דוח המרכז

229

Prediction of Item Difficulty in the English Subtest of Israel's Inter-University Psychometric Entrance Test

Ruth Fortus
Rikki Coriat
Susan Fund

January 1997

מרכז ארצי לבחינות ולהערכה (ע"ר)
NATIONAL INSTITUTE FOR TESTING & EVALUATION
מיסודן של האוניברסיטאות בישראל
ת"ד 26015 ירושלים 91260, טל' 02-6759555

Prediction of Item Difficulty in the English Subtest of Israel's Inter-University Psychometric Entrance Test

Ruth Fortus, Rikki Coriat and Susan Fund

January 1997

We thank Basha Leah Blumberg, Naomi Gafni, Chanan Goldschmidt, Debbie Lerman, Debbie Mishali and Avital Moshinsky for their time and helpful comments.

TABLE OF CONTENTS

ABSTRACT

The aim of this study was to gain a greater understanding of the factors affecting the level of difficulty of the multiple-choice items, particularly the reading comprehension items, appearing in the English subtest of Israel's Inter-University Psychometric Entrance Test. Isolating these factors would result in a number of practical benefits (designing an item pool in accordance with specific needs, writing items for a text that would be tailored to the difficulty level of that text, etc.) as well as provide support for the construct validity of the test items.

Reading comprehension text variables (level of vocabulary and level of grammatical complexity) were found to have the greatest effect on item difficulty, thus providing evidence for the construct validity of multiple-choice reading comprehension tests. Other variables that were found to be significantly correlated with reading comprehension item difficulty were amount of processing (a text-by-item variable), type of item, length of distractors, and level of vocabulary in stem and distractors (item variables). It was found that the correlation between raters' predictions of item difficulty and actual item difficulty (for reading comprehension items) increased from 0.24 to 0.82 following a detailed process of analyzing texts and items. It seems that this type of process contributes to a greater understanding of the factors affecting item difficulty, and that this knowledge can be implemented effectively by raters in the predictions they make. It is highly probable that item writers would also be able to make use of this knowledge.

INTRODUCTION

Large-scale assessment for critical decision-making purposes is often conducted using a test format based on the multiple-choice item type. Therefore, validation studies that explore what determines the difficulty of these types of items can be of great value to the testing community. In addition, given the growing role that English is playing in the international community, investigating what accounts for item difficulty among non-native English speakers taking a multiple-choice test in English is likely to be of particular interest. Isolating the sources of item difficulty for a specific group of non-native English speakers might provide a basis for generalizing across other language groups.

While research exists which has explored factors related to item difficulty, only a few studies have focused on the prediction of difficulty of multiple-choice reading comprehension items in standardized tests, despite the widespread use of this type of item. Drum, Calfee and Cook (1981) predicted difficulty for multiple-choice items on reading comprehension tests by rating texts, item stems and distractors according to mostly structure-related variables (word frequency, number of words with more than one syllable, etc.). They found that text vocabulary factors had a stronger effect on difficulty than did the syntactic properties of the texts; that the appearance of infrequent words in the stem affected difficulty; and that the major factor contributing to item difficulty was the plausibility of the distractors. Embretson and Wetzel (1987) predicted item difficulty for multiple-choice paragraph comprehension items using models which dealt with the processing stages of both text representation and response decision. They found that response decision variables had a greater effect on item difficulty than did text-related variables. Freedle and Kostin (1991, 1992) attempted to predict the difficulty of Scholastic Assessment Test (SAT) and Graduate Record Examination (GRE) reading comprehension items. They found that – among other variables – the abstractness of the text, the number of negations, paragraph length and the rhetorical organization of the text increased the items' difficulty. In their study of Test of English as a Foreign Language (TOEFL) items, Freedle and Kostin (1993) found results similar to those they had found for SAT and GRE items. Perkins, Gupta and Tammana (1995) used an artificial neural network (ANN) to predict item difficulty for 29 multiple-choice reading comprehension items taken from TOEFL tests. The use of an ANN, unlike the traditional statistical procedure of multiple regression, does not require the assumption of a linear relationship

between the predictor variables and the dependent variable, and for this reason was preferred by the authors. Three categories of predictor variables were used in this study: text surface structure variables taken from Drum et al. (1981), propositional analysis variables taken from Scheuneman, Gerritz and Embretson (1989) and from Scheuneman and Gerritz (1990), and a cognitive demand variable from Scheuneman et al. (1989), for a total of 24 variables. Their results showed a reasonable prediction of the item difficulty by the ANN, despite the relatively small size of the set used to train the ANN. However, their study did not specify for which types of variables the prediction was most sensitive.

