Using Readability Formulae for Examination Questions

Final Draft
07/03/2005

Simon Allan
Marie McGhee
Rob van Krieken

SQA
Research and Information Services

Commissioned by the Qualifications and Curriculum Authority.
QCA, 83 Piccadilly, London, W1J 8QA. United Kingdom.

1 *Background*

The purpose of this paper is to consider whether it would be both possible and desirable for QCA to apply a readability measure to GCSE and GCE examinations. It will do this by carrying out a review of the relevant literature available in this area, and considering its relevance to QCA and in particular to examination questions. The paper will also contain a section outlining the use that similar organisations make of readability measures.

The paper begins by considering the existing work in this field, and how it can best be applied to the research question behind this paper, namely, whether it would be appropriate for QCA to advocate using of readability measures as part of the screening of questions in GCSE and GCE examinations. The paper will conclude that much of the existing literature is not directly applicable to this question, and that it is therefore necessary to look more directly at readability measures, and the theory behind them.

It considers the different types of readability measure that are available, and looks at how they are applied. It goes on to consider some theoretical difficulties in applying readability formulae to exam questions. It discusses the variety of different exam question types, and the implications that these have for readability formulae, and concludes that readability formulae can only be of limited use in the setting of examination questions.

2 Readability and Question Difficulty

There is a considerable literature already existing on the role that readability can play in the difficulty of examination questions. A series of papers produced by the University of Cambridge Local Examinations Syndicate (UCLES) looks into this issue, although in general as part of a wider investigation of question difficulty. For example, Pollitt and Ahmed propose a psychological model for answering examination questions in 'Comprehension Failures in Educational Assessment', a paper presented to the European Conference on Educational Research in 2000. The model suggests that the following process for answering examination questions:

- Learning the subject
- Reading the question
- Searching the memory
- Matching question to memory
- Generating an answer
- Writing the answer

The paper looks at a number of context-related issues which impact on a candidate's ability to understand a question, and highlights a number of potential sources of difficulty in examination questions, including putting questions in context, layout and wording. In a 1999 paper, "Curriculum Demands and Question Difficulty", the same authors list a

number of potential sources of difficulty at each stage of the answering process. Some of these, which test the candidate's knowledge of the subject, are legitimate, that is they are part of what the examination is supposed to assess. Others, including whether everyday language is used, are related to readability, whilst others, such as presentation, are neither legitimate nor part of readability as defined by most existing measures.

Mobley, in "Making Ourselves Clearer: Readability in the GCSE", outlines some of the key aspects in making questions a more accurate test of a candidate's ability. The sources of difficulty pointed out are similar to those noted by Pollitt and Ahmed. Pieces of work such as these must be of interest to those involved in producing examinations, and in designing procedures for the setting of exams. However, for this paper, they do not provide us with significant assistance in examining the research question posed. They tell us that readability, and factors connected to it, can form a significant illegitimate source of difficulty in examination questions. What they do not tell us is whether it would be desirable and practical for QCA to consider advocating the use of readability measures by those producing examinations or national tests.

3 Existing Readability Measures

There are two main types of readability measures. The first type considers semantic and syntactic aspects of the text. There is a wealth of formulae of this type; a selection of these will be discussed later in the paper. The second type of procedure, the 'Cloze Test', asks participants to try to predict missing words from the context of a selected passage, and can be used to determine whether an individual has difficulties with reading.

Both types of procedure have valid and recognised merit when assessing the readability of a piece of text; however, the conditions of use imposed by each procedure severely restrict their application to examination questions.

3.1 The Cloze Test

This procedure was first published by Wilson Taylor, University of Illinois, in 1953. Instead of using a specific formula, as had been common in the previously recognised readability measures, the Cloze Test involves individuals filling in missing words within a text; the theory being that the higher the individual's reading ability the greater the success of predicting the missing words. If we know the reading ability of the population taking the Cloze Test, which we could if they were a representative sample of a group such as GCSE/GCE candidates, then the test can then be used to determine the reading difficulty of the text used.

