DOCUMENT RESUME

ED 353 790 FL 020 261

AUTHOR Ferguson, Gibson; Maclean, Joan
TITLE Assessing the Readability of Medical Journal Articles: An Analysis of Teacher Judgements.
REPORT NO ISSN-0959-2253
PUB DATE 91
NOTE 15p.; For serial issue in which this paper appears, see FL 020 251.
PUB TYPE Reports - Research/Technical (143) -- Journal Articles (080)
JOURNAL CIT Edinburgh Working Papers in Linguistics; n2 p112-125 1991

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Difficulty Level; *English (Second Language); Foreign Countries; Interrater Reliability; *Medical Research; *Readability; *Scholarly Journals; Statistical Analysis; Syntax; Vocabulary

ABSTRACT

This study is the first stage of a wider enquiry into alternative ways of assessing the readability of specialist texts. The interest in assessing these texts arose from the need to grade 60 medical journal articles for an individualized English-as-a-Foreign-Language reading scheme for doctors. The study reports on an investigation of subjective judgements of difficulty by "expert" raters. This involved the identification of possible components of difficulty and their independent assessment and scoring by five raters. Subsequent analysis focused on the structure and reliability of these judgements. Preliminary results of the data analysis indicate that four out of the seven components possessed satisfactory levels of inter-rater reliability and that syntactic and lexical difficulty as assessed by the raters may be the best predictors of overall difficulty. Finally, there is statistical evidence that the putative judgement by the raters of seven discrete components may be more adequately modelled as the assessment of two "latent" components of difficulty. (Author/JL)

ASSESSING THE READABILITY OF MEDICAL JOURNAL ARTICLES: AN ANALYSIS OF TEACHER JUDGEMENTS

Gibson Ferguson and Joan Maclean (IALS)

Abstract

This paper is the first stage of a wider enquiry into alternative ways of assessing the readability of specialist texts. In it, we report an investigation of subjective judgements of difficulty by 'expert' raters. This involved the identification of possible components of difficulty and their independent assessment and scoring by five raters. Subsequent analysis focused on the structure and reliability of these judgements. Preliminary results of the data analysis indicate that four out of the seven components possessed satisfactory levels of inter-rater reliability, and that syntactic and lexical difficulty as assessed by the raters may be the best predictors of overall difficulty. Finally, there is statistical evidence that the putative judgement by the raters of seven discrete components may be more adequately modelled as the assessment of two 'latent' components of difficulty.

1. Introduction

Our interest in assessing the readability of medical journal articles arose from a practical concern. We needed to grade 60 medical journal articles for an individualized EFL reading scheme for doctors.

The methods most commonly used for grading texts are readability formulae, cloze procedure, comprehension questions, and subjective judgement. Cloze procedure and comprehension questions were not appropriate for our purposes: quite apart from other considerations, they were excluded on practical grounds, in that we would have required many more student readers than were available to us at the time in order to obtain significant results. Our choice therefore lay between readability formulae and subjective judgement.

Readability formulae have the advantage of objectivity, in that they assign a numerical value to linguistic variables. However, they have limitations. One limitation, of particular relevance for assessing specialist texts, is that the linguistic variables in the most well-known and rigourously tested formulae (Dale-Chall, Flesch, FOG etc.) are defined in terms of "average" general English. Word difficulty, for example, may be assessed by recording the number of words not represented in high frequency word lists, or by using number of syllables as a proxy measure of difficulty. These methods are clearly not suitable for assessing the readability of medical texts, because long words like gastrointestinal, radiotherapy and gynaecologist are typical "core" words in medical English and are not likely to be difficult for the specialist reader.

This example touches on a wider problem in applying readability formulae to specialist texts: they do not take account of the subject knowledge brought to bear on the text by the specialist reader. Yet schema theory indicates that background knowledge has an effect on readability. Finally, most readability formulae were constructed before recent developments in the study of discourse and rhetorical organisation, and therefore take no account of these dimensions.

For all these reasons, we decided to assess the readability of our specialist texts by subjective judgement. The main limitation of this method is the loss of reliability. However, the risk can be reduced by "expert" judgement (usually the judgement of experienced teachers); by further training of judgement, as is typically done with a team of examiners; and by pooling judgements. The principal advantage is that subjective judgement can synthesize a complex range of factors into an overall assessment of difficulty.