Recently, there has been much criticism of multiple-choice tests of reading comprehension. Katz, Lautenschlager, Blackburn and Harris (1990) claimed that examinees do not have to read the text in order to answer many of the questions correctly. This calls into question the construct validity of such tests. However, Freedle et al. (1991, 1992, 1993) found that text or text-by-item variables were better predictors of item difficulty than item-related variables (an example of a text-by-item variable is "lifting," where both the text and the item must be scanned before the item can be rated). The evident contribution of these types of variables to the amount of explained variance provides support for the construct validity of multiple-choice reading comprehension texts. It should be noted that, with the exception of Freedle et al.'s (1993) study and Perkins et al.'s (1995) study, none of the above-mentioned studies focused on predicting difficulty for multiple-choice items in the context of a test taken in a foreign language. It is reasonable to assume that the factors which influence level of difficulty for a non-native speaker might differ from those affecting native speakers.

In the context of language research, relying on the judgement of experts poses some difficulties: the experts may not have sufficient expertise to perform the required task; what they are being asked to evaluate may not be sufficiently clear; and the instrument used to elicit their evaluations may be flawed, to name but a few. In the study conducted by Alderson and Lukmani (1989) on the ranking of reading skills demanded in test items by teachers in the Institute for English Language Education in the University of Lancaster, it appeared that, for more than half the items, there was little agreement among judges on what was being tested by each item. Other studies have produced contradictory findings: Bejar (1983) showed that expert judges cannot reliably estimate the difficulty of test items. Chalifour and Powers (1989) studied GRE analytical reasoning items and concluded that experienced item writers are capable of estimating the

difficulty of this type of item. Alderson et al. suggested that, since an answer to a test question may be arrived at by using different processes, strategies or skills, and that since this is as true for a judge as it is for an examinee, making inferences about underlying processess from the answers to test items is risky. The same may be said for estimations of item difficulty: Factors that affect difficulty in the eyes of a judge may not be those factors that affect difficulty for the examinee, particularly if the judges are native speakers of the language being tested and the examinees are not.

Stansfield and Kenyon (1995) compared the scaling of speaking tasks by three groups of teachers (French and Spanish language teachers and bilingual education teachers) with the scaling posited by the American Council on the Teaching of Foreign Languages (ACTFL) Guidelines. Their results showed that teachers were able to recognize those tasks ranked as Superior level according to the Guidelines, but were unable to rank the Advanced and Intermediate level tasks in a manner consistent with the guidelines. Differences were also found among the groups of teachers in the scaling of the speaking tasks. One explanation the authors presented for these differences is that one group of teachers – the language teachers, whose scaling more closely approached the scaling according to the ACTFL Guidelines than did that of the bilingual education teachers – may have had greater familiarity with the required task because it is one with which they have had experience in their classroom teaching. Stansfield et al. also suggested that native speakers may be less metalinguistically aware of the levels of ability required to perform a speaking task in their native language than are non-native speakers.

The results of three of the studies mentioned above – Bejar (1983), Alderson et al. (1989), and Stansfield et al. (1995) – reflect the inability of judges to reliably perform various tasks: to predict item difficulty, to evaluate what an item is testing, and to scale speaking tasks. However, if a group of judges who are deemed experts in the domain being studied can agree on how particular variables are reflected in a certain type of item, and if their judgement is shown to be statistically relevant, then an additional source of evidence pertaining to the construct validity of that type of item will have been found.

OBJECTIVES

The aim of this study was to gain a greater understanding of the factors that influence the level of difficulty of items appearing in the English subtest of Israel's Inter-University Psychometric Entrance Test (PET), which is developed and administered by the National Institute for Testing and Evaluation (NITE). The PET also includes Verbal Reasoning and Quantitative Reasoning subtests. The English subtest, which is composed of multiple-choice items, is designed to test command of the English language in terms of the ability to read and understand texts at an academic level. The score on the English subtest serves two purposes: It is a component of the total PET score, which is used as an estimate of future success in academic studies in order to select university candidates, and it is also used for placement of students in remedial English classes. To attain the goal of placing students at the level most suited to their knowledge of English, the English subtest must comprise a wide range of items, from very easy to extremely difficult ones. Isolating the factors affecting item difficulty would result in a number of practical benefits: Test developers would be better able to design an item pool in accordance with specific needs (such as the need to differentiate both at very low and very high levels of ability), content specifications for tests would be more precisely defined, and the items written for a particular text would be better adapted to the difficulty level of that text. In addition, the ability of expert judges to identify factors which are relevant to the construct being tested – knowledge of English – would provide support for the construct validity of the test items.

METHOD AND RESULTS

The study was conducted in four stages: In Stage 1, raters were asked to estimate the difficulty of all items in the English subtest; in Stage 2, variables affecting the difficulty of reading comprehension texts were investigated; in Stage 3, variables affecting the difficulty of reading comprehension items were investigated; and Stage 4 was a replication of Stage 1.

The English subtest of the PET comprises 54 items subdivided into two sections of 27 items each. Each subsection includes three item types: sentence completions, restatements and reading comprehension texts with accompanying items (see Appendix A for sample items).

