Establishing difficulty level consistency across texts in exams at four levels using lexical profiling

Dr Daniel Waller University of Central Lancashire, Britain

DWaller@uclan.ac.uk,

Tania Horak University of Central Lancashire, Britain

THorak@uclan.ac.uk

& Sharon Tyrell

Abstract:

This research looked at investigating the consistency of difficulty across the four levels of exams (B1, B2, C1, C2 as per the CEFR) which have been developed by the exams team at the University of Central Lancashire (UCLan). This was completed using two measures of difficulty: readability and word frequency i.e. the research undertook lexical profiling. Results were compared to similar results from Cambridge ESOL exams (Khalifa and Weir 2007) at the equivalent CEFR levels. In addition, this research project aimed to create a database to use as the basis for future benchmarking to improve future exam production. In this initial stage of the project, only the Reading sections of the exams were analysed. Practical implications for exam development conclude the article.

Keywords: lexical profiling, readability, language testing, word frequency, CEFR

Introduction:

The exam team based at the University of Central Lancashire (UCLan) produces EFL exams at 4 levels of the CEFR: B1, B2, C1 and C2. The exams comprise five sections testing the four skills plus a Use of English paper. In order to increase the rigour of the exam production procedure, a research project was instigated to investigate the consistency of the level of difficulty of the texts used in these papers. To date the extensive experience of the exam production team was drawn on to evaluate the difficulty levels, but an empirical basis to complement this intuitive approach was felt to be necessary to enhance exam quality.

While the Common European Framework of Reference (CEFR) has attempted to assist language teaching professionals in clarifying what language ability at different levels means, it is constrained by not being language specific. Thus, the CEFR is not able to provide details of what the language ability at each level may specifically consist of either in terms of language or genre and discourse type. A second problem is that the descriptors in the CEFR (Trim in Green, 2012) were derived from scales, which while widely used, were based on teacher perception, not empirically based (Fulcher, 2010). Thus, we cannot be sure what the progression through levels of ability truly looks like. Both of these issues have implications for test/exam production: if we do not have clear profiles of what we expect students to be able to achieve, we cannot produce precise criteria. Additionally, if we do not have an adequate method for identifying the types of texts and textual features which learners should be exposed to and be able to cope with at different levels, then we cannot be sure that our tests are fair or that they are sufficiently challenging for the intended test-taking population. The CEFR also requires users to consider for themselves what types of texts learners need to negotiate in their context and what the appropriate reading skills might be for these learners. Thus we need other ways to evaluate what an increase in ability up the scales looks like. Consequently, our research project was instigated to tackle this concern.

There were two principle aims to this research. Firstly, we wanted to investigate the consistency across the four levels of exams offered by the exams team (i.e. B1 to C2). We needed reassurance that the difficulty of texts within any one level lay within reasonable boundaries and then also that the difficulty of texts rose appropriately through the levels. In addition, we aimed to create a database to use as the basis for future benchmarking to improve exam production. Thereby, with the EALTA Guidelines for Good Practice (EALTA 2006) in mind, we aimed to specifically address the general area of Quality Control, and apply the outcomes of our investigation to enhance item writing. The aim of this paper is to describe the process undertaken to facilitate such analysis for other exam papers.

Text difficulty and how to evaluate this has been a topic of interest for some time in the fields of applied linguistics and psychology amongst others. Therefore, some of the background to this and considerations we took into account in our project will be set out before describing the research we undertook on this project.

Readability as a measure of text difficulty

As Bailin and Grafstein (2001) comment, rhetoric, that is how language is conveyed clearly to maximise communicative purpose, is an age-old area of study but only more recently, since the 1920’s (Klare 1952: 285), have approaches with an ‘emphasis of quantification’ to analysing text specifically for readability been employed. A variety of formulae have been devised; Klare (1963 cited in Laroche 1979) reported the use of 29 readability formulae alone between 1923 and 1959. Such attempts to find formulae have resulted in the Flesch, Dale-Chall, FOG and other formulae, which have been the most widely used. More recently Lexile, a software programme developed by MetaMetrics (Bailin and Grafstein 2001) and formulae utilising the Coh-Metrix programme (Crossley et al 2011) have been developed, the latter being one of the few designed with L2 readers in mind. Some of the early formulae are still widely used.

Readability indices are used on texts to check information accessibility in various areas: the military, the courts and government (Stockmeyer 2009) and maybe most widely by health professions to check clarity for public users (e.g. see Wang, Miller, Schmitt and Wen, 2013) but also in business to enhance market information delivery (e.g. the wine industry - see Mills and Pitt, 2012). Within the field of education readability formulae have most widely been applied to graded readers and assessment.

Formulae are commonly based on the two precepts of word difficulty and sentence complexity. Establishing word difficulty is problematic in that word frequency lists, on which such ratings are most commonly based, are rarely up to date. New words continually enter the language e.g. words associated with computing and the internet in particular have entered English within the last 15 years and are now widely used although they may not appear as high frequency entries on word lists which the formulae draw on. Nevertheless, although not updated, as word lists are empirically grounded being based on actual corpora, e.g. the British National Corpus (BNC) they are preferential to intuitively derived word lists. Evidence also suggests that once corpora reach a certain size there is some stability in terms of the items within them, albeit these words may well demonstrate considerable differences in their ranking in terms of frequency (Nation, 2004).

