels (Prompt 2 only). All other significant differences between levels were found between non-adjacent levels. Although these measures have been found to discriminate well between test takers at different writing proficiency levels in some previous studies (e.g. Grant & Ginther, 2000; Knoch, 2009), the findings of other studies employing these measures have been mixed. For example, significant results were not achieved by Cumming et al. (2006) and Gebril and Plakans (2009) for average word length, while Knoch et al. (forthcoming) found lexical density to be a poor discriminator of writing score level. In the same study by Knoch et al., lexical sophistication and percentage of AWL words were found to discriminate somewhat between different writing score levels. The main reason for the lexical complexity measures in this study not being successful in distinguishing between different score levels is probably that the writing task relies heavily on the input provided in the case notes and the fact that writers at all score levels have access to the same

input. They are therefore less reliant on providing their own vocabulary as is the case when responding to an independent writing task. The large task effect found for three of the measures of lexical complexity – average word length, lexical density and percentage of AWL words – with the associated finding that discourse produced for Prompt 2 showed greater lexical complexity, could be explained by the greater amount of input (longer case notes) for this prompt which may have led test takers to rely less on their own lexical resources and more on the input provided. A possible relationship between the input material provided with the prompts and observed task effects will be discussed further (below) in relation to the discourse-analytic variables in this study which pertain to cohesion and content.

The measure of ***coherence*** used in this study, the proportion of coherent t-units, successfully discriminated discourse at the lowest level (grade D) from the three higher levels, although only for Prompt 2. As this measure was not able to discriminate any further, and showed no discrimination for Prompt 1 discourse, it is not a suitable measure for differentiating between test takers at different levels of writing proficiency. Although some researchers have found such measures of coherence to distinguish quite well between test taker discourse at different score levels (e.g. Knoch et al., forthcoming), in general, previous work on coherence has focused on L1 writing. The results of the present study may reflect a need for further refinement of available measures of coherence in order to establish measures that are more effective for analysing L2 writing.

Two measures of ***cohesion*** were used in this study: tokens of referential cohesion, and number of connectives. The only significant finding for cohesion was a small task effect for number of connectives. The discourse for Prompt 2 contained more connectives at all grade levels compared with Prompt 1 discourse. It is possible that the longer and more complex input of Prompt 2, which contained case notes for three separate patient visits, may have

triggered greater use of connectives for the purposes of sequencing information pertaining to the three visits, and depicting developments/changes over time. This could be explored through further qualitative analysis of test taker discourse.

The measures chosen for ***content*** did not differentiate between test takers at different writing ability levels. One reason for this could be that the OET assessors (all being language trained rather than health professionals) are not in a position to make adequate judgements about the content of the writing sample. A large effect size for prompt was observed, with Prompt 1 discourse containing greater proportions of required idea units and irrelevant idea units compared with Prompt 2 discourse. An explanation for this finding, that test takers responding to Prompt 1 included greater proportions overall of the given idea units – both irrelevant *and* required – could be simply that the smaller quantity of input for Prompt 1 meant that test takers used a greater proportion of it in order to meet the task expectations for word length.

The final area of discourse analysed in this study was ***structure***, which was analysed qualitatively using a coding scheme developed for this study so as to capture the aspects of structure most specific to the domain and to the referral letter writing task in particular. By characterising structure with the measure, macro-structural elements the study was able to identify the expected sequence of elements (and the acceptable variations to this), as well as deviations from this exemplified in a number of the lowest level scripts (grade D). Further evidence of the characteristics of structure at different writing ability levels is needed to confirm the usefulness of this measure for differentiating test takers at different writing ability levels.

To sum up, the most useful of the discourse-analytic variables used in this study for discriminating between test takers at different writing proficiency levels were the two measures of accuracy:

- Percentage of error-free t-units
- Percentage of error-free clauses

The measure used for structure has potential to be a useful discriminator of test takers at different levels and in the present study was able to distinguish between the discourse of test takers at the highest (grade A) and the lowest levels (grade D). However, as pointed out in the above discussion of the findings for this variable, the macro-structural elements measure developed for this study would need to be applied in further research to better determine its usefulness for this purpose.

To return to the validity argument framework referred to at the start of the literature review section, the two measures of accuracy (percentages of error-free t-units and clauses) can be contextualised with reference to the 'evaluation inference' in this way: the findings for these two variables offer discourse-based evidence which provides empirical backing for the evaluation inference, and thus contribute to a validity argument for the OET. By providing empirical evidence to show that discourse accuracy improves with increasing OET score levels, this study validates the OET rating scale in so far as the criteria *Appropriateness of Language* and *Control of Linguistic Features* and their associated level descriptors, are oriented to the features of discourse documented in this study to vary consistently by writing proficiency level. Four sample scripts (using Prompt 1) which exemplify benchmark performances for discourse accuracy (according to the percentages of error-free t-units and clauses) at each of the levels considered in this study, are provided in Appendix B.

Of further relevance to this study is the 'explanation inference' of the validity argument. This inference rests on the assumption that test tasks engage language abilities similar to those underlying real world tasks in the relevant domain (Xi, 2008). That the OET aims to satisfy this assumption is reflected in the fact that the rating scale invokes task and domain specific discourse features with the criterion, *Appropriateness of Language* i.e. the

descriptor cites ‘control of genre (referral letter)’ and ‘logica[l] organis[ation] in a more-or-less formulaic sequence appropriate to both task and professional context’. Although this study did not demonstrate convincingly that discourse structure (as measured here with ‘macro-structural elements’) can be used to discriminate amongst test takers at different writing ability levels, the findings demonstrate that the Writing task elicits discourse that is clearly domain-relevant and domain-specific. To the extent that it has shown that the test tasks engage real world and domain-relevant features of discourse structure, this study offers discourse-based evidence in support of the explanation inference. Two sample scripts (using Prompt 2) which exemplify benchmark performances for discourse structure (according to inclusion of expected macro-structural elements and adherence to expected sequencing of elements), are provided in Appendix B.