In view of our developing interest in grading specialist texts, we decided that the materials writing project would be a useful platform for a research study examining the reliability and validity of alternative ways of assessing readability; in this particular case readability formulae and subjective expert judgement. Since we had opted for the latter in our materials project, we began with an investigation of the nature of subjective judgement. It is this that is the focus of the present paper.

2. Scope of this paper

We have already mentioned that our main reason for preferring subjective judgements over readability formulae was that the latter took no account of a number of significant factors contributing to difficulty. We felt it likely that our subjective judgments would take account of these factors, and therefore be more sensitive estimates of difficulty, particularly with specialist texts.

In this paper we attempt to test these claims by examining in closer detail some properties of subjective judgement, in particular its capacity to encompass and reliably assess various putative components of difficulty.

The first essential task, then, was the identification on theoretical and intuitive grounds of distinct components of difficulty (see list below). The second was the assessment by a team of five raters of the components of difficulty for each of the graded texts.[1] This assessment provided numerical data for the analysis given in section 5 of this paper.

The questions we wished to ask of these data were as follows:

(i) We wished to establish the degree of agreement between the five judges for each of the seven subjectively assessed components of difficulty (i.e the degree of inter-rater reliability). We reasoned that whilst we might naturally expect some divergence between raters, a low correlation would be evidence that the raters were either conceptualizing the component differently, and consequently applying themselves to its measurement in quite different ways, or that the individual rater was unable to operate consistently with respect to that component. This would suggest, in turn, that the component lacked stability, or possibly coherence.

(ii) We wished to find out which of the seven components of difficulty made the greatest independent contribution to overall text difficulty. Which, in other words, were the best predictors of grade level, and which were relatively redundant?

(iii) We also wanted to know more about the inherent dimensionality of our data. Were the raters in fact operating with seven discrete dimensions of difficulty, or were these to any degree psychologically confounded in the process of judgement? We were asking, in other words, whether a smaller set of underlying dimensions could be so constructed as to more adequately represent what the raters were actually assessing.

3. Identification of distinct components of difficulty

The first stage of our research involved the identification and operational definition of components of text difficulty. Operational definitions were required to distinguish the components and make them accessible to measurement.

Below are the definitions that emerged from our deliberations. We do not claim that they are comprehensive, or even that they render the components, in the strictest sense of the word, observable. The definitions went only as far as we felt necessary at the time to allow us to believe that we shared an understanding of what we were measuring.

(i) Length

Definition: (Approximate) number of words in the body of a text.

Comment: Length *per se* has little independent effect on text difficulty but interacts with other factors (e.g. conceptual complexity) on which it has an "add-on" effect - thus increasing overall text difficulty.

(ii) Print size

Definition: Size of the print in the body of the text and the amount of space between the lines.

Comment: This is an affective rather than a cognitive factor because it mainly works on the motivation of the reader. However, print size may have an independent effect on some of the lower-level reading processes (e.g. visual recognition of letter shapes and word boundaries, and may thus affect the reading of those accustomed to other scripts (e.g. Arab and Japanese students).

(iii) Topic accessibility

Definition: Degree of specialization of text.

Comment: Degree of specialization is used as a proxy for familiarity with topic which, though recognised as an important factor, is difficult to estimate directly owing to variation in reader experience and interest. It is primarily a reader-oriented variable. Degree of specialization can be roughly estimated by the type of technical terminology.

(iv) Organization

Definition: Predictability and degree of conventionality of the rhetorical structure of the text.

Comment: The key word here is "genre". If a text belongs to a genre (e.g. case report, research report, or description of disease) for which there are readily recognizable conventions regarding organization of information then the reader (by virtue of prior knowledge of these conventions) is likely to find it easier to find his way around the text.

(v) Contextual support

Definition: Pictures, diagrams, tables, headings, sub-headings and other features of text lay-out (e.g. boxes, italics, bold print).

Comment: Contextual support is hypothesized to make reading a text easier in a number of possible ways: it may increase redundancy of information, represent information contained in the text in a non-verbal manner, summarize and condense important points etc. Headings and titles are particularly important because they "prime" the reader's expectations, allowing him to engage appropriate schemata, thereby facilitating "top-down" processing.

The contextual support factor may perhaps be enhanced in importance once questions have been set on a text (because of the facilitation it lends to tasks).

A word of warning however: not all tables and diagrams provide contextual support. Some add to the information load.