Then there is the fact that word frequency patterns will not be homogenous across generations, or socio-economic groups (Bailin and Grafstein 2001), or regional or professional groups of speakers of any one language. Frequency measures are thus necessarily broad-sweeping. Consequently, these ratings for words provided in K-lists, used in order to establish difficulty in terms of most/ least frequently used words, have some limitations.

Another assumption underpinning various formulae, including Flesch, is that the longer the word, the more difficult it is (Bailin and Grafstein 2001). This is based on the concept that cognitively the more 'units' in a word to process there are, the more difficult the processing will be. Average word length is typically calculated by number of syllables for 100 words of text. As Bailin and Grafstein (2001) note however, length can in fact aid comprehension when the composition of certain words are the result of affixations which may hold clues to meaning (e.g. negative form –' un', or part of speech - 'ly'). Research by Randall (1988, cited in (Bailin and Grafstein 2001: 290) suggests 'comprehension of complex agents has nothing to do with word length'.

The second aspect of reading difficulty which the readability formulae consist of is sentence complexity, which is equally difficult to adequately quantify. Generally an underlying assumption with the readability formulae is that average sentence length correlates to sentence comprehension difficulty. However, Bailin and Grafstein (2001) again argue that this is not necessarily so. For instance, the presence of co-ordinating conjunctions and other connectors may lengthen a sentence but by doing so they clarify the sentence. Khalifa & Weir (2009:118) concur: 'if the language used is elliptical and the lexis used is highly colloquial, short simple sentences may actually be harder to understand than longer sentences'. It would seem then that for formulae to provide more informative output, other measures of syntactic complexity are needed. In their absence, sentence length is relied on as a general guide.

The percentage of words in a text which readers need to know in order to adequately comprehend it has been disputed by a number of writers. A key problem is the issue of what it means to 'know' a word. A learner might recognise a word like 'set' in terms of its form and know one or two uses, but the word has over thirty entries in the dictionary. Nation (2000) provides a table exploring different aspects of form, meaning and use which need to be in place if it is to be claimed that a word is known and identifies that there is a difference between receptive knowledge and the ability to use the word correctly. There is also the problem set out above of the fact that corpora date rapidly once they cease to be expanded. With these provisos, authors have proposed different percentages for the amount of coverage needed (i.e. the words known) in order for a learner to comprehend a text.

Laufer (1988) found that knowledge of the first 5,000 most common word families (K5 level) would allow a learner to comprehend 95% of a text. Nation & Gu (2007), on the other hand, claimed that knowledge of 8-9,000 word families would cover 98% of texts. Khalifa and Weir (2009) however in their research worked with 97% coverage. It is not made clear in Khalifa & Weir's reporting of this research why the 97% level was used but it can be surmised that a 95% coverage may appear unacceptably low, and that at 98% gaining adequate results may have been difficult.

Although word frequency is a convenient measure for text comprehensibility it is not without problems. Firstly, a word like cohabitation might appear quite difficult as it appears in K12 of the BNC wordlists, however, it is constructed by relatively transparent morphology. By way of contrast, a word like 'fix', which is in K2 and highly frequent can have many different meanings ('fix something, get a fix, be in a fix' etc.).

So frequency is not a measure of transparency or of easiness. It is therefore possible to get cognitively challenging texts which measure as easier in terms of high-frequency lexis than relatively straightforward texts which may contain features which are low frequency or 'off-list' such as proper nouns. 'Off-list' words are those words which are not present in frequency word lists either because the list omits certain features such as proper nouns (e.g. Beijing), or else the word is not frequent enough to feature on the words lists (e.g. if the word is in the 30,000th word list then it would register as 'off-list' on a frequency word list which only went as high as K20. Finally, as pointed out above, if the corpus has dated, then a new word such as 'selfie' will not exist in the corpus on which the word lists were based. The assessment of the conceptual difficulty of a text is where the judgement of test-developers comes into play, but word frequency can provide an additional tool by which to gauge text difficulty.

Bailin and Grafstein (2001:292) propose that 'there is no single, simple measure of readability' since text accessibility for any individual is based on a wide range of inter-relating factors. For the exams under study other features of text which may aid comprehension such as use of different fonts, or font sizes, and illustrations are eliminated as all text is presented in the same font and without illustrations. Whether this makes them unjustifiably inauthentic is another argument for another piece of research, but constitutes one less aspect which may contribute to variance. Other reader-specific factors such as reader schemata for example will affect comprehension but it would seem impossible to factor such an individual trait into any formula to be adopted. As Fulcher (1997) reminds us, sentence length and word frequency are not the factors in themselves which make a text easy or difficult. Nevertheless, they are predictors. This together with the considerations mentioned above make it clear that an accurate evaluation of word difficulty would not in all likelihood be feasible. To be rendered workable, some key features of the actual individual differences in difficulty have to be ignored. Therefore, in order to create some form of usable measure we have to accept that there will be certain limitations on the validity as it cannot entirely reflect the reality of all the elements which contribute to text difficulty.

As stated previously, most of the readability formulae were devised with native speakers' use of texts in mind. Further factors other than those mentioned above which contribute to L2 learners' text comprehension inevitably abound. For example, in the case of word difficulty, cognates can aid, and in turn false cognates can cloud, comprehension enormously (Laroche 1979). Laufer (1997, cited in Khalifa & Weir 2009) cites other factors such as regularity of spelling, amount of polysemy, and morphological/ phonological complexity, amongst others.