The nine raters who participated in the various stages of the study were all involved, to varying degrees, in the preparation of the PET: Seven were native English speakers directly involved in writing and editing items for the English subtest, and two were native Hebrew speakers, highly proficient in English, who work in NITE's test development department.

Stage 1

Six raters (four native English speakers and two native Hebrew speakers) were asked to estimate the level of difficulty of 162 items taken from six English subsections of the PET. Item difficulty, the dependent variable, was measured by calculating the percentage of examinees who correctly answered each item. These percentages were normalized and then converted, by means of a linear transformation, into delta values with a mean of 13 and a standard deviation of 4. All delta values were equated. The raters were asked to provide their estimates of difficulty in terms of delta values ranging from 6 to 17, at intervals of 0.5 deltas. Table 1 presents the distribution of item types in the six subsections. Table 2 presents the inter-rater correlations for prediction of item difficulty in Stage 1.

Table 1 Distribution of item types in six English subsections (Stage 1)

Sentence Completions	65
Restatements	33
Reading Comprehension	57
Total[a]	155

a. Fewer than 162 items were used in the study because the subsections rated were experimental (pilot) subsections; seven items which were subsequently identified as having item analyses that did not meet NITE's standards were deleted.

Table 2 Inter-rater correlations (Stage 1)

	Sentence Completions					Restatements					Reading Comprehension				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
R2	.88					.63					.63				
R3	.87	.79				.78	.54				.41	.26			
R4	.93	.85	.90			.82	.64	.91			.72	.67	.58		
R5	.73	.67	.74	.80		.48	.74	.47	.55		.54	.45	.33	.56	
R6	.75	.68	.71	.79	.69	.64	.58	.61	.75	.62	.61	.63	.35	.65	.52

Relatively high inter-rater correlations were found, ranging from 0.63 to 0.93 for sentence completion items, 0.47 to 0.91 for restatement items and 0.26 to 0.71 for reading comprehension items. Table 3 presents the correlation of the averaged rater predictions with item difficulty.

Table 3 Correlation of averaged rater predictions with item difficulty (Stage 1; number of items = 155)

Sentence Completions	0.57
Restatements	0.64
Reading Comprehension	0.24

It was found that of the three item types comprising the English subtest, the correlation between the averaged prediction of the six raters and actual item difficulty was highest for restatement items. The next best predictions were for sentence completion items, while the correlation for reading comprehension items was much lower. The relatively poor results for prediction of reading comprehension item difficulty – in contrast to the results obtained for the prediction of sentence completion and restatement item difficulty – prompted a closer look at the structure of reading comprehension texts and their related items.

Stage 2

Three expert raters, all native English speakers and extensively involved in the construction of the English subtest, analyzed 24 reading comprehension texts taken from PET tests. A pilot study of eight other texts was first undertaken to ensure that the raters understood the meaning of each variable and how it should be coded, and were calibrated to the same scale. The procedure used (for Stages 2 and 3) was as follows: Each rater first analyzed the text on her own; the final ratings given were those agreed upon by all raters following discussion of discrepancies. This parallels the procedure used in the actual construction of test items, in which a consensus is reached by editors following a discussion of written comments relating to a particular text or item. The raters worked with a list of variables attached to each text and rated each text according to the variables on the list. In general, the number and magnitude of the discrepancies between raters was small.

Some index of text difficulty, as opposed to item difficulty, was required; however, in the context of this study, no way of objectively measuring the former independently of the latter existed. It was decided that the dependent variable would be the average difficulty (in calibrated delta values) of all the items that appeared with a particular text. Each text had between eight and twelve accompanying items.

The texts were rated using the following nine variables (based in part on work done on prose passages by Freedle & Kostin, 1991 and Meyer, 1985):

Text-Related Variables

1. Length of text – the number of lines in a text typed in a standard format. It was hypothesized that the longer the text, the greater its difficulty.
2. Number of negations – the number of times words such as "no" or "not" appeared. Words such as "irregular" or "unknown," which had a negating prefix, were also counted. It was hypothesized that the greater the number of negations, the greater the text's difficulty.
3. Number of referential markers – Words such as "it," "this" or "then," which referred to a previously mentioned entity, were counted; pronouns were also counted. It was hypothesized that the greater the number of referential markers, the greater the text's difficulty.
4. Level of text vocabulary – Texts were rated from 1 (easy vocabulary) to 4 (difficult

vocabulary). The texts were scored based on the raters' experience with the language abilities of a non-native English-speaking population and their estimation of the relative overall difficulty of the words in the texts.[1] It was hypothesized that the more difficult the vocabulary, the greater the difficulty of the text.

