Title	What makes reading comprehension test items difficult? An exploratory analysis
Author(s)	Ho Wah Kam, Lim Tock Keng and Patricia J. Y. Wong
Source	*ERA - AARE Joint Conference, Singapore, 25-29 November 1996*

A Draft Only (new corrs)

Filename: era96.doc

What Makes Reading Comprehension Test Items Difficult? An Exploratory Analysis

by

Ho Wah Kam, Lim Tock Keng and Patricia J. Y. Wong

Abstract

This paper reports on a sub-study based on a bank of test items developed and calibrated for a computerized adaptive reading test for primary and secondary students in Singapore schools. In the main study (as reported in various papers by Lim Tock Keng et al, 1994, 1995), a bank of multiple-choice test items constructed for primary 3 and 5 and secondary 1 and 3 students was found to contain a relatively large number of test items that were quite difficult for these students. While generally it was feasible to guess the (rather obvious) reasons for the very easy items, it was thought necessary to check empirically on the factors that contribute to item difficulty. This sub-study attempts to answer directly the research question (What makes test items in reading comprehension difficult?) in relation to the items developed for the primary 5 students (beween 10 and 11 years old). This was done empirically from the perspectives of the linguistic features of the item texts, propositional analysis, decision processing, and the cognitive demand of the test items. The paper will be in five parts, namely, (a) a brief discussion of the cognitive model of text processing that underpins this exploratory analysis and the idea of text difficulty, (b) the operationalisation of the variables, (c) the statistical findings, (d) the sources of difficulty in English reading comprehension tests of the type studied, for students whose first language is not necessarily English, and (e) implications for test development.

1.Introduction

The strategy of this study represents an approach to what may be called construct definition, which is "the process whereby the meaning of a construct...is specified" Stenner et al (1983:307). It is through constructs that science orders and gives meaning to observations, or to test performance and ratings in the case of psychological and educational testing. According to Stenner et al (1983:306), traditionally psychometricians have been examining person variation rather than variation in item scores as the basis to understand a measurement's meaning or its construct validity. They added that "until the developers of educational and psychological instruments can adequately explain variation in item scale values (ie. item

difficulty), the understanding of what is being measured will remain unsatisfyingly primitive" (Stenner et al, 1983:305). This view provided one of the two reasons for asking the question posed in this paper: What makes test items in reading comprehension difficult? The other reason, a more practical one, is explained a little later. The choice of domain (ie. reading comprehension) was determined by the availability of data.

The typical comprehension test consists of a reading passage followed by a number of questions (open-ended or multiple-choice) based on what is in the preceding passage. Performance on the question depends on sources of difficulty in the passage and in the question itself. Published research (eg. that of Perkins and Brutten, 1993 and

Scheuneman et al., 1991) has shown that the sources of difficulty stem from the language of the text (passage) used, the cognitive demand of the question itself and text-question interactions. There are also the reader factors (background knowledge, language ability), text factors such as content and structure and task factors (such as types of questions). Different studies have found different text factors as having predictive value for reading difficulty - with (as in the case of Perkins and Brutten, 1993:217) the percentage of content words per passage, the number of predicates per passage, passage-predicate density and cognitive demand of the test questions accounting for a significant amount of the test variance. On the other hand, Scheuneman et al. (1991) found that factors such as readability, structure, number of propositions, propositional density and level of questioning accounted for a significant percentage of the test variance. In brief, different combinations of predictive variables accounted for from 65 to 72 per cent of the test variance. However, it must be noted that in these studies reviewed different tests were used and student samples and language environments (EL1 or EL2) differed. Such basic differences make comparisons less helpful.