(vi) Information density

Definition: Number of information items contained in a text, in proportion to the length of the text overall.

Comment: Information items are primarily conceived to be factual details though they may also be points in an argument, or opinions, or "discrete ideas". A text with many descriptive details or containing many numerical figures could, for instance, be said to be informationally dense. The "piling-up" of such details increases the information load that the reader is required to process, and may thereby not only slow the reader down but make reading more difficult.

(vii) Conceptual complexity

Definition: Degree to which concepts, notions, ideas, arguments and relationships between entities are difficult to understand.

Comment: This is thought to be an important general factor affecting text difficulty simply by virtue of the conceptual difficulty involved in processing the meanings of the text. Because of the generality of the factor and the difficulty involved in specifying it precisely, it is likely to be particularly heavily intercorrelated with the variables of information density and topic accessibility.

A text may be thought of as conceptually complex to the extent that (a) it involves a high level of abstraction or idealization away from concrete or context-supported entities or relationships, (b) it involves complex chains of reasoning or argument of an abstract nature, and (c) the concepts presented in the text are novel or unusual, requiring considerable modification of readers' existing schemata. Typically, an abstract argumentative text is more conceptually complex than either a primarily descriptive text or a narrative text.

(viii) Syntactic complexity

Definition: Syntactic complexity refers to grammatical features such as embedding, subordination, clause length, ellipsis and referential substitution, double-negation and noun-phrase complexity (especially in subject function). Where these features are widely distributed in a text, then that text may be said to be syntactically complex.

Comment: Syntactic complexity is relatively easy to assess - being a "text-as-object" related variable. There are objective measures of syntactic complexity which are more or less adequate (e.g. T-units). Syntactic complexity is believed to be an important contributor to overall difficulty.

(ix) Lexical difficulty

Definition: The lexical difficulty of a text may be assessed by reference to the proportion of words which are likely to be unfamiliar or unknown to the reader. The main criterion for this is the relative frequency of occurrence of the word in everyday usage. A word may also be unfamiliar, however, if it is drawn from another field of discourse (e.g. the theatre). Also contributing to lexical difficulty are: a high proportion of words which are in wide general usage but with a high indexical value (i.e. items which have many different senses that vary according to context), and for some readers a high proportion of words of Anglo-Saxon origin (i.e. which do not have a Greek or Latin origin).

Comment: Lexis is widely assumed to have a significant effect on text readability (hence it is frequently controlled in simplified readers).

The first two components of difficulty, length and print size, are excluded from the following analysis and discussion because they are objectively measurable. Our present purpose is the investigation of the subjectively assessed components.

4. Measuring the components of difficulty

Bachman (1990) points out that the first step in measurement is to distinguish the construct you are interested in measuring from other similar constructs by defining it as precisely and unambiguously as possible. The next step is to make the definition operational that is, to define the construct in such a way as to make it observable. Assuming that these two steps have been implemented in the previous section, we come to the following stage the systematic quantification of observations using defined units of measurement.

Our procedure here was as follows. For each text, each of the five raters first recorded their assessment for grade level (see Appendix for an account of the procedure followed for allocating texts to grade level). Once it became apparent that grading was consistent, each rater in addition recorded on a visual analogue scale an assessment for each of the seven components of difficulty. This was done for 31 texts out of the total of 60, and in all cases the allocation to grade level was carried out prior to and independently of the assessment of the components of difficulty. The marks on the visual analogue scales were subsequently converted into numerical scores on a scale 0-126 by measuring the distance in millimetres from the rater's mark to the lowest point on the analogue scale.

We opted to start from a visual analogue scale because we believed it would facilitate the judges' assessment by removing from them the responsibility of assigning a numerical score to each of the seven components for each of the 31 texts in the study. To require them to assign a numerical score would be to make distracting demands of precision.

We recognize that the procedures documented above are not unproblematic. First, the conversion of a visual analogue scale to a numerical scale of 0-126 may give a misleading impression of the precision of the initial assessment of the components. Second, although some would prefer to regard scales deriving from subjective judgement as ordinal scales, we have chosen to treat the scale derived from the visual analogue as an equal interval scale. We feel these procedures are defensible, especially as our investigation is preliminary and exploratory, not confirmatory.