5. Level of grammatical complexity of text – Texts were rated from 1 (simple) to 4 (very complex). This variable was based on factors such as sentence length; number of embedded clauses, number of markers of causality, condition, etc.; the use of ellipsis, the passive voice and punctuation other than commas and periods; and others. It was hypothesized that the greater the grammatical complexity, the greater the difficulty of the text.
6. Level of abstractness – Texts were rated on the abstractness of the content, from 1 (not abstract) to 4 (very abstract). For example, a text discussing research conducted on colds received a 1; a text discussing the historical development of the kiss received a 2; a text on the relationship between jokes and political systems received a 3; and a text on the symbolism inherent in different systems of writing received a 4. It was hypothesized that the more abstract the subject matter, the greater the difficulty of the text.
7. Topic of text – Text content was classified as relating either to the humanities (1), the social sciences (2), or the sciences (3).
8. Rhetorical structure – as described in Meyers (1985); the structure types were description, causal, comparison, response and collection. Following the rating procedure, it was decided to omit this variable from the analyses because (1) the texts being analyzed were shorter than those discussed by Meyers, and thus it was difficult to implement her categorization in this study; (2) a majority of the texts (about 90%) seemed to have a relationship of description, thus greatly reducing the effective discrimination of this variable.
9. Overall text difficulty – In addition to the above-mentioned variables, the raters were asked to provide an estimate of overall text difficulty, from 1 (easy text) to 4 (difficult text).

---

1. Word frequency lists published in English-speaking countries were not considered representative of the difficulty of English words for Hebrew speakers for two reasons: First, the frequency of occurence of particular words may be different in native and non-native English-speaking populations, and second, quite a few English words have been "adopted" into Hebrew, a fact which affects their relative difficulty. NITE has since obtained a list from the Israeli Ministry of Education in which words are designated as being either primary school, intermediate or high school level; this list will be used as a reference in future studies.

Table 4 presents the correlations of variables 1-6 and 9 with the average difficulty of all items related to a text. Table 5 presents the results of an analysis of variance for variable 7.

Table 4 Correlations of variables 1-6 and 9 with the average difficulty of all items related to a text (Stage 2; number of texts = 24)

Variable	r	p value
1. Length of text	0.11	0.5974
2. Number of negations in text	0.02	0.9329
3. Number of referential markers in text	-0.37	0.0778
4. Level of vocabulary of text	0.86	0.0001
5. Level of grammatical complexity of text	0.86	0.0001
6. Level of abstractness of content	0.23	0.2814
9. Estimate of overall text difficulty	0.90	0.0001

Table 5 Analysis of variance for variable 7 (Stage 2; number of text topics = 3; number of texts = 24)

Variable	F value	pr > F
7. Topic of text	0.11	0.5974

Results indicated that of the text-related variables mentioned above, the only ones with a significant correlation with the average difficulty of all items related to a particular text were level of vocabulary (0.86), level of grammatical complexity (0.86) and the estimate of overall text difficulty (.90).

Table 6 presents the results of a step-wise regression analysis conducted on variables 1-7 (variable 9 was excluded from the analysis).

Table 6 Step-wise regression analysis of text-related variables (Stage 2; number of texts = 24)

Dependent variable:	average difficulty of all items related to a text
R-square:	0.81
Adjusted R-square:	0.77

Variable	Estimated B values
Level of vocabulary	0.45
Level of grammatical complexity	0.53
$p < 0.05$	

The step-wise multiple regression analysis provided the same results: only level of vocabulary and level of grammatical complexity contributed significantly to explained variance.

Stage 3

The same three raters who participated in the previous stage attempted to predict the difficulty of the 229 **items** accompanying the 24 texts analyzed in Stage 2. The raters analyzed the items using the following variables:

Reading Comprehension Item-Related Variables

1. Length of stem – the number of words, including articles and prepositions, in the item's stem. It was hypothesized that the longer the stem, the harder the item.
2. Average length of distractors – the total number of words, including articles and prepositions, in all distractors, divided by 4. It was hypothesized that the greater the average length of the distractors, the harder the item.
3. Number of negations in stem – the number of times words such as "no" or "not" appeared; words beginning with a negating prefix, such as "unseen" or "inexpensive," were also counted. It was hypothesized that the greater the number of negations in the stem, the harder the item.
4. Number of negations in distractors – the total number of negations (counted as described for number of negations in stem) in all four distractors. It was hypothesized that the greater the number of negations in the distractors, the harder the item.