Further findings of this study concern differences in the quality of discourse produced across prompts. The greatest differences in discourse quality across prompts were found for the following variables:

- Fluency: numbers of words, t-units and clauses
- Cohesion: number of connectives
- Content: proportions of required and irrelevant idea units
- Lexical complexity: average word length, lexical density, percentage of AWL words

It is likely that these findings may be explained with reference to differences between the two prompts in terms of amount and format of input i.e. Prompt 1 case notes contain less information and fewer words compared with Prompt 2; and Prompt 1 comprises notes of a single patient visit, whereas Prompt 2 comprises notes for three successive visits. Given these differences in input, it might be reasonable to expect *intuitively* that there would be qualitative differences in test taker discourse across the prompts. The findings of this study now contribute empirical evidence of such differences between the discourse triggered by the

two prompts. This finding has implications for task design considerations since the discourse features found to vary by task type are represented in the rating scale; in this way, a prompt effect may be implicated in test taker scores. With the exception of lexical complexity, all of the above aspects of discourse are addressed overtly in the scale as follows: 'fluency' in *Overall Task Fulfilment* ('Answer may be (slightly/far) too long or (far) too short'); 'cohesion' in *Control of Linguistic Features* ('control of…cohesive devices; use of appropriate…connectives'); 'content' in *Comprehension of Stimulus* ('selection of…relevant material from the [case] notes…adequacy of content (coverage of main points)'). Although lexical complexity is not covered overtly in the rating scale, given that lexical complexity essentially refers to the range and size of a lexicon (Wolfe-Quintero et al., 1998), lexical complexity is closely associated with performance in relation to *Appropriateness of Language*.

Conclusion

This study has some limitations which need to be mentioned. Firstly, the data used for this study was provided by OET Centre and it was not possible to obtain an equal number of scripts at all proficiency levels. The difficulty of obtaining sample performances at the full range of proficiency levels is a common problem for studies of this type. As a consequence, scripts at the lowest level (grade D) were unavoidably under represented in the data and no scripts at Level E were included in the study.

Also, the selection of discourse-analytic measures was subject to practical considerations. The automated measures used in this study have proven to be useful in a large number of previous studies, and they are also convenient when it comes to analysing a large number of scripts (such as the 166 scripts used in this study). However, they are not always successful at capturing more nuanced aspects of discourse which lend themselves better to

more time consuming, and expensive, human rating. This issue may account for the fact that very few of the variables used in the study were able to capture consistent differences in discourse quality across adjacent levels of writing proficiency.

A further consideration regarding the findings is the somewhat circular nature of studies such as this one in which the proficiency levels investigated are determined on the basis of raters' scores; in turn, the only validation of these is the findings of the very study relying on them in the first instance. However, as is the case for many studies like this one, in the absence of an external measure of writing proficiency, using the OET grade levels was the best available option for the research design.

One of the outcomes of this study is the identification of benchmarks performances and other sample scripts as cited in the discussion section above. A small number of these can be found in Appendix B. The samples chosen are those pertaining to the level effect for discourse accuracy, and to a discourse variable developed for this study, macro-structure, which has potential to further contribute to the validity argument for the OET. The samples provide a resource for rater training and/or may be displayed by OET Centre as sample performances for the benefit of prospective or in-preparation test takers, and other stakeholders. A future outcome of this study will be the dissemination of the findings internationally to the language testing community via submission to an academic journal. Publication of the study will further enhance the profile of the OET internationally within the research and language testing community.

References

Al-Alfi, M.A, Al-Saigul, A.M., Abed-Elbast, A.M., Sourour, A.M. & Ramzy, H.A. (2007). Quality of primary care referral letters and feedback reports in Buraidah, Qassim region, Saudi Arabia. *Journal of Family and Community Medicine, 14*(3), 113–117. Available: http://www.jfcmonline.com/text.asp?2007/14/3/113/97099 [accessed 2013 Aug 20].

Anspach, R. (1988). Notes on the sociology of medical discourse: the language of case presentation. *Journal of Health and Social Behaviour, 29*, 357–375.

Arnaud, P.J.L. (1992). Objective lexical and grammatical characteristics of L2 written and the validity of separate-component tests. In P.J.L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 133–145). London: Macmillan.

Bachman, L.F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1-34.

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

Banerjee, J., & Franceschina, F. (2006, February). *Documenting features of written language production typical at different IELTS band score levels*. Paper presented at the Workshop sponsored by the European Science Foundation entitled 'Bridging the gap between research on second language acquisition and research on language testing', Amsterdam.

Banerjee, J., Franceschina, F., & Smith, A.M. (2007). *Documenting features of written language production typical at different IELTS band score levels*: British Council and IELTS Australia.

Bardovi-Harlig, K., & Bofman, T. (1989). Attainment of syntactic and morphological accuracy by advanced language learners. *Studies in Second Language Acquisition, 11*, 17-34.

Bhatia, V.K. (1993). *Analysing genre: language use in professional settings.* Essex, England; New York: Longman.

Bhatia, V.K. (2004). *Worlds of written discourse*. London; New York: Continuum.

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

Brown, A.L., Day, J.D., & Jones, R.S. (1983). The development of plans for summarizing for texts. *Child Development, 54*, 968–79.

Burbach, F.R. & Harding, S. (1997). GP referral letters to a community mental health team: an analysis of the quality and quantity of information. *International Journal of Health Care Quality Assurance, 10*(2), 67–72.

Chappelle, C. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), The *Routledge handbook of language testing* (pp. 21–33). Abington, Oxon; New York: Routledge.

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

Chenoweth, N.A., & Hayes, J.R. (2001). Fluency in writing: Generating text in L1 and L2. *Written Communication, 18*(1), 80–98.

Cohen, A.D. (1993). The role of instructions in testing summarizing ability. In D. Douglas & C. Chapelle (Eds.), *A new decade of language testing research* (pp. 132-160). Arlington, VA: TESOL.

Cobb, T. (2002). Web VocabProfile. Retrieved 12 December 2013, from http://www.lextutor.ca/vp/

Coffman, G.A. (1994). The influence of question and story variations on sixth graders' summarization behaviors. *Reading Research and Instruction, 34*(1), 19–38.

Cooper, T.C. (1976). Measuring written syntactic patterns of second language learners of German. *Journal of Educational Research, 69*(5), 176–183.

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

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238.

Cumming, A.H., Kantor, R., Baba, K., Eouanzoui, K., Erdosy, U., & James, M. (2006). *Analysis of discourse features and verification of scoring levels for independent and integrated tasks for the new TOEFL*. (TOEFL Monograph Series 30 Rm-05-13). Princeton, NJ: Educational Testing Service.

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

Crowhurst, M. (1987). Cohesion in argument and narration at three grade levels. *Research in the Teaching of English, 21*(2), 185-201.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford: Oxford University Press.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139–155.

Enright, M.K., Bridgeman, B., Eignor, D., Kantor, R., Mollaun, P., & Nissan, S. (2008). Prototyping new assessment tasks. In C.A. Chapelle, M.K. Enright & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 145-186). New York: Routledge.

Field, Y., & Yip, L. (1992). A comparison of internal cohesive conjunction in the English essay writing of Cantonese speakers and native speakers of English. *RELC Journal, 23*(1), 15-28.