In the original Cloze, the deleted words from the text are taken at regular intervals, and every fifth or tenth word is standard. The reliability of the test is increased the more missing words there are, with a minimum of at least 50 is recommended. Given this, to obtain a reliable outcome from the test, a text of at least 250 words would be required.

The accuracy of the test would also increase if different versions of text were used and

tested by different groups of readers. Each version would have a different sequence of deleted words so that all words within the text were deleted within the complement of versions. The main disadvantage of this approach is that it is significantly more labour intensive than the simpler approach.

The cloze test is likely to be inappropriate for most examination questions due to the minimum length requirement. Used with questions shorter than 250 words, it would give a very unreliable and inaccurate indication of the difficulty a candidate might have in reading the question.

3.2 Readability formulae

There is a variety of different readability formulae which consider aspects of a piece of text, such as word length, sentence length, number of syllables etc. The outcome of these formulae often provides a level of reading age suitable for the text, at which the contents of the passage should not pose problems for the reader.

Some of these formulae incorporate the use of vocabulary frequency lists. One of the first of these lists was compiled to 1921 by Edward Thorndike titled 'Teacher's Word Book'. His book listed 10,000 words by frequency of use which allowed teachers to measure the difficulty of words and text based on the assumption that the knowledge of words is a strong measure of the ability of the reader.

Since then he has published further word lists, as have other authors. Of note is the 1981 publication 'The Living Word Vocabulary: A National Vocabulary Inventory' by Edgar Dale and Joseph O'Rouke. This publication was based on the previous work of Thorndike as well as their own extensive research. It contained 44,000 words, each with grade level score based on the familiarity of the different word meanings.

Another approach is based on research carried out by Gray and Leary in the 1930s, which attempted to discover the key factors that made a text easy to read. In *What Makes a Book Readable?* (1935), they selected a number of texts, and established the level of difficulty of the texts by giving reading comprehension tests to a group of around 800 adults. They then compared various stylistic features of the texts, such as sentence length, the percentage of monosyllabic words, and the number of easy words, and found 17 variables that produced a correlation of greater than 0.35. They then combined five of these variables to create a formula for predicting the readability of a text. The five variables they used were average sentence length, number of different "Hard" words, number of first, second and third person pronouns, percentage of different words and number of prepositional phrases. The resultant formula had a correlation of 0.645 with readability as defined by the comprehension tests.

Gray and Leary recognised that many of the factors which influenced readability were not stylistic, and they categorised these into the three groupings of design, structure, and content. They suggested, however, that it was not possible to measure content, structure, or design statistically, and they therefore concentrated on style as the only area that could

be looked at mathematically. This has been queried more recently but it still remains the basis for most work in readability.

The next section provides details of a number of readability formulae, which work by looking at specific features of a text. This is not an exhaustive list but it covers some of the better-known formulae, as well as some which may prove most effective for use with examination questions.

3.2.1 Flesch- Kincaid Grade Level Formula

This readability formula is an adaptation of the Flesch Reading Ease formula, constructed by Robert Flesch in 1948, which gives grade school level of a piece of text.

The original Flesch Reading Ease formula, below, considers the average sentence length and average number of syllables per word of a 100-word sample.

Score = 206.835 - (1.015 x Average sentence length)
-(84.6 x Average number of syllables per word)

The calculated score lies between 0 and 100, where 0 is very difficult and 100 is very easy. A guide to the outcome of the scores and be found below.

Reading Ease Score	Description
0 to 30	Very Difficult
30 to 50	Difficult
50 to 60	Fairly Difficult
60 to 70	Plain English
70 to 80	Fairly Easy
80 to 90	Easy
90 to 100	Very Easy

The Flesch Reading Ease formula was recalculated for the US Navy in the mid 1970s. This recalculation uses the same variables as the original formula however was made simpler for user and the output is in terms of grade level. The new formula is listed below.

Grade Level = (0.39 x Average sentence length)
+ (11.8 x Average number of syllables per word)
- 15.59

This formula gives a result in terms of US school grade levels, which will not be directly comparable with English school years. A number of other formulae produce results in this form, and the difficulties that this causes are discussed further later on in this paper.