5. Type of stem – If the stem was phrased as an incomplete sentence, it was rated 1; if it was a full question, 2; and if structured as a sentence completion, it was rated 3. It was hypothesized that items with stems phrased as incomplete sentences would be easier than those whose stems were phrased as a complete question, and that items with stems phrased as complete questions would be easier than those whose stems were structured as sentence completions.
6. Position of key – rated as the position of the correct answer (1, 2, 3 or 4). Key positions were randomly assigned during test construction, and it was assumed that this variable would not be correlated in any way with item difficulty.
7. Level of vocabulary of stem – Stems were rated from 1 (easy vocabulary) to 4 (difficult vocabulary) according to the same criteria used to rate the texts for level of vocabulary (see Stage 2). It was hypothesized that the more difficult the vocabulary in the stem, the greater the difficulty of the item.
8. Level of vocabulary of distractors – Distractors were rated from 1 (easy vocabulary) to 4 (difficult vocabulary) according to the same criteria used to rate the texts for level of vocabulary (see Stage 2). All four distractors received one rating. It was hypothesized that the more difficult the vocabulary in the distractors, the greater the difficulty of the item.
9. Level of grammatical complexity of stem – Stems were rated from 1 (simple) to 4 (very complex) according to the same criteria used to rate the texts for level of grammatical complexity (see Stage 2). It was hypothesized that the greater the complexity of the stem, the greater the difficulty of the item.
10. Level of grammatical complexity of distractors – Distractors were rated from 1 (simple) to 4 (very complex) according to the same criteria used to rate the texts for level of grammatical complexity (see Stage 2). All 4 distractors were rated together. It was hypothesized that the greater the complexity of the distractors, the greater the difficulty of the item. However, as the length of the stems and distractors was, of course, much shorter than that of the texts, the variance in raters' scores for variables 9 and 10 was relatively small.
11. Type of item – Raters classified each question as being one of 11 different types: (1) main idea; (2) title; (3) inference; (4) sentence processing; (5) reference; (6) paragraph purpose; (7) which of the following; (8) comparative; (9) author's tone; (10) vocabulary; and (11) continuation of the text (see Appendix B for examples). It was hypothesized that items of a more global nature (such as main idea items, title items, author's tone, etc.) would be more difficult than items of a local nature (such

as vocabulary items, reference items, etc.). It was hypothesized that the difficulty of some item types (such as inference items or paragraph purpose items) could not be predicted relative to other types, and that their difficulty would depend on an interaction with other factors, such as globality or amount of processing.

12. Globality – refers to the amount of text that had to be read to answer the item. Items were rated from 1 (very local) to 4 (very global). For example, a main idea item was always scored 4, on the assumption that the whole text needed to be read to answer the question; a vocabulary question might be scored 1 if reading only one sentence was sufficient to answer it. It was hypothesized that the more global the item, the harder the item would be.
13. Amount of processing – refers to the estimated amount of thinking, or cognitive processing, necessary to answer the question. Items were scored from 1 (little processing) to 4 (great deal of processing). While global items (main idea, title) almost always involved a great deal of processing, local items differed in this respect. For example, a reference item ("In line 6, 'it' refers to - ") might require varying amounts of processing, depending on how far away the referent was from its antecedent and on the complexity of the sentence structure in which it appeared. Thus, amount of processing, while related to globality, is not necessarily equal to it. It was hypothesized that the more processing involved in answering a question, the harder the item would be.
14. "Badness" – Raters scored the items from 1 (good item) to 4 (bad item). As anyone involved in test construction knows, it is possible to discover faults in an item even after it has been pre-tested and found statistically acceptable. An item was considered "bad" if the question in the stem was phrased awkwardly or if what was being asked of the examinee was unclear; if the intended key was not a completely accurate response; or if one (or more) of the distractors was believed to be too close to being a correct response.
15. Vagueness of stem – Raters scored the item stems from 1 (very explicit) to 4 (very vague). Vagueness was defined as the extent to which the stem asked a specific question or directed examinees so that, even before reading the distractors, they had an idea of the answer required. For example, "The author claims that - " is a very vague stem. "Research has shown that the frequency of colds can be reduced by - " is an example of an explicit stem. It was hypothesized that the vaguer the stem, the harder the item would be.

Reading Comprehension Text-Related Variables

16-18. Level of text vocabulary, level of grammatical complexity of text, and estimate of overall text difficulty (variables 4, 5 and 9 in Stage 2). It was assumed that examinees answered each item after reading the related text;[2] therefore, it was clear that the variables which were shown (in Stage 2) to influence text difficulty, as measured by averaging item difficulties, would also have some effect on the text's related items. The values for these variables were taken from the analysis conducted in Stage 2.

As in Stage 2, each rater first analyzed the items on her own; the final ratings given were those agreed upon by all raters. Raters were also asked to provide estimates of each item's difficulty in terms of delta values ranging from 6 to 17 at intervals of 0.5 deltas (as in Stage 1).

Table 7 presents the inter-rater correlations for prediction of item difficulty in Stage 3.

Table 7 Inter-rater correlations for prediction of reading comprehension item difficulties (Stage 3)

	R1	R2
R2	0.62	
R3	0.71	0.67

Table 8 presents the correlations of variables 1-4, 7-10, and 12-18 with item difficulty. Table 9 presents the results of an analysis of variance for variables 5, 6 and 11.

2. A study conducted by Farr, Pritchard and Smitten (1990) showed that subjects taking a multiple-choice reading-comprehension test did not read the text first and then answer the questions; rather, they used the questions to direct a search of the text to locate the information required for answering the questions. The authors noted, however, that the subjects still had to understand the sections of the text that were relevant to selecting the right answer.