Fisher, R.A. (1984). Testing written communicative competence in French. *Modern Language Journal, 68*(1), 13–20.

Flahive, D.E., & Snow, B.G. (1980). Measures of syntactic complexity in evaluating ESL compositions. In J. W. Oller & K. Perkins (Eds.), *Research in language testing* (pp. 171–176). Rowley, MA: Newbury House.

Frase, L., Faletti, J., Ginther, L., & Grant, L. (1999). *Computer analysis of the TOEFL® Test of Written English™*. TOEFL Research Report No. RR-64. Princeton, NJ: Educational Testing Service.

Friedlander, A. (1990). Composing in English: Effects of a first language on writing in English as a Second Language. In B. Kroll (Ed.), *Second langauge writing: Research insights for the classroom* (pp. 109-125). Cambridge: Cambridge University Press.

Gebril, A., & Plakans, L. (2009). Investigating source use, discourse features, and process in integrated writing tests. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 7*, 47–84.

Gipps, C., & Ewen, E. (1974). Scoring written work in English as a second language: The use of the T-unit. *Educational Research, 16*(2), 121–125.

Graesser, A.C. & McNamara, D.S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science, 3*, 371–398.

Granger, S., & Tyson, S. (1996). Connector usage in the English essay writing of native and non native EFL speakers of English. *World Englishes, 15*(1), 17-27.

Grant, L., & Ginther, A. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing, 9*(2), 123–145.

Halliday, M.A.K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Harvey, K. & Koteyko, N. (2013). *Exploring health communication language in action*. Abington, Oxon; New York: Routledge.

Henry, K. (1996). Early L2 writing development: A study of autobiographical essays by university-level students of Russian. *Modern Language Journal, 80*(3), 309–326.

Hirano, K. (1991). The effect of audience on the efficacy of objective measures of EFL proficiency in Japanese university students. *Annual Review of English Language Education in Japan, 2*(1), 21–30.

Hobbs, P. (2003). The use of evidentiality in physician's progress notes. *Discourse Studies, 5*, 451–78.

Hoenisch, S. (1996). The theory and method of topical structure analysis. Retrieved 30 April 2007, from http://www.criticism.com/da/tsa-method.php

Ho-Peng, L. (1983). Using T-unit measures to assess writing proficiency of university ESL students. *RELC Journal, 14*(2), 35–43.

Homburg, T.J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18*(1), 87–107.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461–473.

Ishikawa, S. (1995). Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing, 4*, 51–70.

Iwashita, N. & Grove, E. (2003). A comparison of analytic and holistic scales in the contextof a specific-purpose speaking test. *Prospect, 18*(3), 25-35.

Jafarpur, A. (1991). Cohesiveness as a basis for evaluating compositions. *System, 19*(4), 459-465.

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

Kameen, P.T. (1979). Syntactic skill and ESL writing quality. In C. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79: The learner in focus* (pp. 343–364). Washington, D.C.: TESOL.

Kane, M.T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement*, 4th edn. (pp. 17-64). Westport: CT: American Council on Education and Praeger.

Kane, M.T. (2012) Articulating a validity argument. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 34–47). Abington, Oxon; New York: Routledge.

Kawata, K. (1992). Evaluation of free English composition. *CASELE Research Bulletin, 22,* 49–53.

Kennedy, C., & Thorp, D. (2002). *A corpus-based investigation of linguistic responses to an IELTS academic writing task*: University of Birmingham.

Kepner, C. (1991). An experiment in the relationship of types of written feedback to the development of second-langauge writing skills. *Modern Language Journal, 75*, 305–313.

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

Knoch, U. (2007a). *The development and validation of an empirically-developed rating scale for academic writing.* Unpublished PhD, University of Auckland.

Knoch, U. (2007b). 'Little coherence, considerable strain for reader': A comparison between two rating scales for the assessment of coherence. *Assessing Writing, 12*, 108-128.

Knoch, U. (2009). *Diagnostic writing assessment: The development and validation of a rating scale*. Frankfurt am Main: Peter Lang.

Knoch, U. (2011). Investigating the effectiveness of individualized feedback to rating behavior – a longitudinal study. *Language Testing, 28*(2), 179-200.

Knoch, U., Macqueen, S., O'Hagan, S. (forthcoming). *An investigation of the effect of task type on the discourse produced by students at various score levels in the TOEFL iBT writing test*. TOEFL Research Report.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12*(4), 439–448.

Larsen-Freeman, D. (1983). Assessing global second language proficiency. In H. W. Seliger & M. H. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 287–304). Rowley, MA: Newbury House.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning, 27*(1), 123–134.

Lautamatti, L. (1987). Observations on the development of the topic of simplified discourse. In U. Connor & R. B. Kaplan (Eds.), *Writing across languages: Analysis of L2 text*.

Linné, Y. & Rössner, S. (1998). What is 'obesity' - an analysis of referral letters to an obesity unit. *International Journal of Obesity Related Metabolic Disorders, 22*(12), 1231-1233. PMID: 9877259

Lumley, T. (1995). The judgements of language-trained raters and doctors in a test of English for health professionals. *Melbourne Papers in Language Testing 4*(1), 74-98.

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan & A. Wray (Eds.), *Evolving models of language* (pp. 58–71). Clevedon, Philadelphia: Multilingual Matters.

Malvern, D.D., & Richards, B.J. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*(1), 85–104.

Malvern, D.D., Richards, B.J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, Palgrave Macmillan.

McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2005, January 1). Coh-Metrix version 1.4. Retrieved 3 March 2013, from http://tool.cohmetrix.com.

Meara, P.M. & Miralpeix, I. (n.d). D_Tools: the Manual. Retrieved 6 February 2013, from http://www.lognostics.co.uk/index.htm.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13–103). Washington, DC: American Council on Education and National Council on Measurement in Education.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749.

Monroe, J.H. (1975). Measuring and enhancing syntactic fluency in French. *The French Review, 48*(6), 1023–1031.

Moselhy, H.F. & Salem, M.O. (2009). Referrals to psychiatric service in United Arab Emirates: an analysis of the content of referral letters. *International Journal of Health Sciences, Qassim University, 3*(1), 13-18.

Neuner, J.L. (1987). Cohesive ties and chains in good and poor freshman essays. *Research in the Teaching of English, 21*(1), 92-105.

Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555–578.

Newton, J., Hutchinson, A., Hayes, V., McColl, E., Mackee, I. & Holland, C. (1994). Do clinicians tell each other enough? An analysis of referral communications in two specialties. *Family Practice, 11*(1), 15–20.

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly, 14*(1), 61–69.