The use of the traditional readability formulae to evaluate text difficulty for L2 readers has been criticised (e.g. see Brown's 1998 study of Japanese learners). However research by Greenfield (2004) contradicted Brown's findings. As Crossley et al (2011:98) state: 'more research is needed to develop formulas that contain more linguistic features and that better match text readability for various genres, readers and levels'.

Methodology:

To break the project down into manageable stages, just the Reading section of the exams was initially analysed and this is what will be reported on here. Reading was chosen as this is obviously the most 'text-heavy' section. All exam sections (except Speaking) were prepared however and will be analysed in later stages of the project.

The tests under examination are a suite of proficiency tests administered at the CEFR levels B1, B2, C1 and C2. Text length varies from around 400 words per text at B1 to 600 words at C2. The reading section at each level contains two texts. At the B-levels there are a mixture of heading matching items and multiple choice items, while the C-levels use multiple choice items only.

Data preparation:

In order to prepare the material for analysis, a series of four papers from across the suite, at the four levels mentioned above were 'dissected' into the various sections of each exam, i.e.

Reading, Use of English and Listening. The writing section was not analysed at this stage of the project partly due to the minimal text content. Once a paper had been divided into sections, these were then further divided into their different parts (see table 1). A further sub-division was to divide the parts according to the permutations, namely text *with* the questions and text *without* the questions. This was to further enhance the quality control to evaluate whether the language of the questions was appropriate for the level of the paper. In other words, within each part each text was treated as a separate entity for analysis purposes, so within the Reading section there would typically be two texts per paper for analysis. Standard rubric which was identical across papers was not included.

Paper	**Sections**	**Parts**	**Components**
B2 Level Paper	Listening	Listening Part 1	Transcript
			Questions
		Listening Part 2 A	Transcript
			Questions
		Listening Part 2 B	Transcript
			Questions
	Reading	Reading Part 1	Text
			Questions
		Reading Part 2	Text
			Questions
	Use of English	UoE Part 1	Questions
		UoE Part 2	Questions
		UoE Part 3	Text
			Questions
		UoE Part 4	Questions
		UoE Part 5	Questions

Table 1: Initial division of papers by section, part and component

In terms of text manipulation for the process of analysis, extra-textual numbers written in figure form (e.g. question numbers) were omitted as this distorted the word frequency counts since Lextutor VocabProfiler, the analysis software, counted all these instances. Where they were intra-textual they remained but were converted to text form (e.g. 3 to three) where necessary. Finally the texts, which are produced in Microsoft Word, were then converted to basic Word files text files for the analysis, as required by the software (see below).

The relevance of examining ‘off -list’ words is that their presence may skew the K levels. For example, if a text has a disproportionate amount of off-list words it may appear more difficult than it may in fact be. The ‘off-list’ includes proper nouns so the repeated use of proper nouns could misrepresent the lexical profiling, for example, texts and questions might repeatedly use a name (e.g. Beijing or Susan) which would then count as an ‘off-list’ word and this might give the impression that the questions (which often refer to individuals explicitly by name so as to prevent them from being ambiguous) were harder than the actual text. In order to limit this effect, repeated proper nouns were changed after their second instance to a suitable pronoun, the assumption being that a reader or listener would understand to whom the text was referring. All such changes were recorded in the data preparation log and profiling was undertaken both with and without the change so as to allow for comparison. Table 2 below gives an example of how this was carried out for one section.

Paper	**Sections**	**Parts**	**Components**	**Permutations for analysis**
B2 Level Paper	Reading	Reading Part 1	Text Questions	• Text without questions • Text and questions • Text with repeated proper nouns converted to pronouns • Text and questions with repeated proper nouns converted to pronouns

Table 2: Example of permutations for analysis

Data Analysis:

The analysis comprised three stages: firstly, obtaining readability scores for all texts; secondly, obtaining scores for word frequency range of all texts and finally a comparison of similar measures for Cambridge ESOL exams from B1 to C2 was made.

Readability

The readability formula mentioned above devised by Crossley et al (2011) based on Coh-Metrix is very promising, being devised specifically for evaluating L2 comprehensibility, and taking into consideration the flaws of more traditional formulae. However as the authors state, the work reported to date was based on academic texts alone. Since our aims for this project concern exams of general English ability, not of specific genres, this index was not deemed suitable at this stage.

Crossley et al (2011) also state that while traditional formulae were less accurate than the ones they themselves had devised, they were indeed useful at discriminating texts' difficulty. Therefore selecting from the established, traditional formulae seemed the best approach for this project, despite the inconclusive research results into the effectiveness of traditional formulae for L2 users. In the absence of a significantly superior, accessible, easily usable formula based on such research, we thus decided to use the Flesh Reading Ease and Flesh-Kincaid Grade Level formulae, with some reservations and fully aware of their faults and inadequacies. As Masi (2002) says, they have been used recently and widely as indicators of text difficulty in various studies.

These two Readability Tools are available with Microsoft Word software. The Flesch Reading Ease score is computed from the average number of syllables per word and the number of words per sentence and the result is presented in percentage form. The higher the score, the more readable (i.e. the 'easier') the text is. A score of 90.0–100.0 means the text would be easily understood by an average 11-year-old student, a score of 60.0–70.0 means the text would be easily understood by 13- to 15-year-old students, and lastly, 0.0–30.0 means the text would be best understood by university graduates (Flesch, 1979).