Table 8 Correlations of variables 1-4, 7-10 and 12-18 with item difficulty (Stage 3; number of items = 229)

Variable	r	p value
1. Length of stem	0.12	0.0718
2. Average length of distractors	0.22	0.0008
3. Number of negations in stem	0.00	0.9594
4. Number of negations in distractors	-0.01	0.8446
7. Level of vocabulary of stem	0.22	0.0009
8. Level of vocabulary of distractors	0.42	0.0001
9. Level of grammatical complexity of stem	0.13	0.0525
10. Level of grammatical complexity of distractors	0.04	0.5877
12. Globality	0.20	0.0028
13. Amount of processing	0.34	0.0001
14. Badness	0.26	0.0001
15. Vagueness of stem	0.01	0.8653
16. Level of text vocabulary	0.59	0.0001
17. Level of grammatical complexity of text	0.60	0.0001
18. Estimate of overall text difficulty	0.62	0.0001

Table 9 Analysis of variance for variables 5, 6 and 11 (Stage 3; number of types of stem = 3, number of key positions = 4, number of types of item = 11; number of items = 229)

Variable	F value	pr > F
5. Type of stem	1.13	0.3262
6. Position of key	2.03	0.1098
11. Type of item	2.86	0.0023

Results indicated that, of the variables listed above, the following were significantly correlated with item difficulty: average length of distractors (variable 2); level of vocabulary of stem (variable 7); level of vocabulary of distractors (variable 8); amount of processing (variable 13); badness (variable 14); and all text variables (variables 16-18). Differences between item types were also found to be significant.[3]

3. While it appeared that main purpose questions were significantly harder than other types, there were not enough items of this type (only 11 out of 229 items) to substantiate this claim. Further research is planned in which there will be an equal number of items of each type, thus enabling a better comparison to be made.

Table 10 presents the results of a step-wise regression analysis conducted on variables 1-18.

Table 10 Step-wise regression analysis of variables 1-18 (Stage 3; number of items = 229)

Dependent variable:	difficulty of each item
R-square:	0.47
Adjusted R-square:	0.44

Variable	Standardized estimated β values
Overall difficulty of text	0.47
Level of vocabulary of distractors	0.14
Amount of processing	0.16
Badness	0.13
$p < 0.05$	

It was found that 45% of the item difficulty variance could be explained by the following four variables: estimate of overall text difficulty; level of vocabulary of distractors; amount of processing; and badness. These results indicate that (as found in Stage 2) semantic and grammatical factors (overall text difficulty and level of vocabulary of distractors) play the greatest role in determining item difficulty. However, what can perhaps be defined as a measure of reasoning ability (amount of processing) plays a not insignificant role as well. Badness, which in an ideal test would not exist, also contributes significantly to item difficulty.

Table 11 presents the correlation between the raters' predictions of item difficulty and actual item difficulty.

Table 11 Correlation of rater predictions with actual item difficulty (Stage 3; number of items = 229)

R1	0.56
R2	0.61
R3	0.67
$p < 0.0001$	

In view of the marked improvement in the raters' ability to predict reading comprehension item difficulty relative to that found in Stage 1 (see Table 3), it was decided to check whether the information obtained from the three stages of the study would have an effect on the raters' ability to predict item difficulty for all types of items.

Stage 4

In essence, this stage was a replication of Stage 1, using different items. Six expert raters (five native English speakers and one native Hebrew speaker), five of whom had participated in Stage 1, were asked to estimate the level of difficulty of 162 items. Of the raters, three had performed the analyses conducted in Stages 2 and 3 of the study. To ensure that all raters had the same level of knowledge, the "new" raters were informed of the results found in those stages. The dependent variable was item difficulty, in calibrated delta values.

Table 12 presents the distribution of item types. Table 13 presents the inter-rater correlations for prediction of item difficulty in Stage 4.

Table 12 Distribution of item types in six English subsections (Stage 4)

Sentence Completions	59
Restatements	35
Reading Comprehension	61
Total[b]	155

b. See Table 1, note a, for explanation of total number of items.

Table 13 Inter-rater correlations (Stage 4)

	Sentence Completions					Restatements					Reading Comprehension				
	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5	R1	R2	R3	R4	R5
R2	.54					.72					.83				
R3	.52	.56				.76	.71				.87	.85			
R4	.53	.44	.40			.71	.81	.74			.63	.69	.61		
R5	.44	.58	.38	.38		.58	.68	.59	.76		.58	.61	.55	.56	
R6	.58	.61	.59	.50	.50	.72	.48	.49	.53	.49	.75	.73	.79	.55	.37

The inter-rater correlations found ranged from 0.33 to 0.61 for sentence completion items (much lower than what had been found in Stage 1), 0.48 to 0.81 for restatement items (about the same as in Stage 1) and 0.37 to 0.87 for reading comprehension items (higher than they had been in Stage 1).