Perkins, K., & Leahy, R. (1980). Using objective measures of composition to compare native and non-native compositions. In R. Silverstein (Ed.), *Occasional Papers in Linguistics, No. 6* (pp. 306–316). Carbondale: Southern Illinois University.

Polio, C.G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101-143.

Reid, J. (1992). A computer text analysis of four cohesion devices in English discourse by native and nonnative writers. *Journal of Second Language Writing, 1*(2), 79-107.

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

Schneider, M., & Connor, U. (1990). Analyzing topical structure in ESL essays. *Studies in Second Language Acquisition, 12*(4), 411-427.

Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*(1), 1–14.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532.

Shaw, I., Clegg Smith, K.M., Middleton, H. & Woodward, L. (2005). A letter of consequence: referral letters from general practitioners to secondary mental health services. *Qualitative Health Research, 15*(1), 116c128.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings* Cambridge: Cambridge University Press.

Tattersall, M.N.H, Griffin, A-M., Dunn, S.M. et al. (1995). Writing to referring doctors after a new patient consultation. What is wanted and what was contained in letters from one medical oncologist. *Australian and New Zealand Journal of Medicine, 25*, 479–482.

Tattersall, M.N.H., Butow, P.N., Brown, J.E. & Thompson, J.F. (2002). Improving doctors' letters. *Medical Journal of Australia, 177*(9), 516–520.

Tedick, D. J. (1990). ELS writing assignment: Subject-matter knowledge and its impact on performance. *English for Specific Purposes, 9*, 123–143.

Tomita, Y. (1990). T-unit o mochiita kokosei no jiyu eisaku bun noryoku no sokutei (Assessing the writing ability of high school students with the use of t-units). *Step Bulletin, 2,* 14–28.

van Dijk, T.A. (1988). *News as discourse*. Hillsdale, NJ: Erlbaum Associates.

Vann, R.J. (1979). Oral and written syntactic relationships in second language learning. In C. Yorio, K. Perkins, & J. Schachter (Eds.), *On TESOL '79: The learner in focus* (pp. 322–329). Washington, D.C.: TESOL.

Widdowson, H.G. (1979). *Explorations in applied linguistics*. Oxford; New York: Oxford University Press.

Wigglesworth, G., & Foster, P. (2008, April). *Measuring accuracy in second language performance*. Paper presented at the TESOL Convention, New York.

Witte, S., & Faigley, L. (1981). Cohesion, coherence and writing quality. *College Composition and Communication, 32*(2), 189-204.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy & complexity*. Honolulu, HI: University of Hawai'i at Mānoa.

Xi, X. (2008). Methods of test validation. In E. Shohamy & N.H. Hornberger (Eds.), *Encyclopedia of Language and Education*. New York: Springer.

Yau, M. (1991). The role of language factors in second language writing. In L. M. Malavé & G. Duquette (Eds.), *Language, culture, and cognition: A collection of studies in first and second language acquisition* (pp. 266–283). Clevedon, Philadelphia: Multilingual Matters.

Yu, G. (2007). Students' voices in the evaluation of their written summaries: empowerment and democracy for test takers? *Language Testing, 24*(4), 539–572.

Yu, G. (2008). Reading to summarize in English and Chinese: a tale of two languages? *Language Testing*, 25(4), 521–551.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics, 31*, 236–259.

Appendix A: Writing tasks

Prompt 1:

OCCUPATIONAL ENGLISH TEST

WRITING SUB-TEST: MEDICINE

TIME ALLOWED: READING TIME: 5 MINUTES

WRITING TIME: 40 MINUTES

Read the case notes and complete the writing task which follows.

Notes:

Mr Brian Edwards (born on 16 May 1945) is a patient in your General Practice.

Patient Details

Name:	Mr Brian Edwards
Residence:	42 Grandview Drive Mountain Valley
Social background:	65-year-old retired accountant Married, lives at home with wife Non-smoker, non-drinker

Patient History

12.03.2011

Subjective:

Presenting complaint:	12/12 Hx progressively enlarging skin lesion on L lower leg
	Associated 2/52 Hx swelling & erythema
	Patient is anxious
	No pain/tenderness; no bone pain
	No LOW, no fever, no change in bowel habits
	Otherwise healthy, independent in activities of daily living
Past medical Hx:	2005 - Glaucoma
	2008 - SCC R pre-tibial skin lesion excised
Medications:	timolol maleate 0.25% solution, 1 drop twice daily
	No known allergies
Family Hx:	Nil skin Ca or other Ca; nil CVD or Diabetes

TURN OVER 2

Objective:	T - 36.7ºC; P - 80, regular; Ht - 172cm; Wt - 78kg
	Alert & orientated
	3cm skin lesion on L tibia area
	Irregular edge, ulcerated, erythematous, purulent discharge
	Systems review – normal
Assessment:	? SCC
Plan:	Refer to plastic surgeon for assessment (?infection) and excision
	Counsel patient on surgery (incl. skin grafts) and provide reassurance that lesion is most probably localised
	Start oral flucloxacillin

Writing Task:

Using the information given in the case notes, write a letter of referral to plastic surgeon, Mr Jon Liew, seeking follow-up assessment.

Address the letter to: Mr Jon Liew, Department of Plastic Surgery, Main Hospital, Royal Avenue, Newtown.

In your answer:

- **Expand the relevant notes into complete sentences**
- **Do not use note form**
- **Use letter format**

The body of the letter should be approximately 180–200 words.

3

Prompt 2:

OCCUPATIONAL ENGLISH TEST

WRITING SUB-TEST: MEDICINE

TIME ALLOWED: READING TIME: 5 MINUTES
WRITING TIME: 40 MINUTES

Read the case notes below and complete the writing task which follows.

Notes:

Mrs May Hong is a 43-year-old patient in your general practice.