The second tool used to establish a readability index was the Flesch Kinkaid Grade Level score. The higher the score, the more difficult the text is deemed to be, in contrast to the Flesch Reading Ease score, explained above. The score (1-10) indicates an estimation of the number of years of education required to understand a certain text, therefore, for example, a score of 8.2 = 8th grade (usually 12–14 years old in the USA). It is also calculated on the basis of number of syllable and words (see endnote 1).

Thus the sentence: 'Education is an admirable thing, but it is well to remember from time to time that nothing that is worth knowing can be taught' gains a Flesch Reading Ease score of 62.6%, Flesch Kincaid Grade Level score of 10.4 having 24 words and 34 syllables. Equally, The popular children's story book, The Very Hungry Caterpillar by Eric Carle consists of 232 Words, 280 syllables which creates a readability score of 83.2% and a Flesch Kincaid Grade Level of 4.4.

Word Frequency

The second feature to be analysed, namely word frequency, was undertaken using the Compleat Lexical Tutor software version 6.2 which is a vocabulary profiler tool (http://www.lextutor.ca/). At the time, this software provided a comparison of the inputted text against texts of the British National Corpus (BNC) representing a vocabulary profile of K1 to K20 frequency lists, i.e. the first 1000 most frequent word families, up to 20, 000 most frequent words. It provides a frequency profile for texts including 'off-list' words. This refers to words which are not found in the K20 BNC lists, i.e. more obscure words, and will typically consist of proper nouns.

Levels of readability at three rates of coverage were examined: 95%, 97% and 98%, based on the work of previously mentioned studies by Lauffer (1988), Nation and Gu (2007) and Khalifa and Weir (2009). We decided to analyse at all three levels and to then evaluate the results to see which level we found to be most informative and useful.

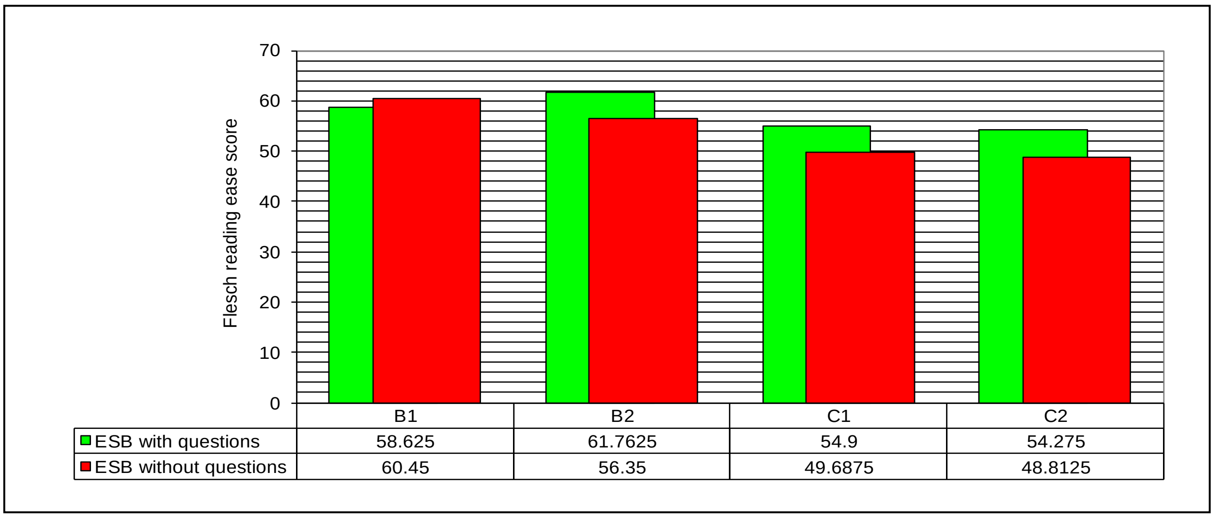
Cambridge Comparison

The next stage was then to compare our results to the results from similar analysis of Cambridge ESOL main suite exams PET, FCE, CAE, CPE (reported in Khalifa & Weir 2009). The rationale for this comparison was that these exams are amongst the few widely available, reputable public exams to have undergone profiling, using similar techniques, and with easily available published results.

Results:

Readability

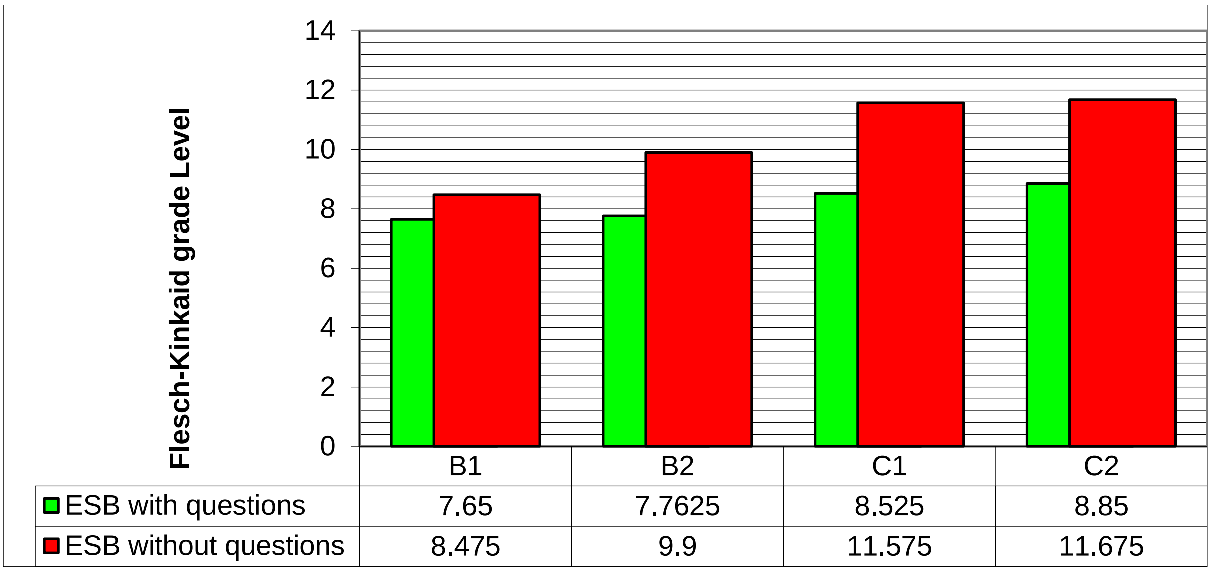
First of all the Reading Ease Scores for texts both with and without the relevant question will be discussed. In Figure 1 it can be seen that the trend is in the right direction, B1 being more easily readable than B2, which in turn is more readable than the C level papers. Unfortunately there is poor differentiation between the C level papers in terms of readability, either with or without the questions. The B2 papers are more readable when questions are included, but this is not the case with the B1 papers, which is not desirable. The language levels of the items should in principle be lower than that of the texts they accompany (Alderson et al 1995).



	B1	B2	C1	C2
🟩 ESB with questions	58.625	61.7625	54.9	54.275
🟥 ESB without questions	60.45	56.35	49.6875	48.8125

Figure 1: Flesch reading ease scores

Figure 2 on the other hand shows slightly better differentiation if we consider the texts without the questions. C1 and C2 however are not differentiated as much as would be hoped. When the texts with their questions are considered, again the trend is in the right direction, but more differentiation between B1 and B2, and between C1 and C2 would be desirable.



	B1	B2	C1	C2
🟩 ESB with questions	7.65	7.7625	8.525	8.85
🟥 ESB without questions	8.475	9.9	11.575	11.675

Figure 2: Flesch Kincaid grade level scores

Another way of examining the grade level score is to consider the range of scores (see Figure 3). When the texts on their own are considered, we can see that there is some differentiation in the difficulty levels between B1 and B2 and between B2 and C1. However, the range of difficulty between C1 and C2 is minimal. When the texts are considered with the questions included in the profile, the range of difficulty between levels is far narrower. The C papers cover a much wider range than the B-level texts, especially the B2 texts which cover the narrowest range of scores. A key observation at this stage is that the differential between the C1 and C2 texts, both with and without questions, is extremely narrow.

It must be noted that the comparison between the texts with questions does require some caution in interpretation due to task effect; there are different task types used at different levels, with the C levels consisting of multiple choice entirely. This is another reason why it was important to consider the texts without questions.