Table 14 presents the correlation of the averaged rater predictions with item difficulty for Stage 4.

Table 14 Correlation of averaged rater predictions with item difficulty (Stage 4; number of items = 155)

Sentence Completions	0.72
Restatements	0.69
Reading Comprehension	0.82

It was found that for the three item types, the correlation between the average prediction of all six raters and actual item difficulty was highest for reading comprehension items (0.82, compared to 0.24 as found in Stage 1; see Table 3). The next best predictions were for sentence completion items (0.72, compared to 0.57 in Stage 1); the predictions for restatement items were slightly lower (0.69, compared to 0.64 in Stage 1).

CONCLUSIONS

The aim of this study was to gain a greater understanding of the factors that affect the level of difficulty of items appearing in the English subtest of the PET. The study was conducted in four stages. In Stage 1, raters predicted the difficulty of three types of items: sentence completion, restatement and reading comprehension. In Stage 2, reading comprehension texts were rated on eight variables. In Stage 3, items accompanying the texts analyzed in Stage 2 were rated on eighteen variables. Stage 4 was a replication of Stage 1, using different items.

It was found that the ability of expert raters to predict multiple-choice item difficulty varied according to the item type. In Stage 1, for sentence completions and restatements, the predicting ability was fairly high, indicating a relatively good grasp of the factors underlying item difficulty. For reading comprehension items, the initial predicting ability was much lower. In Stage 2, it was found that raters were extremely good at predicting the overall text difficulty. In view of their poor ability to predict reading comprehension item difficulty as found in Stage 1 of the study, it would have seemed likely that additional, item-related factors would have accounted for much of the difficulty in prediction. However, the raters' ability to predict reading comprehension item difficulty as found in Stage 3 was fairly high (as good as their ability to predict sentence completion and restatement item difficulty in Stage 1). It was concluded that the process of analyzing texts and items and the ensuing discussions improved the raters' ability to identify and integrate the factors relevant to difficulty and make use of them in their predictions.

The results of the analysis of the reading comprehension texts carried out in the second stage of the study showed that the two factors which were most influential in determining the level of text difficulty were level of vocabulary and level of grammatical complexity. This corroborates a conclusion reached by Klare (1984) in his review of readability research: Historically, word difficulty and syntactic difficulty have accounted for substantial variance in text difficulty scores.

Length of text was not significantly correlated with the average difficulty of all items related to a particular text (see Stage 2). While length of text might seem to be a good predictor of difficulty (a text of 20 lines would obviously be easier in a time-limited test than one of 200 lines, the length of the texts analyzed here ranged only from 17 to 28 lines (mean=22.7 and s.d.=3.24), and the variance was probably too small for this variable to be significant. Another variable which was not significantly correlated with

the average difficulty of all items related to a particular text was the number of referential markers. However, the direction of the correlation was opposite to that hypothesized (see Stage 2), and warrants some explanation. It may be may be that, in easier texts, sentence length is usually shorter, and the subjects and objects of preceding sentences are more frequently referred to by pronouns in order to avoid repetition.

Reasearch has shown that an examinee's prior knowledge does influence learning in tests of reading comprehension when the texts are in the examinee's native language (see Johnston, 1984; McNamara, Kintsch, Songer & Kintsch, 1996; Voss & Silfies, 1996). The fact that text topic, number of negations and text abstractness were not found to be significant factors in predicting the average difficulty of all items related to a particular text suggests that background knowledge in specific topics and reasoning ability may play a smaller role in determining an examinee's ability to understand a text in a foreign language than do linguistic variables such as semantic and grammatical knowledge. This statement must be qualified by the fact that, as stated above, an objective measurement of the text's difficulty could not be obtained without reference to the accompanying items.

None of the categories of text-related variables that influenced the difficulty of items, as reported by Freedle and Kostin (1991, 1992), were found significant in this study. Among other reasons, this difference may stem from the fact that, in these two studies, Freedle et al. analyzed items answered by examinees in their native language; quite different factors may influence examinees taking a test in a foreign language. However, this does not explain why few similarities were found between the significant variables influencing item difficulty as reported in this study and those contributing independent predictive information as reported in Freedle et al.'s (1993) study of TOEFL items. There may be a number of reasons for the lack of similarity: (1) Freedle et al. used a large number of very specific variables (for example, "the number of words in the key text sentence for supporting idea items," p.162) to predict item difficulty, thus making it difficult to match the two sets of variables. (2) Freedle et al. studied only three types of items – main idea, inference and supporting items – while this study included a broader range of item types. (3) The TOEFL is taken by examinees whose native languages vary considerably, whereas this study was conducted on a Hebrew-speaking population only. It is possible that what constitutes difficult vocabulary, or even grammatical complexity, for one language group is not what constitutes difficulty for other language groups. For example, different language groups take the same English subtest of the PET (the other subtests of the PET, Verbal and Quantitative Reasoning, are translated into Arabic, Russian, English, Spanish and French). Not only are there differences between the

groups' overall performance on the English subtest, but the difficulty of particular items occasionally varies greatly from group to group. It would be of interest to see what – if any – differences there are between the factors affecting item difficulty for each of these groups, and whether the homogeneity (or lack thereof) of the language spoken by the test-taking population is a factor which should be considered in future studies.