5. Results and discussion

5.1 Inter-rater reliability for the assessment of the seven subjectively assessed component

Inter-rater reliability was investigated by inter-correlating each rater's score on each of the seven components in turn. A summary of the correlations obtained for each of the components (Table 1) shows that a respectable level of agreement between raters was attained with regard to four of the seven components. We reason that this indicates that these constructs have at least some stability, and that their operational definitions possess at least some adequacy. By contrast, there is a low degree of agreement with two of the components, suggesting that they lack stability as constructs. In fact, since raters diverge so markedly in their scores for these components, it is possible that they are operating with quite different notions of what the terms 'information density' and 'topic accessibility' denote.

Table 1: Inter-rater reliability on assessment of the seven components of difficulty

Range of correlations between raters scores	Component
(a) High agreement 0.60 - 0.90	Conceptual Complexity Lexical difficulty Syntactic Complexity Contextual Support
(b) Moderate agreement 0.40 - 0.60	Rhetorical organisation
(c) Low agreement 0.00 - 0.40	Information Density Topic Accessibility

These results were not surprising, as the components with a high degree of inter-rater correlation were the easiest to define operationally. They are relatively well-established constructs featuring in numerous discussions of readability, and, with the possible exception of conceptual complexity, they are relatively directly observable. By contrast, the unstable components, showing a low degree of inter-rater agreement, always

presented problems of operational definition. In the case of 'information density', for example, there is some indeterminacy about what constitutes a unit of information. It is hardly very surprising, then, that raters differ over when a text is informationally dense.

The construct of 'topic accessibility' is problematic for two reasons. First, it is highly reader-dependent since what is accessible for one individual may not be for another. Divergence, then, between individual raters is only natural. Second, its operational definition is very indirect. The type of technical terminology indicates degree of specialization which in turn is related to familiarity, and finally to topic accessibility.

We suggest, therefore, that the components of 'topic accessibility' and 'information density' be excluded from consideration in any future judgements of components of text difficulty unless their operational definition can be improved and more direct indices devised. Note, however, that we are not denying their potential significance. However, it seems worth persevering with the four components whose assessment exhibits higher inter-rater reliability.

5.2. The contribution of the seven components as predictors of text difficulty

Our next research objective was to find out which of the seven 'subjective' components were the best predictors of overall text difficulty as indicated by grade level. Remember that all the texts in our study had already been reliably allocated between seven grade levels as part of the construction of the medical English reading scheme. It was possible, then, for us to consider grade level as our dependent variable and the seven subjectively assessed components as predictor or independent variables. Our question was now as follows: what was the unique contribution of each independent variable in the determination of grade level, or, more technically, how might the variance in the dependent variable be apportioned between the various predictor variables?

The reason for desiring an answer to this question will be familiar to those with experience of readability formulae research and development. It is that if we can isolate variables with minimal predictive power, we can dispense with them so making the whole business of determination of level of difficulty more economical.

The preferred technique for answering this question is normally multiple regression analysis because it has the power of isolating the unique correlation of each independent variable with the dependent variable by taking account of and excluding the intercorrelation of that independent variable with others. Prior to undertaking the analysis, there is, however, one important decision to be made which has to with the very nature of the analysis. For technical reasons, (see Woods, Fletcher and Hughes 1986; Hatch and Farhaday 1982), it is important which variable is entered first into the regression equation. There are two methods of proceeding. The first is to enter the variables according to some order of importance that is believed to be theoretically and logically defensible. The second is to operate entirely empirically and allow the computer to decide the order in which the variables enter the equation. Normally, of course, the variable with the highest correlation with dependent variable is entered first. This second method, stepwise regression, is the one we have adopted for this study.

Before presenting the results of the stepwise regression, it may be useful to include two further pieces of information to facilitate interpretation of the multiple regression. These are (a) the inter-correlations between the seven subjectively assessed variables, and (b) the correlation of each predictor variable with the dependent variable before the computation of their unique correlation with the dependent variable.