Nonetheless, it seems obvious that test writers for national examinations and teachers who construct tests for school examinations do not usually go beyond a subjective judgement of test difficulty towards "quantifying" or controlling the level of difficulty of test passages. While careful control of test content, which is what teachers have done very well, the approach is still very subjective. However, for teachers to "quantify" or control passage difficulty less subjectively, they should know (empirically) what factors/variables account for the observed difficulty of comprehension passages. The authors of this paper were motivated to address such a question (What makes English comprehension test items difficult?) as a follow-up to the calibration of sets of test items for developing an item bank in an on-going project on computerized adaptive testing in the areas of reading comprehension (see Lim et al., 1994 and 1995) and mathematics. A primary purpose of this sub-study is to find out how the sources of

difficulty in reading comprehension test items may be quantified by the ordinary teacher/test setter using the computer. Our belief is that reading test writers would benefit much from studies aimed at answering the question asked in the title of this paper because while it is important to know how pupils performed on items, it is equally useful, if not more so from the psychometric point of view, to know how the items behaved.

2.A Reading Comprehension Model and Difficulty Factors

Cognitive psychologists and reading specialists generally agree that text comprehension involves the two main factors of how the surface features of a text are "translated" by the reader into underlying conceptual propositions, and also how the reader in using his/her background knowledge to identify referents of the concepts in the text infers causal connections in the sequence of actions in a narrative, according to Bower and Morrow (1990: 44). They referred to their referential representation as a "mental model or situation model".

As most teachers know, there are many potential sources of difficulty in a reading comprehension task. For the purpose of this study, the complexity factors taken into account were related to the characteristics of the reading comprehension passages. For example, the surface features of a text (eg. the number of content words, the number of sentences in a passage) have been shown to account for much of the difficulty in understanding a text or passage set for comprehension. Nouns, verbs, adjectives and adverbs were counted as

content words. Some of these features have to do with the reader's capacity to process textual information. In brief, text properties related to sentence structure, semantic content and readability ease were considered.

Under semantic content, it has been suggested (see Kintsch and van Dijk, 1978) that the number of propositions in a text is related to item difficulty; so is propositional density. The work in propositional analysis assumes that memory processes and text characteristics affect the time taken to process a text. A proposition, a basic unit of meaning, is considered to be a more psychologically significant feature of a text than a surface feature. According to Scheuneman et al (1991, 14--15), who used propositional analysis in their study, propositions are made up of concepts, of which the first element is the predicate. In turn, predicates relate arguments which are subjects and objects. Another category of propositions is the modifier which can be an adjective or adverb or any phrase that modifies the arguments in the text. The number of predicates, arguments and modifiers was computed separately, from which predicate density, argument density, modifier and combined (propositional) density were also determined. Freedle and Fellbaum (1987) also found that lexical overlap helped to account for

multiple-choice item difficulty in the TOEFL test of single sentence comprehension such that item options that contain greater lexical overlap with the stimulus sentence tended to be associated with the easier items.

The overall level of difficulty of a text was, in addition, determined through an index described as readability ease. This was the one Flesch had developed, based on a scale from 100 (very easy to read) to 0 (very difficult to read). It incorporated text properties such as average sentence length and word complexity.

The cognitive demand of the questions in comprehension can also cause difficulty. Based on previous work (eg. Scheuneman and Gerritz, 1990 and others), cognitive demand was defined afresh in terms of three levels, simple inference (inferring the meaning of words from the context), complex inference (requiring the interpretation of relationships between elements in the text not explicitly stated), and evaluative (making an evaluative assessment of the elements in the text). These three categories representing the demand level of the questions and the cognitive process involved formed a convenient "taxonomy" of cognitive skills in reading comprehension.

3.The Test

In the computerized adaptive testing project, 60 per cent of the comprehension test format used what may be called the short-context technique (Jafarpur, 1987), a technique in which a brief passage (narrative or expository) is followed by a single multiple-choice question. The writers of the test items chose this mode because this technique allowed for the use of a variety of passages and contexts in order that no particular group of test takers would be favoured because of prior knowledge of the topic or content. Because of the brevity of each passage, it allows the test writer to test total comprehension rather than parts of the passage.