3.2.2 Fog Formula

The fog index was first published by Robert Gunning in the paper 'The Technique of Clear Writing' in 1952. It became a popular formula amongst users due to its simplicity. As with the Flesch Kincaid formula, only two variables are used, which consider sentence length and syllables within the text.

A recommended sample size for use is a text of 100 words. Accuracy improves the more samples of text taken. The formula is stated below.

Grade Level = 0.4 x (Average sentence length + Average number of Hard Words)

Where "Hard Words" are words of more than two syllables.

Given the structure of this formula, short sentences written in plain English achieve a lower score than long sentences written in complicated language.

3.2.3 Homan-Hewitt Formula

This formula was developed in the 1980s by M Hewitt and S Homan. It uses slightly different variables than in previous mentioned formulae to measure vocabulary load and syntactic complexity. The three variables used within this formula are

WUNF – Number of difficult words. The difficulty of words is measured using the 'The Living Word Vocabulary' by Edgar Dale and Joseph O'Rouke has mentioned earlier.

WLON - Number of words with seven or more letters.

WNUM - The average number of words per Hunt's T-Unit. The measure T-Unit is used instead of sentence, as it is believed that this is a more accurate measure of syntactic complexity. A T-Unit was described by Hunt as the 'shortest grammatically allowable sentence into which (writing can be segmented) or minimally terminable unit". Therefore, a simple or complex sentence constitutes a T-unit, a compound sentence has more than one T-unit.

The formula is stated below.

Readability Level = 1.76 + (0.15 x WNUM) + (0.69 x WUNF) – (0.51 x WLON)

The outcome of this formula relates to Grade level; a readability level of 2.0 to 2.9 equates to 2nd Grade; a readability level of 3.0 to 3.9 equates to 3rd Grade, etc.

In 1994 a paper was published in the Journal of Educational Measurement, 'The Development and Validation of a Formula for Measuring Single sentence test item readability' by Homan, Hewitt and Linder. This paper found that Homan-Hewitt

readability formula was successful in predicting question difficulty, although further research was required.

4 Comments on Readability Formulae

Although readability formulae do give an approximate measure of the reading level of a piece of text there are a few points which should be taken into account when using these formulae.

Readability formulae merely give an estimate of difficulty. Different formulae employ different indicators of text difficulty and often result in different outcomes. This is due to the factors in place when formulae are constructed, such as a criterion scores. A criterion text is given a rating or grade level by researchers based on results from reading tests on that text, and then the formula can be calculated using that piece of text. This results in the constructed formula predicting the grade level required to answer correctly a percentage of questions on a reading test based on the criterion passage. The percentage of correct answers can vary between 50% and 100%. Given this variation, and the different variables used in constructing formulae, different formulae can and do provide significantly different estimates of the reading difficulty of a given piece of text.

4.1 Length of Questions

As exam questions are inevitably going to be short when compared with books, essays, or substantive articles, this has to be considered when looking at questions of readability. As can be seen in the sections above on the various readability formulae in common use, a number of them feature frequencies of particular features of writing, such as multi-syllabic and different words.

Applying such formulae to short fragments of text, such as exam questions is going to prove difficult. A readability formula attempts to provide as accurate an approximation to the readability of a text as possible, but is not, in itself, a measure of readability. Depending on the particular features of a very short piece of text, such as the vocabulary used, or the number of length of sentences, the result given by a readability formula could be skewed very significantly. There is no way to know whether such features actually have as large an impact on the readability of the text as they do on the result given by a particular readability formula.

Cloze tests require a minimum length of 250 words to provide a reliable result for the readability of a given piece of text. Given the effort involved in such an approach, compared with using readability formulae, it is likely that QCA would have to use a sampling approach for those questions which met the minimum length requirement, if no other effective measure of readability was found.

Most formulae require the user to sample the text being scrutinised. These samples are typically sections of 100 words, and to ensure that the sample is typical, it would be

advisable to take more than one such sample. Therefore, to apply most formulae to a piece of text, and feel confident in their reliability, it is likely that the minimum length of the text would have to be around 400-500 words.