Table 2: Inter-correlation matrix for seven subjectively assessed predictor variables

	infod	lex	concept	sytax	organ	topic	context
lex	0.399						
concept	0.542	0.819					
sytax	0.555	0.844	0.942				
organ	-0.108	0.458	0.526	0.437			
topic	0.537	0.629	0.748	0.727	0.361		
context	-0.208	0.276	0.297	0.212	0.766	0.160	
gradelev	0.607	0.896	0.924	0.925	0.430	0.668	0.213

Table 2 shows a high correlation between a number of predictor variables, in particular the following pairs: conceptual complexity/syntactic complexity 0.94, syntactic complexity/lexical difficulty 0.84, conceptual complexity/lexical difficulty 0.82. This indicates that to a large extent they overlap and substitute for each other. In other words it is relatively easy to infer the value of one given the value of the other. More problematic is determining the source of the high correlations. It may be that conceptual complexity and syntactic complexity are indeed perceived by raters as distinct dimensions which happen in fact to be closely associated. On the other hand, it may be that raters are simply unable to distinguish clearly between the two, and covertly or subconsciously use the more observable, syntax, as an index of the other.

Table 3: Correlations of predictor variables with the dependent variable (grade level)

Syntactic complexity	0.925
Conceptual complexity	0.924
Lexical difficulty	0.896
Topic accessibility	0.668
Information density	0.607
Rhetorical organisation	0.430
Contextual support	0.213

Table 3 shows the correlations of the predictor variables with the dependent variable. It is interesting to note above that the three predictor variables with the highest inter-correlations are also those with the highest correlations with the dependent variable, grade level. This might be in part an artefact of the high intercorrelations between them, which is precisely what the multiple regression factors out in determining the unique correlations with the dependent variable. Nonetheless, it is worth noting that syntax and lexis, which feature prominently in readability formulae, also have high correlations with grade level in this study. They are also the variables for which there was high inter-rater reliability.

Finally, the results of the stepwise regression (as delivered by MINITAB) are given in Table 4. As might be expected, the first entered variable, syntax, appears to account for

the highest proportion of variance in the dependent variable (85.64). Lexis and information density follow with the remaining variables, conceptual complexity and topic accessibility, accounting for progressively less of the variance.

Table 4. Stepwise regression of gradelevel on seven predictor variables

Step	1	2	3	4	5	6
constant		-0.5462	-0.4304	-1.3355	-0.8507	-0.5490
syntax	0.0777	0.0493	0.0378	0.0133		
T-Ratio	13.15	5.34	4.10	0.99		
lex		0.0271	0.0300	0.0279	0.0304	0.0301
T-Ratio		3.65	4.46	4.42	5.25	5.75
infod			0.0199	0.0185	0.0198	0.0232
T-Ratio			2.80	2.79	3.07	3.87
concept				0.0235	0.0312	0.0405
T-Ratio				2.35	4.92	6.00
topic						-0.0187
T-Ratio						-2.63
S	0.783	0.656	0.588	0.544	0.544	0.493
R-SQ	85.64	90.27	92.46	93.78	93.55	94.91

It would seem, then, that syntax, lexis and information density are the components making the largest contribution in the determination of grade level. One has to remember, however, that information density has already been shown to be an unstable construct and thus its contribution should remain an object of the utmost scepticism.

Possibly more important than the details of the regression analysis is the broad tendency it indicates. Taken together with the evidence mentioned earlier, one is left with the impression that in this study syntax and lexis are the components of difficulty that really count. They are the best predictors of grade level and they are also components with high inter-rater reliability. Other components do make a contribution but adding them in does not much increase the predictive power.

Interestingly, this is broadly consistent with the findings of readability formulae constructors, that syntax and lexis were the best indicators of text difficulty. One has to pose the question, then, whether it is worth persevering with dimensions of difficulty which offer little predictive yield, and which may be difficult to measure reliably.

5.3 The underlying dimensionality of the data

Our third area of investigation concerns the inherent dimensionality of our data. The judges assessed seven putatively distinct and discrete dimensions of text difficulty. However, given that some of the components are highly inter-correlated, one can ask if there is a smaller set of dimensions underlying the data. Or, to put the question another

way: can the seven dimensions that the raters thought they were assessing be collapsed to a smaller number of dimensions which are so constructed that they are more representative of what the raters were actually assessing? If so, then identifying the number and nature of these underlying dimensions may be expected then to provide clues as to the nature of the judgement process.

The statistical procedure we have chosen for this part of our investigation is principal components analysis. This differs from the better known factor analysis in that it is (a) more objective, though less flexible, (b) more robust in its assumptions about the distribution of scores in a population, and (c) more suited to initial exploratory analysis of the kind we are pursuing here.