21/11/2012

Subjective:
Noted a productive cough over last 3/7
No dyspnoea or pain
Feverish
Continues to smoke 10 cigarettes/day!
History: Rheumatic carditis in childhood, resulting in mitral regurgitation & atrial fibrillation (AF)

Objective:
Looks tired
T: 38℃
P: 80, AF
BP: 140/80
Ear, nose, throat (ENT) – NAD
Moist cough
Scattered rhonchi through chest, otherwise ok
Apical pansystolic murmur

Assessment: Acute bronchitis; cigarettes ↑ condition severity ++

Plan:
Advised – cease smoking
Amoxycillin 500mg; orally t.d.s.
Other medications unchanged (digoxin 0.125mg mane, warfarin 4mg nocte)
No known allergies (NKA)
Review 2/7
Check prothrombin ratio next visit

23/11/2012

Subjective:
Cough ↑, thick yellow phlegm
Feels quite run-down
Not dyspnoeic
Taking all medications
No cigarettes for last 48/24

Objective:
Looks worn-out
T: 38.5℃
P: 92, AF
BP: 120/80
Mild crackles noted at R lung base posteriorly
Occasional scattered crackles. Otherwise unchanged

TURN OVER 2

Assessment:	Bronchitis ↑; early R basal pneumonia
Plan:	Sputum sample for microscopy and culture (M&C) FBE, chest X-ray Chest physiotherapy Prothrombin ratio today (result in tomorrow) Review tomorrow

24/11/2012

Subjective:	Brought in by son Quite a bad night Symptoms ↑ Pleuritic R-sided chest pain, febrile, dyspnoea Prothrombin ratio result 2.4 (target 2.5-3.5)
Objective:	Unwell, tachypnoeic T: 38°C P: 110, AF BP: 110/75 Jugular venous pressure (JVP) not elevated R lower lobe dull to percussion with overlying crackles L basal crackles present Pansystolic murmur is louder M&C: gram + streptococcus pneumoniae, sensitive – clarithromycin & erythromycin Amoxicillin resistant Chest X-ray: Opacity R lower lobe (RLL) FBE: Leukocytosis 11.0 x 109/L
Assessment:	R lower lobar pneumonia
Plan:	Urgent hospital admission. Spoke with Dr Roberts, admitting officer, Newtown Hospital Ambulance transport organised

Writing Task:

Using the information given in the case notes, write a letter of referral to Dr L Roberts, the Admitting Officer at Newtown Hospital, 1 Main Street, Newtown, for advice, further assessment and treatment.

In your answer:

- **Expand the relevant notes into complete sentences**
- **Do not use note form**
- **Use letter format**

The body of the letter should be approximately 180–200 words.

Appendix B: Sample scripts

Discourse measure: *Accuracy* at grades D, C, B, A (Prompt 1)

Below are four sample scripts chosen to exemplify increasing discourse accuracy with increasing writing proficiency level. For the variables, percentage of error-free t-units and percentage of error-free clauses, each sample essay below scored within one standard-deviation of the mean score for essays at the relevant grade level. Note, for these and all further sample scripts, only the main body of each sample is reproduced; address, date, opening salutations and closings have been removed for the sake of presenting relevant parts of the scripts only. Fair scores and raw scores are shown (raw scores for 1st, 2nd and subsequent ratings separated by a forward slash, and criterion labels abbreviated as: OTF = Overall Task Fulfilment, AoL = Appropriateness of Language, CoS = Comprehension of Stimulus, LF = Linguistic Features, PF = Presentation Features.

Script 77, Grade D

Fair score: 4.11 Raw scores: OTF: 4 / 4, AoL: 5 / 4, CoS: 4 / 5, LF: 4 / 4, PF: 4 / 4

Thank you for seeing Mr Edwards, a 65year old retired accountant who lives at home with his wife The patient presented to me complaining of a progressively enlarging skin lesion on the left lower leg wich he had noticed over the last 10 month. The condition is associated with swelling and erythema over the last two month . The pothiens has been anxious consegneally

Regarding his posst medical history significant is that he was suffering gloucoma in 2005, and pretibial skin lesion was excised in 2008. For gloucoma treatment with timolol maleate 0,25% solution has been included. In his family history nothing significant regardin skin carcinoma, He is not smoker no drinker.

On examination , systemic review is normal, his height is 172 cm and weight is 78 kg. Puls rate is normol and body temperature is 36,7◦C On the left tibial area there is a skin lesion who measures 3 cm , with irregular edges, purulent discharge and ulcerated.

Based on the above I bilieve that the lesion could be SCC.

Advice for the grafts sugery hos been given to the patient vlor - g with reasurance that lotion hool not spreaded . Pleas e mole that treatment nso glucloxacillin has been started

I mould g reatly appreciat your urgent assessmn bnd addmition of this patient and your monagem us well .

Script 54, Grade C

Fair score: 4.78 Raw scores: OTF: 5 / 4, AoL: 5 / 5, CoS: 6 / 5, LF: 5 / 4, PF: 5 / 4

Thank you for seeing Mr. Brian Edwards, a married retired accountant , who I suspect has squamous cell carcinoma and infection. He is non-smoker, non-drinker and lives at home with wife .

Today he presented to my general practice complaining of a progressively enlarging skin lesion on left lower leg for the last twelve months and associated swelling and erythema for the last two weeks. Mr. Edwards was anxious for his condition however he is healthy and independent in activities of daily living. He did not had fever, bone pain, or tenderness. His bowel habits are normal and there is no change in his weight.

On examination, his general examination was unremarkable. His height is 172 cm and weight is 78 kg. Local examination revealed , three centimeter skin lesion on left lower tibial area which has irregular edge , erythema and purulent discharge. However , his systemic examination was normal.

I prescribed him oral flucloxacillin. I had provided counselling for surgery inclu-ding skin grafting and reassurared him that this lesion is probably localised.

He had glaucoma since 2005 for which he is taking Timolol maleate 0.25% solutions , 1 drop twice daily. He had SCC in pretibial area in 2008 which was excised. He did not have any significant family history.

I would be grateful, if you could do follow-up assesment for suspected SCC and associated infection and manage accordingly of Mr. Edwar

Script 38, Grade B

Fair score: 5.00 Raw scores: OTF: 5 / 5, AoL: 5 / 5, CoS: 5 / 6, LF: 5 / 5, PF: 5 / 5

I am writing to refer one of mine patient, Mr. Edwards, to you . He is a 65-year-old retired accountant , married , who presented with symptoms suggestive of Squamous Cell Carcinoma (SCC).

He first visited me today with a twelve months history of enlarging skin lesion on left lower leg . The swelling was associated swelling and erythema.

Examination revealed that the patient was anxious and a 3 cm skin lesion found on left tibial region. In addition, the lesion had irregular edge with ulceration and purulent discharge . Apart from his other findings were unremarkable . He has no known allergies and no family history of carcinoma . Please note that he has a past history of SCC on right pre-tibial skin which was excised in 2008 . He is currently on timolol maleate 0.25% solution , 1 drop twice daily .

On review today , I commenced him on flucloxacillin orally. Mr. Edward was counselled on surgery (including skin grafts) and reassurance was provided that the lesion is most probably localised .

Should you have further queries, please do not hesitate to contact me. I would appreciate your assessment and management .

Script 15, Grade A

Fair score: 5.75 Raw scores: OTF: 6 / 5, AoL: 6 / 6, CoS: 5 / 5, LF: 6 / 5, PF: 6 / 6

I am writing to refer Mr. Brian Edwards, a 65-year-old retired accountant, who is suffering from suspected squamous cell carcinoma for assessment and management.