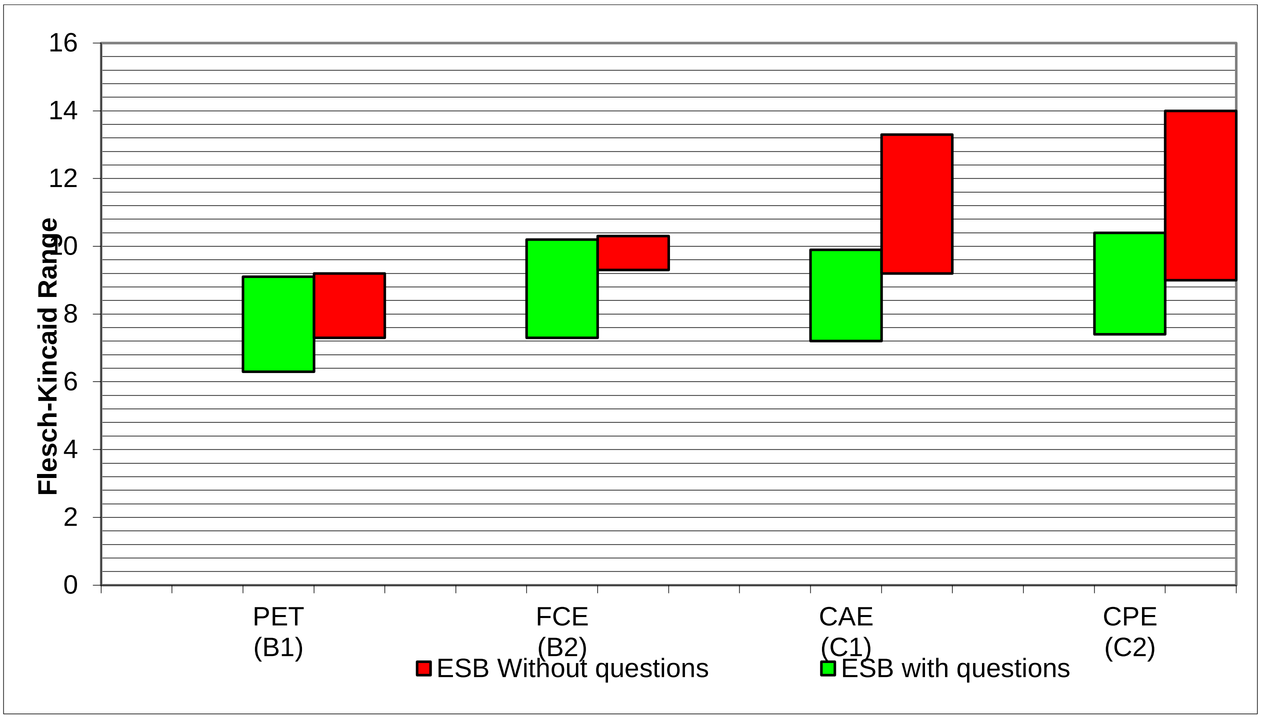


Figure 3: Flesch Kincaid reading scores range

Word Frequency

Next, the vocabulary profiles (from the LexTutor analysis) will be discussed. As discussed earlier the benchmarks of 95% and 98% text coverage were used for the comparisons.

The mean scores across levels indicated some interesting results, as seen in the following tables.

	With Questions (95%)	Without Questions (95%)	With Questions (97%)	Without Questions (97%)	With Questions (98%)	Without Questions (98%)
B1	6.25	5.13	7.00	6.50	9.92	8.63
B2	3.08	3.12	5.16	5.75	6.75	6.88
C1	8.00	8.88	12.75	13.00	15.88	15.38
C2	7.50	7.50	11.50	10.88	15.75	14.38

Table 3: Mean profile of texts by K-level at 95%, 97% and 98% coverage

At level B1 the profile was potentially problematic as the questions seem to be making the text more demanding in terms of frequency. B2 was better, with the text on its own having a slightly higher K-level. However the B1 pattern was repeated at the C2 level wth the questions adding to the K-level of the task. C1 appeared different in that the text had a higher K-level without the questions until the 98% level. It was hypothesised that one reason why the language of the questions tends to raise the levels was probably because less frequent lexis such as proper nouns is often utilised when synonyms for target lexis are chosen by test developers, especially in multiple choice style questions. This is because test-writers are required to ensure that such questions are unambiguous about the identities in the text they are describing.

Given the unwanted nature of the patterns seen in these results, as stated in the methodology, we eliminated the repeated proper nouns (i.e. after the first mention any proper nouns were replaced with appropriate pronouns), to test whether they were exerting a skewing effect as discussed above. The results are shown in Table 4 below.

	With Questions (95%)	Without Questions (95%)	With Questions (97%)	Without Questions (97%)	With Questions (98%)	Without Questions (98%)
B1	3.71	4	4.57	4.2	4.57	4.4
B2	3.33	2.33	4.33	3	5	3.33
C1	4.86	4.5	5.86	7	7.29	7.3
C2	4.57	4.29	5.43	5.00	7	6.71

Table 4: Average vocabulary profiles of texts by level - with repeated noun replaced

The results made little impact in terms of improving the level of difficulty in terms of K-level to the relationship of the text with or without questions and similar patterns from the previous analysis (Table 3) were found. At B1, at 95% coverage the text rated as higher than the text without the questions but this was not continued at 97% or 98%. The changes also resulted in B2 losing its previously positive profile while at C1 the text did rise in difficulty against the text and questions at the 97% and 98% coverage levels.

In terms of comparison between levels some problems were still apparent, namely the trend from B1 to C1 is not even; there is a jump between the B and C levels. This may reflect the very large range of performance within the B2 level as manifested in the range of IELTS scores which fall within B2 (www.ielts.org/cefr). The B2 K-level remained lower than B1 in all cases. In addition, and more worryingly, C2 appears to be easier than C1. As we had previously suspected, based on reviewer feedback, the differentiation of difficulty levels of the C level papers was problematic and these results confirmed this.

As per the caveat previously discussed word frequency is not necessarily a failsafe indicator of difficulty. However, general trends do provide predictors at least. Consequently, the average word frequency levels for the various papers were compared. The higher the K-level, the less frequent (and therefore 'harder') the word is.   Tables 3 and 4 compares the results for each of the four levels of exams at both the 95%, 97% and 98% coverage, both for the texts plus their questions and texts without their accompanying questions.

Cambridge Comparison

Finally, as stated previously, we felt it would be interesting to compare the results from our examinations against the Cambridge exams, at equivalent CEFR levels, since similar research had been undertaken on these exams. The reading ease results (**Error! Reference source not found.**) showed a similar trend to the UCLan exams, in that they increase in difficulty up the levels, but the C levels are not well discriminated from each other. The outcomes of calculating the Flesh-Kincaid grade level scores are similar (see Table 6).