Earlier it was suggested that the Homan-Hewitt formula is more appropriate for short pieces of text and questions than some of the others mentioned in this article. In a 1994 paper, Homan, Hewitt, and Linder analyse whether the results of the Homan-Hewitt formula correlate to the difficulty level of given examination questions. They found a substantial correlation between higher formula scores and more difficult questions.

It should be clearly noted that their article talks about the difficulty of questions, rather than their readability, and that no effort is made to look at why these questions are difficult. The reasons for their difficulty may be nothing to do with readability, which is an issue which is not addressed in the paper.

It should be clearly noted that success in predicting difficulty of questions using a readability formula does not, of itself, imply that this difficulty is caused directly and only by the increased level of readability. It is at least possible that there is a correlation between the reading difficulty of a text and the difficulty of its concepts, which means that a question about harder concepts would be intrinsically harder. Before it would be possible to say that the Homan-Hewitt formula was an accurate measure for readability, significantly more research would have to be undertaken. This would have to analyse issues such as the reasons that specific questions are difficult, whether questions that deal with more difficult aspects of a syllabus are likely to be less readable and how much variation in formula results can be caused by specific, small changes in the text.

Although this formula does not require samples of 100 words to be taken, the problems surrounding the other formulae seem to apply to the Homan-Hewitt equation as well. The results given by this formula will be affected substantially by minor changes in the text of a specific question, which may have little or no impact on the readability of the question.

4.2 Subject Specific Factors

Many subjects have terms which have particular meanings within that area of knowledge. It is likely that knowledge of some of these subject-specific terms will be required in the curriculum for that subject. For example, candidates in physics are likely to have to know the meanings of words such as ‘inertia’, ‘momentum’, and ‘resistance’. Under a number of readability measures, such words would have a significant impact on the overall score for a particular passage. Any formula which counted the total number of syllables, number of words with three or more syllables, or number of long words would be affected by the use of these words, which would be required by the curriculum for physics. A competent candidate would not, however, be disadvantaged by these words appearing in the exam. A more extreme example would be in a nursing exam, where the word ‘electroencephalogram’ is used. This should not be difficult for most candidates, but would add substantially to syllable and letter counts, thus badly distorting the use of some readability formulae for predicting the difficulty of questions.

Within some subjects, students may be asked to read information from a graph, extract information from a diagram, use a given formula stated in the question, or understand a question which uses subject specific notation. These types of questions are common within Mathematics and science subjects. It is particularly difficult to see how a readability formula can accurately take all these factors into account. This point is effectively illustrated by a paper by Shorrocks-Taylor and Hargreaves, "Measuring the Language Demands of Mathematics Tests: the Case of the Statutory Tests for 11 Year Olds in England and Wales", which finds that the results of readability formulae are significantly at variance with the views of experienced mathematics teachers in determining the reading difficulties of questions from statutory mathematics tests.

4.3 Measuring the Language Demands of Mathematics Tests: the Case of the Statutory Tests for 11 Year Olds in England and Wales

The paper by Shorrocks-Taylor and Hargreaves is particularly important as it is one of the very few published attempts to apply readability measures to examination questions that was found in the course of this piece of research. It applied a number of readability measures to questions taken from the statutory tests in mathematics for 11 year olds, and compared the ratings gained to the results of a survey of experienced mathematics teachers. They found that there was a significant variation between the results of the formulae and the views of the teachers.

The methodology of Shorrocks-Taylor and Hargreaves can be criticised on two grounds. These are:

i. The manner in which readability formulae are applied to the questions.

There are inevitable difficulties in applying readability formulae to examination questions, which are often short, and contain graphs and formulae. Such difficulties would be acknowledged by the authors of the formulae, who would acknowledge that their formulae can only be applied to text of a minimum length, and may not be appropriate to passages containing a large number of graphs and formulae. Nonetheless, if formulae are to be applied to such texts, then the link to the criterion used in their construction and validation is lost, thus undermining the results obtained. Given the structure of exam questions, there is little that can be done to prevent this.

ii. It is unclear whether the teachers in the project managed to distinguish between mathematical difficulty and reading difficulty.