The relative merits of close-ended, multiple-choice items in the traditional comprehension format of the North American variety have been debated for a long time in the literature. In brief, Elley and Mangubhai (1992) tested the argument of no significant difference between multiple-choice and open-ended items in a study carried out in both Australia and New Zealand among nine-year old students and found

that "item format exerts no significant influence on the outcomes in a large-scale survey of readers, when a major purpose is to establish levels of comprehension in comparable groups of students" (p. 198). While there are sound pedagogical arguments for using open-ended questions in reading instruction, Elley and Mangubhai (1992: 198) added that a distinction has to be made between regular practice and formal testing on a large scale. For this reason, this testing project adopted

the multiple-choice format as it was testing more than 1300 students.

4.The Data

The data for this exploratory study came from 42 out of the 70 items in a reading comprehension test administered to 1313 primary five students from a cross-section of schools. The data were used to calibrate newly constructed test items for the development of an item bank for a computerized adaptive testing project for primary and secondary school students. The project was aimed at providing a practical alternative to paper-and-pencil testing, making use of the latest computer technology. The computerized adaptive test is aimed at measuring the reading comprehension "ability" of the students as efficiently as possible, matching to the extent possible, using currently available software, the difficulties of the test items to the ability of the sample (both of which are measured on the same scale using a unit called logit). The difficulty levels of the items were determined from the administration of the test to a pilot sample, using the paper-and-pencil format.

A map of the person-ability and item-difficulty values of a pilot sample of 1313 students is shown on the next page (see Figure 1). From the map, it is clear that the ability distribution of the sample is skewed towards the higher ability end and, because of this, there would be in comparison more relatively easier items than difficult ones. Nonetheless, there were items which proved to be quite difficult for the majority of the sample, which was the reason for this sub-study. On the whole, the sample found the inferential type of items and those testing main ideas difficult, as shown in Table 1 on page 6. Although the testing plan provided for items covering a wide range of difficulty, the item selection criteria adopted would have sieved out a number of very difficult items.

Fig. 1. Map of Person-Ability and Item-Difficulty Values

Table 1. Distribution of Logit Values for the Items/Questions

For the purpose of data analysis, item difficulty was computed from the responses of this sample of primary five pupils. This was the p-value or proportion of correct responses. This p-value was in turn

transformed to the delta scale, which has a mean of 13 and a standard deviation of 4. The delta scale, which (according to Scheuneman and Gerritz, 1990) is an index of item difficulty used by the Educational Testing Service for test construction, yields a statistic that is always positive, ranging in practice from 5 to 24, with the higher values representing easier items.

5.Results

Table 2 presents the means and standard deviations of the variables categorized under three sub-headings. These variables were chosen based on previous research.

Table 2. Means and Standard Deviations ofVariables

Variable	Mean	SD
Test		
Item diff (classic, p-val, non-delta)	0.28	10.21
Text Structure		
No. of words per passage	80.40	31.10
No. of content words per passage	40.36	16.66
No. of sentences per passage	8.19	4.74
Word/sentence ratio of passage	11.14	3.34
Perc of content words per passage	0.50	0.06
Type-token ratio	0.73	0.09
No. of passive sentences	9.85	13.59
Readability	88.58	9.15
No. of words per sentence	11.01	3.10
Propositional Analysis		
No. of arguments per passage	6.40	3.24
No. of predicates per passage	9.31	4.66
No. of modifiers per passage	3.60	3.11
No of propositions per passage	19.31	7.42
Predicate density per passage	1.32	0.75
Modifier density per passage	0.51	0.39
Argument density per passage	0.09	0.44
Combined density per passage	2.74	1.04
The questions		
Cognitive demd (of qns/options)	1.48	0.59
Lexical overlap (betwn pasge & qns)	1.00	1.08

In Table 3, the coefficients of the variables when correlated with item difficulty are displayed. Five of the variables correlated significantly ($p < 0.05$) with item difficulty although not quite substantially. They were the number of sentences in a passage, readability ease, number of predicate propositions, number of content words in a passage word/sentence ratio and the total number of propositions. However, in the nature of things, a variable might correlate (in its 'raw' form) quite highly with item difficulty on its own but, as we shall see, in multiple regression the quantum of its value in a model depends on its ability to add extra predictive power over and above that already provided by variables that preceded it. The effect of multicollinearity became apparent in the results displayed in Table 4.