Today , Mr. Edwards presented with a one year history of progressively enlarging skin lesion on his right lower leg. He also noticed some associated erythema and swelling over the last two weeks. Other than that he is fit and healthy.

On examination, revealed that he has a 3 c.m. size lesion over the left tibial area. It has an ulcerated , irregular edge with purulent discharge and erythema. All other examinations are normal .

It is of note that he has had glaucoma since 2005 . He also had a S.C.C. excised from his right pre-tibial skin in 2008. He does not have any family history of skin carcinoma or major illness.

In view of the above condition, I believe, his condition requires specialist opinion. I counselled him on surgery including skin graft and commenced on oral flucloxacillin for the infection.

I would greatly appreciate your assessment. and management of this patient. Please do not hesitate to contact me , if you need any further information .

Discourse measure: *Structure* at grades A, D (Prompt 2)

Below are two sample scripts chosen to exemplify expected macro-structure in terms of elements included and the sequencing of those elements (in a grade A script), compared with deviation from the expected macro-structure (in a grade D script).

Script 99, Grade A

Fair score: 5.65 Raw scores: OTF: 5 / 6, AoL: 6 / 6, CoS: 5 / 6, LF: 6 / 5, PF: 6 / 5

I am writing to refer Mrs Hong , who is suffering from an acute right lower lobar pneumonia and needs immediate hospitalization and further management . She had rheumatic carditis in childhood which resulted in mitral regurgitation and atril fibrilation and she takes digoxin and warfarin accordingly.

Mrs Hong initially presented to me three days ago with a productive cough and fever . Her chest examination revealed scattered rhonchi through her chest along with an apical pansystolic murmur. She was also feverish (38˚ C). I diagnosed an acute bronchitis and ordered Amoxicillin and advised her to discontinue smoking.

Two days later she returned with an increase in her cough and reported having a thick yellow phlegm. The temperature was increased to 38.5˚ C and mild crackles were audible in the posterior side of her right lung base. I assessed her condition possibly an early right basal pneumonia and ordered sputum sample culture and microscopy, and chest x-ray , in addition to full blood test. I recommended that she should begin chest physiotherapy as well. Fortunately she had quited smoking since previous visit.

Today she was brought in with worsening of the symptoms. She developed a right-sided pleuritic chest pain and was dyspnoeic and febrile. Examination indicated dullness in the right lower lobe in percussion accompanied by overlying crackles. I have attached her test results, however , the sputum culture showed an amoxicillin resistant streptococcus pneumoniae.

The grade A script above exemplifies the letter initial sequencing of the element, Diagnosis (together with Purpose) in the opening sentence of the referral letter. This is followed by a prototypical sequencing of Presenting complaint-Findings/Results-Management in the middle paragraphs. The sample script at level D on the other hand (next page), positions Diagnosis in the middle section of the letter (within the second cycle of Presenting complaint-Findings/Results-Management). Note also in the D level sample, Presenting complaint (at

visit number 1), is given in the opening paragraph, and the expected cycle of Presenting complaint-Findings/Results-Management is disrupted by History at paragraph two.

Script 164, Grade D

Fair score: 3.70 Raw scores: OTF: 4 / 3, AoL: 4 / 4, CoS: 4 / 5, LF: 4 / 3, PF: 4 / 4

Thank you for accepting Mrs May Hong , 43-year-old patient. initially, she is presenting in 21. 11. 2012 with productive cough lasting for 3 days, in addition to fever, she was smoker 10 cigarettes daily.

his past medical history nothing significant only sh had history J Mitral Regurgitation and atrial fibrillat Cawsed by Rhematic carditis .

, At that time, there was scattered rhonchi through his chest. Therefore my diagnosis was acute Bronchitis which aggrivated with increased cigarettes smoking, So advised to ceased smoking besaid antibioticAs, amoxycillin 500 mg three time daily besaid continued on here Medications , which he took digoxin 0.125 mg and warfarin 4mg daily , to bc followed after 2 days with check prothrombic ratio to hire Unfortunately , he came after 2 days with worsen symptoms, his cough increase associated with yellow phlegm

At this time , she stopped cigarettes intake . On examination here feverish , his tempture 38.5 c˚, his chest examination there were mild crackles noted especially at Rt. lunge base. posteriorly with towS cattered crackles in bath lunge

So his Bronchitis increased , with early Right Basal pnemonia, there fore sputum sample took from hism for microscopy and culture. , send for chest. X-Ray, prothrombin ratio ordered to be Review for Next day, advice him to do physiotherapy to his chest .

on Review today , Mrs May accompanied by his Son , his Symptom increased and associated with Right pleuritic chest pain, fur and shortnery Breath.

The Resulty prothrombin radio Rednced 2.4 (taget2.5-3.5[] On examination he is look unwell, his Respiration rate increased , his tempture 38c˚ , pulse 110 min. and Rt lower lobe dull in percussion with crackles besaid, left basal crackles present be said pan systolic mumur is louder.

unfortunately , The Result my lab. shows gram + streptococcus which Resistant to Amoxiciilin ,But it sensitive to clarithromycin & enythraycin. CXray Result shows opacity y Right lower lobe and his leucocytion 11.0 x 10 /L Base on above , I am worried about Mrs. May I think she has Rt- lower about pnemonia , I appreciate your further assessment and managment .

please do not hesitate to contact me for further clarification

Discourse samples at grade levels A, B, C, D

The following sample scripts have been included to exemplify a range of levels of quality for selected discourse features.

Script 91, Grade A (Prompt 2)

Fair score: 5.77. Raw scores - OTF: 6 / 6, AoL: 6 / 6, Cos: 6 / 6, LF: 6 / 6, PF: 5 / 6

Thank you for seeing Mrs. Hong , a 43 year-old patient, who is presenting with right lower lobe pneumonia. Her medical history includes Rheumatic carditis complicated with mitral regurgitation and atrial Fibrillation for which she takes digoxin and warafin. She smokes 10 cigarettes a day.

Initially she presented on 21 . 11 . 2012 suffering from acute bronchitis over the last 3 days . Her examination revealed a tired patient with scattered wheezing through the chest, apical pansystolic murmur and a temperature of 38 C˚. Therefore , she was prescribed amoxycillin and advised to stop smoking.

yesterday she visited me again with worsening symptoms as her temperature has increased to 38.5 C˚ , her cough has been worse and she has developed yellow phlegm . As a result, she was requested to undertake some investigations.