Main Suite Level	Flesch-Kincaid grade level	UCLan Tests Average Flesch-Kincaid grade level	
	Cambridge	With questions	Without questions
B1	7.9	7.65	8.475
B2	8.4	7.7625	9.9
C1	9.6	8.525	11.575
C2	9.9	8.85	11.675

Table 5: Flesch Kincaid grade level scores

Level	Flesch Reading ease score	UCLan Tests Average Flesch Reading ease score	
	Cambridge	With questions	Without questions
B1	64.7	58.625	60.45
B2	66.5	61.7625	56.35
C1	58.4	54.9	49.6875
C2	57.7	54.275	48.8125

Table 6: Reading grade level results compared with Cambridge results

When compared to the Cambridge results (Table 7) it can be seen that the texts we offer lie within a narrower range of difficulty at each level than those of the Cambridge exams. It must be noted that we include a smaller number of texts in our exam since our reading paper on B1 and B2 has two sections and Cambridge B level exams have four sections. This may go some way to accounting for this discrepancy. Our exam is much shorter than the equivalent Cambridge paper and therefore has less variability in the texts. Cambridge exams being longer can draw on harder and easier texts to test different skills and range of ability within each level. The Cambridge exams, again, as well as the UCLan exams do not appear to discriminate effectively between C1 and C2 in this regard.

Main Suite Level	Flesch-Kincaid range	UCLan Average Range	
	Cambridge	With questions	Without questions
B1	5 - 10.1	6.3 - 9.1	7.3 - 9.2
B2	5 - 12.3	7.3 - 10.2	9.3 - 10.3
C1	5.7 - 16	7.2 - 9.9	9.2 -13.3
C2	5.6 - 16.1	7.4 - 10.4	9.0 - 14.0

Table 7: Flesch Kincaid score ranges

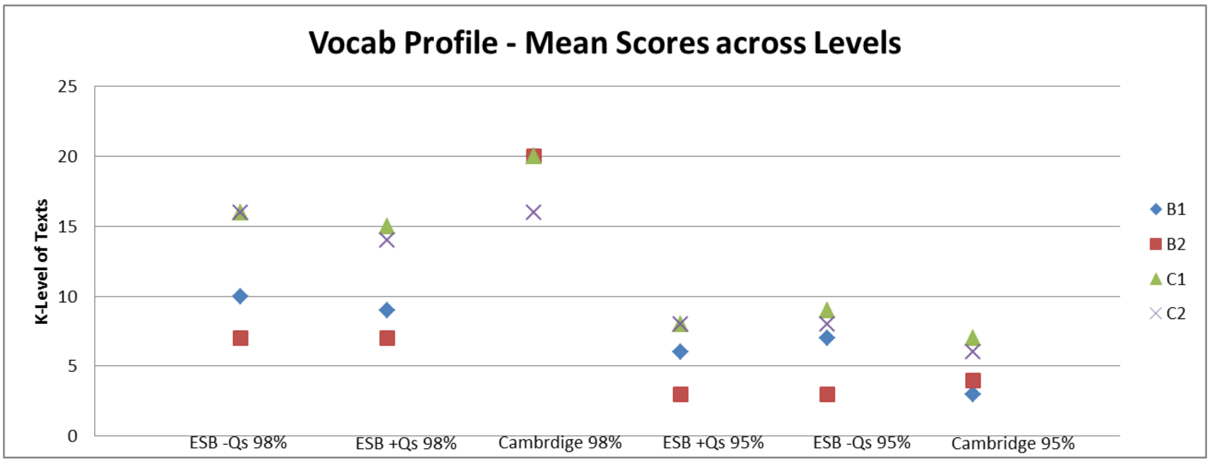


Figure 4: Vocabulary Profile - mean scores across levels - at 95% and 98% coverage - compared with Cambridge results

It appears that from the data provided (Khalifa & Weir 2009) the Cambridge exams overall manifest clustering at the 98% coverage, namely they do not discriminate well as far as lexical profiling is concerned. The C2 exam appears to be ‘easier’ in terms of word difficulty than the other levels at 98% coverage, and C1 appears slightly more ‘difficult’ than C2 at 95% coverage. At 95% coverage the levels are, slightly, discriminated, which they are not at 98%

We have to treat this analysis with caution since we are not sure of the exact methodology used in the Cambridge study to produce their figures, for example whether questions and /or rubrics were included in the word analysis or not.

Two questions for consideration, in light of the differences between the results from Cambridge and UCLan exams is a) what is the effect of Cambridge using more texts than we do? and b) do we filter too much in the exam production process? In other words, in manipulating the language of the exams in particular in the formation of questions, is too much altered in terms of lower-frequency lexis and in the process are the texts ‘downgraded’? This is a clear point for action in future exam production guidelines and training.

To summarise our findings to date, the initial results show that, in terms of readability the UCLan Exam Reading sections appear to demonstrate gradation regarding level of difficulty. The reading sections also appear to be more challenging than those produced by Cambridge ESOL based on readability scores. However, in terms of the lexical frequency results, there are inconsistencies between levels and between texts with and without questions.

Implications:

The results of this research have had several very clear outcomes for the development of the examinations at UCLan.

Firstly, there were lessons for item development as the issue of text manipulation to form questions arose from the data as an area for concern. The main area to be considered is the unintentional raising of difficulty levels through searching for synonyms for key lexis in a target text, in order to write the items. Both guidelines and text checking mechanisms can be put in place to improve this in future.

A second matter regarding text manipulation is to be aware of how the role of proper nouns in a text may affect difficulty. By replacing proper nouns after first mention with pronouns, the level of text difficulty in terms of lexis was reduced but it cannot be assumed that this is necessarily a skill that the target readers would bring to the task. Further research on such text manipulation should be undertaken to investigate whether this assumption was correct.

Another aspect of exam development which we can now pursue is improved discrimination between C1 and C2 levels. Better discrimination between C1 and C2 levels is needed and this will be the topic of further research. If, as appears to be the case, the distinction between C1 and C2 is not necessarily lexical, then issues around other text-features such as discourse mode, cultural assumptions and conceptual difficulty need to be investigated as potential level-distinguishing properties.