Table 3. Correlations with Item Difficulty (Facility)

No. of sentences per passage	.38*
Readability	.34*
No. of predicates	.34*
Word/sentence ratio	-.33*
No. of propositions	.33*
No. of words per sentence	-.30
No. of words per passage	.28
Proposition density per passage	-.26
Argument density	-.25
Modifier density	-.19
Type/token ratio	

-.16

No. of arguments
.15

No. of passive sentences per passage
-.12

Predicate density
-.12

No. of modifiers
.11

No. of content words per passage
-.08

Percentage of content words per passage
.07

Extent of lexical overlap
.05

Cognitive density
.01

*Statistically significant at the 0.05 level.

Table 4 presents the results of a stepwise regression analysis. Owing to the multicollinearity problem as explained earlier, a few variables were off-loaded from the analysis as a result. In turn, the variable, number of sentences in a passage, was entered into the analysis first

and it accounted for 14 per cent of the variance in test difficulty. The readability index was entered next, accounting for another four per cent of the variance in item difficulty, after which came type-token ratio, accounting for another three per cent and the number of predicates a further two per cent of the variance. The first three variables in the regression equation came from text structure and the fourth from propositional analysis.

Table 4.Results of Stepwise Regression for Dependent Variable, Item Difficulty

Variable Entered	R-Square	R2 increase	F-value
No.sentences in passage	0.14	0.14	6.56*
Readability index	0.18	0.04	4.18*
Type-token ratio	0.21	0.03	3.24*

No. of predicates	0.23	0.02	2.27*

* p< 0.05

Table 5. Four Variables in the RegressionEquation

The F-test for each unstandardized coefficient in Table 5 gives credit to each variable only for its incremental contribution when all the other variables have been introduced in the equation. Based on this four-variable model, the general regression equation to determine item difficulty is:

$$Yi = -7.08 + .29X1i + .09X2i + 11.09X3i + .17X4i$$

where X1 = number of sentences in the passage, X2 = readability ease, X3= type-token ratio, and X4 = number of predicate propositions.

So, to obtain a predicted item difficulty value for any given value of the four predictor variables, the authors would need to employ the constants in the column under regression coefficients in Table 5. Nonetheless, it has to be noted that these four predictors explained only a relatively small percentage of the difficulty variance; a large part of it remains unexplained.

6. Discussion of the Results

This study was intended to be exploratory since the results would at best be preliminary, and not conclusive at all.. These results therefore cannot be generalized to all reading comprehension tests in English for primary children. The main purpose of this study was to explore the idea that intrinsic features of the items can throw some light on the difficulty of the items. The focus was on the behaviour of the items rather than on that of the test-takers. The strategy used was to choose features of the items (ie. the passages) that could be classified into categories. The method adopted was largely quantitative and it was concerned principally with text characteristics. Stepwise multiple regression was used to evaluate the simultaneous contribution of the variables to item difficulty.

Clearly the number of sentences in a passage explained a fair amount of the difficulty of the test; in other words, when passage length is held

constant, the shorter the sentences the easier the passage. Next was

the readability of a passage, which was measured by the Flesch readability index. This index includes some measure of word difficulty. Type-token ratio refers to the percentage of word types in a passage, ie. the lower the ratio, the fewer the word types, which in turn would make the passage easier to understand or process. The number of predicates in a passage, which explained some two per cent of the variance, has to do with the coherence of a text. The predicate is what Embretson and Wetzel (1987) have called superordinate propositions. According to them, superordinates usually link the text for coherence and to some extent predicates play that role. A predicate, in fact, relates arguments. In this case, unlike in Embretson and Wetzel's (1987) study, the number of predicate propositions in a passage did have a statistically significant correlation with item difficulty and is retained in the regression equation..