Table 5: Principal components analysis of seven subjectively assessed dimensions of text difficulty (by covariance method)

Eigenanalysis of the Covariance Matrix

Eigenvalue	2889.78	1229.9	260.0	182.8	128.5	96.3
Proportion	0.600	0.255	0.054	0.038	0.027	0.020
Cumulative	0.600	0.855	0.909	0.947	0.973	0.993

Eigenvalue	32.8
Proportion	0.007
Cumulative	1.000

Variable	PC1	PC2	PC3	PC4	PC5	PC6
indfod	0.148	0.330	-0.468	-0.479	0.280	-0.586
lex	0.496	0.169	0.708	-0.231	-0.280	-0.293
concept	0.527	0.186	-0.178	0.130	0.280	0.389
sytax	0.411	0.209	0.030	0.010	0.312	0 360
organ	0.327	0.443	-0.027	0.602	0.260	-0.511
thlopic	0.299	0.175	-0.465	0.246	-0.766	-0.018
context	0.299	0.747	-0.175	-0.527	-0.116	0.171

Variable	PC7
indfod	0.017
lex	-0.007
concept	-0.640
sytax	0.748
organ	0.074
thlopic	0.130
context	0.025

Table 5 shows the results of the principal components analysis (PCA), which are the basis for the subsequent discussion. There are two key issues in the interpretation of the results of principal components analysis: (a) how many principal components (or underlying dimensions) should be dignified with recognition and (b) what meaning should be attached to the resulting components? Let us take the number problem first.

The first component (eigenvalue 2889.8) accounts for 0.6 (60%) of the total variance, while the second component (eigenvalue 1229.9) accounts for a further 0.25 (25%) of

the total variance. These two components together account for just over 85% of the total variance. The remaining components account for progressively less variance.

We also need to know how many components to accept. There is an objective method of doing so, but it requires such complex computation that a rule of thumb is commonly used: "if the original data has p dimensions, assume that components which account for less than a fraction 1/p of the total variance should be discarded" (Woods, Fletcher and Hughes 1986: 283)

We found that we should discard components accounting for less than 14% of the total variance, which left us with only two of the components. In other words only two distinct dimensions underlie the original seven. Perhaps, then, the judges were in fact operating with two dimensions though they may have believed they were independently assessing seven.

The next problem is to interpret what these two components mean. Technically speaking, they stand in need of reification. If we examine how the principal components correlate with the original seven variables, we see that the first principal component correlates positively with all the original variables. This is unsurprising since the first principal component is typically a 'general factor'. It is, if you like, a kind of composite of the original variables. There is, however, a bias towards higher correlation with the variables of conceptual complexity (0.527), lexical difficulty (0.496), and syntactic complexity (0.411). This suggests that the first principal component may represent a factor of 'general language difficulty'.

The second principal component presents even greater problems of interpretation. It has a relatively high positive correlation with information density (0.330), and relatively high negative correlations with contextual support (-0.747) and rhetorical organization (-0.443). Thus, a text with a high score on this component will tend to have an above average score for information density, a below average score for rhetorical organization and a markedly below average score for contextual support. All this may perhaps point to the second component having something to do with contextual support and rhetorical organization. Even this, however, leaves problems of interpretation. Perhaps, one might give the label 'textual lay-out/organization' to the second component.

Our analysis, then, has led us to conclude that there are two significant dimensions in our data. We have chosen to interpret these as (a) 'general language difficulty', and (b) 'textual lay-out/organization'. The purpose of undertaking the analysis, however, goes was to uncover clues as to the nature of the process of judging text difficulty. Here we can speculate that what the judges were actually doing was (a) estimating the general language difficulty of the text and (b) deriving an impression of the relative ease of the text by sampling fairly superficial characteristics of appearance (diagrams, tables, subheadings etc) as well as rhetorical organization.

We must be cautious, however. The difficulty in interpreting the second component draws attention to Woods et al's (1986) point that PCA is an exploratory instrument. Clearly, further detailed investigation of the components is required.

6. Conclusion

It will be clear that this is indeed very much a working paper. Our analysis of the data is preliminary and exploratory, and much work remains to be done. Nevertheless, we have, we believe, turned up some results of interest.

We have shown that some of our components of text difficulty possess greater stability and reliability than others. In particular, the components of topic accessibility and information density are so unstable as to be hardly worth persevering with in their present form. A general lesson is that it is one thing to identify possible components of difficulty on theoretical and intuitive grounds but quite another to define them operationally and render them accessible to reliable measurement. Perhaps, then, there is something to be said in favour of the concentration of readability formulae on lexical and syntactic difficulty.