A potentially more serious criticism of Shorrocks-Taylor and Hargreaves' work is that there is no way of knowing whether the teachers involved successfully managed to distinguish between the mathematical difficulty of a question, and the difficulty involved in reading and understanding a question. While there is no easy way to evaluate how successfully this has been done, it is likely that an experienced group of teachers would be able to make such distinctions.

Given that these criticisms are largely unfounded, we find that there are significant correlations between a number of the formulae used in this piece of research, although interestingly, the correlations between many of the formulae and the judgement of the teachers were not significant. It is not immediately clear whether readability formulae should be seen as more accurate than teacher's judgements, or vice versa. However, given the discrepancies between the various methods of assessing readability, this research casts significant doubt on the practical possibilities of applying readability measures to examination questions.

4.4 Formulae Results and Grade Levels

Readability formulae will predict the reading level required to understand the piece of text. What this means in practice is that, for instance, if a formula results in a grade level of 6, then 6th grade students would only just be able to understand that piece of text. Many formulae give their results in terms of American school grades, and we are not aware of any work which looks at comparisons between reading ability in US grades and in UK schools.

As many of the formulae result in a grade level or reading age of text comprehension, the question arises of when and how these grade levels and reading ages were calculated. The results of these formulae are frequently expressed in terms of expected reading levels from that time. No consideration has been given to whether standards of reading, and the requirements to meet those standards, have changed since the formulae were produced.

Without further, recent research into these issues, readability formulae which produce results in US school grade terms are therefore likely to be of very limited use in English schools.

4.5 Question Types

This section describes some of the main question types used in exam situations, some question types are subject specific and each have their own complexities. These questions types discussed are categorised by the question structure rather than the answer format required, as is normally the case.

At the heart of each examination question is the inclusion of a Limited Text Question. Many exam questions use this format as it is, for example

'Describe the economic changes within Britain between 1910 and 1950.'

However, there are variations or extensions to this basic exam question which add additional complexity with regard to the use of readability procedures.

4.5.1 Limited Text Questions

In most cases this is the format used within examination questions; short and direct with use of common verbs which are frequently used, i.e. Analyse, Describe, and Compare etc.

The levels of expectation to the extent of answer from questions using these verbs vary from subject to subject. A 'describe' question in history may require an essay while in biology it may only require a short paragraph. Students will be taught exam technique within a particular subject to allow them to distinguish the level of complexity of answer required.

4.5.2 Questions supported by Source Material or Substantial Text

Within some subjects limited text questions are supported with a large piece of text such as source material or case studies. The student will be expected to read, understand, and extract relevant information from this text to answer the exam question. This additional reading adds to the demands of the examination situation. Readability formulae could reasonably be applied to such texts to ensure that students are not being required to read a piece of text which is beyond the reading skills that they could conceivably be expected to have.

4.5.3 Multiple Choice Questions

Multiple-choice questions have more complex question structures than basic limited text questions. In addition to a basic question, optional answers are given.

The name of the organ which pumps blood round the body is

a) Lungs
b) Heart
c) Liver
d) Kidneys

Each complete sentence option must be considered to consider fully the reading involved for the student tackling a multiple-choice question,. Therefore the question above should be considered as having a length of 4 sentences.

5 Use of readability formulas with items in other organisations

Several colleagues involved in research or production of examinations were asked if their organisation has ever contemplated the use of readability formulas when screening raw questions, not only in languages, but in other subjects as well. They were

New Zealand Qualifications Authority
Board of studies in New South Wales, Australia
Secondary Schools Assessment Board of South Australia
Victorian Curriculum and Assessment Authority
Hong Kong Examinations and Assessment Authority
Educational Testing Service
Citogroep, the Netherlands
Marita Moll, assessment expert in Canada

The timing of the request was not ideal, being close to the Christmas break. The organisations round the Pacific were also busy reporting the results of their examinations Only ETS reported any use of readability formulas. This was restricted to texts accompanying items, and did not extend to the items themselves.