On review today , she attended with her son who reported that she had a bad night last night. Her examination showed pleuritic chest pain on the right side , an increased respiratory rate and right lower lobe dullness on percussion associated with wheezing. Moreover, her investigations demonstrated gram positive streptococcus pneumonia resistant to amoxycillin and sensitive to clarimycin and erythromycin, an x-ray Findings consistent with right lower lobe pneumonia and white blood cells count of 11.0 x 109 L.

Based on the above , I am referring her for urgent hospital admission for further assessment and mangement.

Please do not hesitate to contact me if you require any further information

Sample script 91 above, rated at the highest level (grade A), exemplifies for Prompt 2 a number of discourse features at higher levels of quality. The script demonstrates the higher levels of discourse quality for accuracy, fluency (number of words only), syntactic complexity (words and clauses per t-unit), lexical complexity (average word length, lexical sophistication, percentage of AWL words), coherence, cohesion and content (proportion of irrelevant idea units only). Note for some measures, this script does not exemplify the expected level of quality: i.e. numbers of t-units and clauses (fluency), words per clause (syntactic complexity), d-value, lexical density (lexical sophistication) and proportion of required idea units (content).

Script 3, Grade A (Prompt 1)

Fair score: 5.95 Raw scores - OTF: 6 / 6, AoL: 6 / 6, Cos: 6 / 6, LF: 6 / 6, PF: 6 / 6

Thank you for seeing Mrs. Hong , a 43 year-old patient, who is presenting with right lower lobe pneumonia. Her medical history includes Rheumatic carditis complicated with mitral regurgitation and atrial Fibrillation for which she takes digoxin and warafin. She smokes 10 cigarettes a day.

Initially she presented on 21 . 11 . 2012 suffering from acute bronchitis over the last 3 days . Her examination revealed a tired patient with scattered wheezing through the chest, apical pansystolic murmur and a temperature of 38 C˚. Therefore , she was prescribed amoxycillin and advised to stop smoking.

yesterday she visited me again with worsening symptoms as her temperature has increased to 38.5 C˚ , her cough has been worse and she has developed yellow phlegm . As a result, she was requested to undertake some investigations.

On review today , she attended with her son who reported that she had a bad night last night. Her examination showed pleuritic chest pain on the right side , an increased respiratory rate and right lower lobe dullness on percussion associated with wheezing. Moreover, her investigations demonstrated gram positive streptococcus pneumonia resistant to amoxycillin and sensitive to clarimycin and erythromycin, an x-ray Findings consistent with right lower lobe pneumonia and white blood cells count of 11.0 x 109 L.

Based on the above , I am referring her for urgent hospital admission for further assessment and mangement.

Please do not hesitate to contact me if you require any further information

Sample script 3 above, also rated at the highest level (grade A), demonstrates for Prompt 1 the higher levels of quality in terms of accuracy, fluency, syntactic complexity, coherence, cohesion and content. For the measures of lexical sophistication, this script does not exemplify the expected level of quality.

Script 121, Grade B (Prompt 2)

Fair score: 5.42 Raw scores - OTF: 6 / 5, AoL: 6 / 6, Cos: 6 / 6, LF: 5 / 5, PF: 5 / 5

I am writing to refer this 43 year old patient , to you , who has been diagnosed as suffering from Right lobar pneumonia and needs urgent admission in your hospital.

Initially Mrs Hong came on 21/11/2012 with a complaint of productive cough for 3 days associated with mild fevere. However, she denied any dyspnoea or pain. She is a smoker (10 cigarettes/day) and has a history of Rheumatic carditis in the childhood . On examination, she looked tired, her temperature was raised (38˚C) , and there were Scattered rhonchi with an apical pansystolic murmur . Other findings were normal . The possible diagnosis was Acute bronchitis . Therefore, she was advised to cease smoking and prescribed Amoxycillin . Her other medications for carditis remained unchanged . The next visit was planned in 2 days.

In her next visit,on 23112012, her condition deteriorated and there were mild crackles at right lung? base . Investigations were advised including sputum, blood , chest x-ray ad prothrombin ratio tests.

Today, she has been brought by her son, having right sided pleuritic chest pain . There was a dull percussion note with overlying crackles on the right chest. Her x-ray shows an opacity of the righ lower lobe . Other test results has been sent to you .

Base on the above , I would appreciate if you would admit the patient and treat accordingly.

Sample script 121 above, rated at the second highest level (grade B), exemplifies for Prompt 2 a number of discourse features at relatively high levels of quality for each of the features listed below: accuracy, fluency (number of words only), syntactic complexity (words and clauses per t-unit), lexical complexity (average word length, lexical sophistication, percentage of AWL words), coherence, cohesion (number of connectives only) and content (proportion of irrelevant idea units only). Note for some measures, this script does not exemplify the expected level of quality: number of words (fluency), words per clause (syntactic complexity), d-value, lexical density (lexical complexity), referential cohesion (cohesion) and content (proportion of required idea units).

Script 44, Grade B (Prompt 1)

Fair score: 5.53 Raw scores - OTF: 5 / 6, AoL: 6 / 6, Cos: 4 / 5, LF: 6 / 6, PF: 5 / 5

Thank you for seeing this 65 year old retired accountant , who has demonstrated features of Squamus Cell Carcinoma (S.C.C) or an infectious skin lesion. He needs further investigation and follow up assessment.

Mr Edwards noticed this skin lesion one year ago, on the lower part of his left leg. Since then, this lesion has been enlarging and two weeks ago, it developed some swelling as well as erythema. He did not complaine of any bone pain ; however he was anxious about these current changes. On examination , I found a 3 cm skin lesion on the left tibia with irregular edge , associated with some ulceration, erythema and purulent discharge . Other examinations including his temperature were normal.

It is of note that, the patient has a history of S.C.C ,but on his right pre-tibial area , 3 years ago ; for which he was treated with timolol maleate solution (25% , one drop twice daily). His familial history is other wise unremarkable and he has no known allergies.

In view of the above , my differential diagnosis is either an infectious skin lesion or S.C.C. I started oral flucloxacillin and reassured him that the lesion is localised , but I consider that the possibility of skin cancer needs to be ruled out. I would appreciate your further assessment and possibly exision of the lesion.

Sample script 44 above, also rated at the second highest level (grade B), demonstrates for Prompt 1 a relatively high level of quality in terms of fluency, syntactic complexity, lexical complexity (average word length and lexical density only) and cohesion (number of connectives only. Note however, this script shows slightly higher levels of quality for accuracy, coherence, cohesion (referential cohesion) and content (proportion required and irrelevant idea units) than expected for a script at this level.