This latter point is also borne out by the evidence from our multiple regression analysis. This, again, suggests that the linguistic components of syntactic complexity and lexical difficulty are the best single predictors of overall text difficulty (grade level). Information density comes third, but we have noted how unstable it is. The remaining components appear to be something of a luxury. Individually, and collectively, they account for little of the remaining variance in grade level, and some are assessed with barely satisfactory inter-rater reliability.

In a sense, then, many of our findings are negative but interestingly so. We started from the belief that subjective judgements could encompass a wider range of factors than readability formulae. More detailed comparability studies are needed.

PCA constitutes another area of investigation. Using this technique, we sought to get below the surface of things - never an easy thing. In particular, we wanted to find out more about the structure of judgements of text difficulty. Our tentative findings were that raters appeared to operate with two dimensions: (a) a dimension of 'general language difficulty' and (b) one of 'textual lay-out/organization'.

Having gone this far, we can identify at least three possible directions for further research. First, we need to refine and elaborate some of our existing statistical analyses, particularly in relation to principal components analysis. Second, we wish to compare the estimates of text difficulty of different categories of rater, e.g. medical English teachers, non-native doctors, and native speaker doctors. Third, we would like to compare raters with readability formulae in terms of how they rank texts according to difficulty. Taken together these investigations may offer some further clues as to the nature of readability in relation to medical journal articles.

Note

1. The five raters were: Gibson Ferguson, Ron Howard, Joan Maclean, Anne Murray and Alison Oates.

Acknowledgements

The writers would like to thank Ron Howard, Anne Murray and Alison Oates, both for their part in the judging process and also for their contribution to the discussions that led to the operational definitions of the components of difficulty.

The writers would like also to thank Dr Clive Criper, who supported the first stages of this project with funding, advice and encouragement.

References

Bachman L. 1990. Fundamental Considerations in Language Testing. Oxford: OUP.

Hatch E. and Farhady H. 1982. Research Design and Statistics for Applied Linguistics. Rowley, Massachusetts: Newbury House.

Woods A., Fletcher P. and Hughes A. 1986. Statistics in Language Studies. Cambridge: CUP.

Appendix

The allocation of texts to seven levels of difficulty: Grading procedure.

The judges were five experienced medical English teachers (one of whom was also a qualified doctor).

We selected 60 articles from medical journals, representing what appeared to us to be the range of difficulty in these journals. Articles which were not "mainstream medicine" (such as discussions of medico-legal matters, ethical aspects of medical practice, or salary and working conditions) were excluded from the selection. The source journals were in the main generalist rather than specialist medical journals, i.e. the target readership was anyone with medical qualifications rather than doctors from a particular specialty. We hoped thereby to go some way towards controlling for difficulty due to inappropriacy of the text for the reader with regard to subject matter.

(i) We viewed the collection of 60 texts and agreed on which text was likely to be the easiest, and which the most difficult, for an "ideal non-native speaker doctor". These two texts therefore represented the end-points of our range of difficulty.

(ii) We then individually made judgments about the difficulty of three other texts in relation to each other and the endpoint texts, i.e. we separately ranked the five texts. One of the texts was agreed to be roughly in the middle of the range of difficulty, and so we now had three "criterial" texts (bottom, middle, and top) to be used as reference points for ranking other texts.

(iii) We tried to view the range as a continuum but found that we were so much influenced by working with the seven levels of an earlier version of the graded reading programme that we tended to think in terms of seven levels. We therefore proceeded to allocate texts to a grade level (from 1 to 7). It should be noted that we made no claim that the grades divide the continuum into equal intervals of difficulty. The grades represent ranks of difficulty.

(iv) This process was repeated with small numbers of texts until we were confident that we were judging consistently. After 11 texts were ranked, we were able to proceed more quickly with the remaining 49. Further "criterial" texts (to an eventual total of seven, one for each grade level) were identified as reference points as we proceeded.

(v) In order to establish the grading for the ESP reading programme, we then pooled the overall judgments of the five judges. In fact, although there were occasional discrepancies between judges, the rate of agreement was extremely high.

(vi) As a point of interest, about two months after the graders were finished, each judge independently ranked all 60 texts (without looking at the grading papers). The correlations across judges were high (about 0.95).