The major UK examination boards were asked the same question. All boards responded that they have screening procedures in place to ensure that questions can be read without difficulty. They also discuss difficulties experienced by candidates with special needs specialists, in order to inform the screening of questions as well as provide well-judged adaptations. None of the boards however, used readability formulas to check the reading level of questions. Some staff were well aware of these formulas and of recent literature in this field, but had never considered a general application of them.

6 Conclusions

There are significant elements in the theoretical framework which would cast doubt on the effectiveness of using readability formulae on exam questions. These can be summed up as follows:

- Questions are likely to be too short to obtain reliable readability results.
- There are difficulties applying readability formulae to questions with graphs, diagrams and formulae.
- Subject-based language is not taken into consideration in readability formulae, especially where specific vocabulary is part of the knowledge required for a subject.
- The results given by readability measures may be difficult to apply to a UK context, if they are based on US school grades or reading ages.

This view is supported by the evidence produced by Shorrock-Taylor and Hargreaves in their paper on statutory mathematics tests for 11 year olds in England & Wales. It is

further bolstered by the fact that other organisations involved in similar activities to QCA do not generally make use of readability formulae in setting examination papers

On the theory that is currently available, it is recommended that readability formulae could only be applied to questions which consist of a substantial quantity of text. It is suggested that a suitable minimum would be around 400 words.

It is important to note at this point that texts (or long questions) should not be written to fit specific readability formulae. Research has been carried out which suggests that, in such cases, the formulae become increasingly inaccurate, and that such strategies do not significantly improve readability.

A number of typographical issues can impact on a candidate's ease of understanding of a given question. These are outlined in the relevant literature, and include

- Contrast
- Type size
- Spacing
- Type face
- Justification
- Blank space

Another specific issue that impacts on how easy a question is to understand is the candidate's familiarity with the context of the question. Examples taken from 'real life' can be substantially more difficult for candidates who are not familiar with the circumstances described because of a different social, regional, religious or gender background.

A particular point that should be noted is that readability and readability indices do not necessarily assist candidates with special educational needs. The fact that a document is more readable is not necessarily of assistance to a candidate with specific special needs, and QCA may be better to consult with relevant groups representing such candidates to determine what changes would be particularly helpful in this area.

BIBLIOGRAPHY

Ahmed, A. & Pollitt, A. 1999 *Curriculum Demands and Question Difficulty*, IAEA Conference Paper

Dubay, W. H. 2004 *The Principles of Readability*, Impact Information

Fisher-Hoch, H. & Hughes, S. 1996 *What Makes Mathematics Exam Questions Difficult?* BERA Conference Paper

Flesch, R. 1948 *A New Readability Yardstick*, Journal of Applied Psychology 32: 221-233

Gray, W. S. & Leary, B 1935 *What Makes a Book Readable?* Chicago University Press

Gunning, R. 1952 *The Technique of Clear Writing* McGraw-Hill

Homan, S., Hewitt, M. & Linder, J. 1994 *The Development and Validation of a Formula for Measuring Single-Sentence Test Item Readability*, Journal for Education Measurement Vol 31, No 4 pp349-358

Pollitt, A. & Ahmed, A. 2000 *Comprehension Failures in Educational Assessment*, ECER Conference Paper

Shorrocks-Taylor, D. & Hargreaves, M. 2000 *Measuring the Language Demands of Mathematics Tests: the Case of the Statutory Tests for 11 Year Olds in England and Wales* Assessment in Education Vol 7, No 1 pp39-60

Taylor, W. 1953 *Cloze Procedure: A New Tool for Measuring Readability*, Jornalism Quarterly 30:415-433

Thorndike, E.L. 1921 *The Teacher's Word Book*, Bureau of Publications, Teachers College, Columbia University

Thomson, S. & Thurlow, M. 2002 *Universally Designed Assessments: Better Tests for Everyone!* NCEO Policy Directions No.14