Script 136, Grade C (Prompt 2)

Fair score: 4.63 Raw scores - OTF: 4 / 5, AoL: 5 / 5, Cos: 5 / 6, LF: 4 / 5, PF: 3/ 4

Thank you for urgent admission of Mrs. Hong , a 43-year-old woman whom I suspect of having Right lober Pneumonia.

Initially ,she presented to me on 21.11.2012 complaining of productive cough over the last 3 day. She was smoking 10 cigarettes a day and had a childhood history of Rheumatic carditis which caused a mitral regurgitation and atrial fibrillation. On examination , she looked tired and feverish (38˚) . upper respiratory systems were clear but I noticed a scattered rhonchi through her chest along with apical pansystolic murmur . with a possibility of Acute bronchitis , I advised her to give up the smoking and added Amoxycillin 500 mg, 3 times a day to her current medication includs digoxin on warfaring . Yesterday , May returned with complaint of severe cough with yellow phlegm while her puls had increased around 92 . A mild crackels at right base of her lung on postrior position was dettected. I diagnosed Bronchitice and ordered sputum sample test as well as Prothrombin ratio test. Chest physiotherapy was asked and I arranged a review for the next day.

Unfortunately, Mrs. Hong was brought to me today , with a severe pain in the chest. Her pulse rate was 110 and blood pressure decreased at 110/75 . On examination, crackles were noticed on the right lung with loud pansystolic murmer in the chest. He M & C test showed the germ is resistant to Amoxicillin . Furthermore , I noticed an opacity in her Right lower lobe. Leukocytois was reported on the test.

Based on the above information, I believe she has a Right lower lobar Pneumonia. I would appreciate if you could urgently assess and treatment this patient in your hospital.

Sample script 136 above, rated at the second lowest level (grade C), exemplifies the following discourse features at relatively low levels of quality: accuracy (percentages of error-free t-units and clauses), complexity (lexical sophistication, percentage of AWL words), cohesion (number of connectives) and content (proportion of irrelevant content). For this sample, discourse quality did not correspond with expectations for level for the following variables: fluency (numbers of words, t-units and clauses), complexity (words per clause, average word length), cohesion (referential cohesion) and content (proportion of required content).

Script 57, Grade C (Prompt 1)

Fair score: 4.72 Raw scores - OTF: 5 / 5, AoL: 5 / 5, Cos: 5 / 5, LF: 5 / 5, PF: 4 / 5

Thank you for seeing Mr Brian Edwards,a 65 year old retired accountant , married , lives at home with his wife .

His past medical history is unremarkable except for glaucoma which has been diagnosed in 2005 , and has been treated with Tim--olol maleate 0.25% Solution, one drop twice daily. He had Squamous Cell carcinoma on his right pre-tibial skin which was removed in 2008. He is non drinker and non smoker . He has no known allergies . His family history is unremarkable.

Today, Mr Edwards came to see me , complai--ning of enlarged skin lesion on lower leg which he had had it for the previous few months. It has been associated with swelling and erythemia for two wee--ks. There has been no pain , tenderness or fever. On examination, the patient looks anxious. His temperature is 36.7◦C with pulse rate of 80/min and regulare. His hight and weight are normal. Examination of leg revealed a 3cm lesion on the left tibia area which is irregular ulcerated , erythematous wit puruland discharge. Otherwise systemic review and examinations were normal . No change in his bowel habits.

In my openion Mr Edwards is suffering from Squamous cell carcinoma which has been associated with infection and ulceration. I have started oral flucloxacillin for infection and reassured him.

I would appreciate your assessment and management of this patient.

Sample script 57 above, also rated at the second lowest level (grade C), demonstrates for Prompt 1 a relatively low level of quality in terms of accuracy, syntactic complexity (numbers of words per clause and t-unit), cohesion and content (proportion of irrelevant idea units). Note however, this script shows slightly higher levels of quality for fluency, lexical complexity, coherence and content (proportion required idea units) than expected for a script at this level.

Script 165, Grade D (Prompt 2)

Fair score: 3.84 Raw scores - OTF: 3 / 4, AoL: 3 / 4, Cos: 4/ 4, LF: 4 / 4, PF: 4 / 4

Thank you for admitting Mrs. May Hong a 43 year old lady. Mrs Hong was presented to me with a cute bronchitis three days ago which was

Treated with Amoxycillin 500 mg Tds , and advice smoking . yesterday I saw Mrs hong with bronchitis and early R basul pneum[]

Today Mrs Hong presented to me Tachy pnoeic and R chest pain [] [..] chest x My and Blood lest. [..] the result of urine and chest x My and blood test with ha file with Hong

my assessment R lower lobai pneumonu

please admit to hospital for managemt [..] result

Sample script 165 above, rated at the lowest level (grade D), exemplifies the following discourse features at the lower levels of quality: accuracy, fluency, complexity (d-value, average word length, lexical sophistication, percentage of AWL words), coherence and content (proportion of required idea units). For this sample, discourse quality did not correspond with expectations for level for the following variables: complexity (words per clause, lexical density) and cohesion.

Script 78, Grade D (Prompt 1)

Fair score: 4.05 Raw scores - OTF: 4 / 5, AoL: 4 / 5, Cos: 4 / 5, LF: 3 / 5, PF: 3 / 4

I would like to refer my patien , Mr. Brian Edwards , a 65 year old retired accountant, who has suspicious 3 cm skin lesion on his left tibia.

Today , he presents with a 12 month history of progressive skin lesion on his left leg , associated with a 2 week s welling and erythema . Apart of this lesion he has not any complaint . He is anxious . On examition, he is alert and orientated. His vital signs are normal. He weight is 78 kg. I found a 3 cm lesion on his left tibia with an irregular edge a long with erythema and purulent discharge . Also the lesion on has un ulceration . The systemic examination is normal.

I suspect that , he may have squamous cell carcinoma or SCC.

I discussed that , the skin lesion most probably localised. I prescribed flucloxacillin.

He is married and lievs with his wife . Mr . Edwards had Glaucoma in 2005 . He had a squimous cell carcinoma pre tibial lesion excision in 2008 He takes timolol meleate 0,25% solution one drop twice daily He has no known allergies He has no Ramily history of any cancer .

I would appreciate it if you could assess this patient with a view of surgery.

Sample script 78 above, also rated at the lowest level (grade D), demonstrates for Prompt 1 a low level of quality in terms of syntactic complexity, lexical complexity (average word length and percentage of AWL words), cohesion (number of connectives) and content (proportion of irrelevant idea units). Note this script shows slightly higher levels of quality for accuracy, fluency, lexical complexity (d-value, lexical density and sophistication), coherence, cohesion (referential cohesion) and content (proportion required idea units) than expected for a script at this level.