In summary, the results show that 23 per cent of the variance of the item difficulty of the 42 items could be accounted for by four features: the number of sentences in a passage, the readability index of each passage, the type-taken ratio of the passage and the number of predicate propositions. In other words, the easier items had passages that used more sentences, high on the readability index, low on the type-taken ratio and had relatively more predicate propositions.

In retrospect, we can see that there were two problems in using features of passages to predict the difficulty of passages. On one hand, some of the features did not appear in all of the passages, and on the other hand, textual features may contribute little to passage or item difficulty by themselves (ie. individually) but may interact with other features to create greater difficulty or that a difficulty occurs when other features of a text were present. The only type of "interaction" we tried to measure was lexical overlap between the text/passage and the questions. However, the correlation extent of lexical overlap and item difficulty was low (re = .06).

The preliminary nature of these findings stresses the need for refinement of the techniques and for inclusion of other text variables that are less gross and also reinforces the point that the task of quantification can be further pursued in the area of reading comprehension.

7. Implication for Test Construction

For the moment, there is a case for suggesting that English comprehension tests that are set for school examinations would benefit from some attempt at quantifying the difficulty characteristics of passages used. This type of test validation can be built into the school's test construction procedures. Two of the characteristics identified in this study, namely, number of sentences in a passage and readability ease, are usually computed in any good word-processing program without the user asking for them. Type-token ratio and the

number of predicate propositions should not be too difficult to compute even by hand. English test setters can therefore move away from total dependence on a subjective judgement of passage difficulty.

References

Bower, G. H. and Morrow, D. G. (1990). Mental models in narrative comprehension. Science, 247, pp 44-48.

Elley, W. & Mangubhai, F. (1992). Multiple-choice and open-ended items in reading tests: Same or different? Studies in Educational Evaluation, 18, pp 191--199.

Embretson, S E & Wetzel, C D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11 (2), pp175--193.

Freedle, R and Fellbaum, C. 1987. An exploratory study of the relative difficulty of TOEFL's listening comprehension items. In Freedle, R and Duran, R (eds). Cognitive and Linguistic Analysis of Test Performance. Norwood, N J : Ablex.

Jafarpur, Abdoljavad. (1987). The short-context technique: an alternative for testing reading comprehension. Language Testing, 4(2), pp 195-220.

Kintsch, W. and van Dijk, Teun A. (1978). Towards a model of text comprehension and production. Psychological Review, 85(5), pp 363-394.

Lim Tock Keng, Ho Wah Kam and Wong, Patricia J. Y. (1994). The computerized adaptive reading test project. Paper presented at the 8th Annual Conference of the Educational Research Association, 24-26 November, Singapore.

Lim Tock Keng, Ho Wah Kam and Wong, Patricia J. Y. (1995). Computerized Test in Reading Comprehension. In Proceedings of the ICCE 95 - Applications Track, organized by the Asia-Pacific Chapter of AACE, Singapore 5--8 December.

Perkins, K. and Brutten, S. R. (1993). A model of ESL reading comprehension difficulty. In Ari Huhta, Kari Sajavaara and Sauli Takala (eds), Language Testing: New Openings. Finland: Institute for Educational Research.

Scheuneman, J and Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. Journal of Educational Measurement, 27 (2), pp, 109-131.

Scheuneman, J, Gerritz, K and Embretson, S. (1991). Effects of

paragraph complexity on achievement test item difficulty. Princeton, N J: Educational Testing Service.
Stenner, A. J., Smith III, M and Burdick, D.S. (1983). Toward a theory of construct definition. Journal of Educational Measurement, 20, pp. 305--317.