The most immediate practical application of the results will be to revise procedures for Item Writers. They will be trained in the use of the analytical tools used in this research so they can ensure that the texts they select are of the appropriate level. They can then also check, once texts have been manipulated to make them suitable for exam papers that they still lie within appropriate level boundaries. Training will be provided and the instructions for the use of the tools have already been produced and pilots for their use undertaken. This should make the exam production process much more effective and efficient. As recognised by Fulcher (1997), decisions regarding text modification for pedagogical purposes are often made primarily on intuitive grounds. We can now better support that intuition to decide whether a certain text is of an appropriate level. While our results found this intuition was generally within acceptable boundaries, a more reliable systematic method to support this is now available to us.

Next stage

In the next stage of the project we plan to analyse the texts used elsewhere in the exams, namely the Use of English sections, Writing exam prompts, and the Speaking prompts will be prepared and analysed for the sake of consistency. Finally the Listening section texts will also be analysed. We recognise it would be unwise to use measures of reading ease on a Listening paper as such measures are devised for texts manifesting features typical of written discourse whereas the Listening section should contain texts typical of spoken discourse. We therefore aim to search for a more appropriate tool to analyse the listening texts.

We also aim to search for a better measure of reading ease, given the reservations we hold about the Flesch and Flesch-Kincaid formulae for L2 use. For example the CohMetrix L2 Reading Index is worth exploring. Any such tool needs to be easily available and readily usable if we are to expect the item writing team to use them routinely as part of exam production.

We could develop this process further, given reservations about the use of the traditional readability formulae, by including a parallel evaluation of texts by expert judges as Fulcher (1997) did, testing for intra-rater reliability of judgments as well as inter-rater reliability, and then comparing to the scores calculated, as per our research.

In the future, such research as described in this paper, should be carried out on a regular basis as part of the quality control measures for UCLan Exams, and we would suggest for any set of exams. The research has been a useful exercise to evaluate current practice and inform item writer development, and contribute to quality assurance measures.

References:

Alderson, J C, Clapham, C & Wall, D (1995) *Language Test Construction and Evaluation*, CUP

Crossley, S A, Allen, D B and McNamara, D S (2011) Text readability and intuitive simplifications: A comparison of readability formulas, *Reading in a Foreign Language*, Vol 23, No1, pp84-101

EALTA (2006) , Guidelines for Good Testing Practice and Assessment, URL: *http://www.ealta.eu.org/documents/archive/guidelines/English.pdf [accessed13.12.20013]*

Flesch, R F, (1949) A new readability yardstick, *Journal of Applied Psychology*, Vol 32 , Issue 3, pp. 221–233

Flesch, R (1979) "How to write plain English." *URL: http://www. mang. canterbury. ac. nz/courseinfo/AcademicWriting/Flesch. htm [accessed 10.12.2013]*

Fulcher, G (1997) Text Difficulty and accessibility: Expert Judgement, *System*, Vol 25, Issue 4, pp 497–513

Fulcher, 2010 (2010) Practical Language Testing. London: Hodder Education

Hiebert, E H (2011) Beyond single readability measures: Using multiple sources of information in establishing text complexity, *Journal of Education*, Vol 191, Issue 2, pp33-42

Khalifa, H and Weir, C J (2009) *Examining Reading – Research and Practice Examining Second Language Reading*, SILT#29, Cambridge: Cambridge University Press

Klare, G (1952) Measures of the readability of written communication: an evaluation, *The Journal of Educational Psychology*, No 7, Vol 43 pp 385-399

Klare, G (1963) *The Measurement of readability*, Iowa State University Press.

Laroche, J M (1979) Readability Measurement for Foreign Language Materials, System, Vol 7, pp131-135

Laufer, B (1997) What's in a word that makes it hard or easy? Some intralexical factors that affect the learning of words, in Schmitt, N (Ed) *Vocabulary: Description, Acquisition, and Pedagogy,* CUP

Laufer, B (1988) What Percentage of Text-Lexis essential for Comprehension? In Lauren, C and Nordmann, W (Eds) *Special Language : From Humans to Machine*, Clevedon: Multilingual Matters, 316 -23.

Mills, A J and Pitt, L (2012) Reading between the vines: analysing the readability of consumer brand wine web sites, *International Journal of Wine Business Research*, Vol 24, No 3, pp 169-182

Nation I S P (2001) *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nation I S P (2004) A study of the most frequent word families in the British National Corpus. In: Bogaards, P and Laufer, B (Eds) *Vocabulary in a Second Language*. John Benjamin, 3-14.

Nation, I S P and Gu, P Y (2007) *Focus on Vocabulary*, Sydney: NCELTR

Stockmeyer, N O (2009) Using Microsoft Word's Readability Programme, *Michigan Bar Journal,* January, pp 46-47

Trim, J L M (2010) Some earlier developments in the description of levels of language proficiency. In Green, A (2012) *Language Functions Revisited: Theoretical and Empirical Bases for Language Construct Definition Across the Ability Range*. Cambridge: Cambridge University Press.

Wang, L-W, Miller, M J, Schmitt, M R.and Wen, F K (2013) Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Research in Social & Administrative Pharmacy,* Vol. 9, Issue 5, pp503-516