VALIDATION EVIDENCE

Despite the lack of similarity between specific variables found to be significant in this study and those in Freedle et al.'s (1993) study, there is one important point of agreement: In both studies, text or text-by-item variables were found to be the best predictors of item difficulty. In this study, it was found that reading comprehension text variables had the greatest effect on reading comprehension item difficulty, thus providing evidence for the construct validity of multiple-choice reading comprehension tests. Item-related factors that were also significant were type of item, amount of processing, badness, length of distractors, and level of vocabulary in stem and in distractors. However, some of the variables defined in this study as item variables could be said to be text-by-item variables. For instance, amount of processing and badness require scanning of both the text and the item in order to rate them. In general, it seems that both variables relating to language skills (level of vocabulary of text, level of grammatical complexity of text, level of vocabulary of stem and of distractors) and variables relating to cognitive ability (amount of processing, badness) affect the difficulty of reading comprehension items, although the former play a more significant role than the latter.

The variable called "amount of processing" was defined as the amount of cognitive processing required to answer a reading comprehension item. Although no attempt was made in this study to define a hierarchy of processing skills, it is possible that items given a low rating for amount of processing might be testing lower-order reading skills (such as word recognition, identification of specific information in text, etc.; see Alderson et al., 1989) while those given a high rating for amount of processing might be testing higher-order reading skills (such as inference-making and synthesizing abilities). In contrast to Alderson et al.'s findings, raters were able to rate items effectively according to this variable; in addition, it was shown to play a significant role in the determination of item difficulty.

The effect of the variable called "badness" on item difficulty highlights what is, perhaps, an obvious point: Great care must be taken by item writers in the phrasing of stem, key and distractors. An distractor that is "too close," a less than completely accurate key or a poorly phrased stem all increase an item's difficulty. Since this type of effect is what Messick (1993) called "construct-irrelevant difficulty," it should be avoided whenever possible.

The results of the fourth stage of the study show that, while there was a general improvement in the raters' ability to predict all three types of items, by far the greatest increment was for reading comprehension items, thus reinforcing the conclusion mentioned above: The process of analyzing texts contributes to a greater understanding of the factors affecting item difficulty. Moreover, once raters are made aware of the existence of these factors and their effects on item difficulty, they can implement this knowledge effectively in their predictions. It is highly probable that item-writers too, once they become aware of the factors affecting an item's difficulty, will be able to integrate this knowledge into their writing and will increase their ability to produce items tailored to a specific level of difficulty.

Future research will focus on increasing our understanding of the factors underlying the difficulty of sentence completion and restatement items. A comprehensive analysis of all three types of items appearing in the English subtest will further contribute to the PET's validity.

REFERENCES

Alderson, J.C., & Lukmani, Y. (1989). Cognition and reading: cognitive levels as embodied in test questions. Reading in a Foreign Language, 5, 253-270.

Bejar, I. (1983). Subject matter experts' assessment of item statistics. Applied Psychological Measurement, 7, 303-310.

Chalifour, C.L., & Powers, D.E. (1989). The relationship of content characteristics of GRE analytical reasoning items to their difficulties and discriminations. Journal of Educational Measurement, 26, 120-132.

Drum, P.A., Calfee, R.C., & Cook, L.K. (1981). The effects of surface structure variables on performance in reading comprehension tests. Reading Research Quarterly, 16, 486-514.

Embretson, S.E., & Wetzel, C.D. (1987). Component latent trait models for paragraph comprehension tests. Applied Psychological Measurement, 11, 175-193.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. Journal of Educational Measurement, 27, 209-226.

Freedle, R., & Kostin, I. (1991). The prediction of SAT reading comprehension item difficulty for expository prose passages. Princeton, NJ: ETS Research Report RR-91-29.

Freedle, R., & Kostin, I. (1992). The prediction of GRE reading comprehension item difficulty for expository prose passages for each of three item types: main ideas, inferences and explicit statements. Princeton, NJ: ETS Research Report RR-91-59.

Freedle, R., & Kostin, I. (1993). The prediction of TOEFL reading item difficulty: implications for construct validity. Language Testing, 10, 133-170.

Freedle, R., & Kostin, I. (1994). Can multiple-choice reading tests be construct-valid? Psychological Science, 5, 107-110.

Johnston, P. (1984). Prior knowledge and reading comprehension in test bias. Reading Research Quarterly, 19, 219-239.

Katz, S., Lautenschlager, G., Blackburn, A., & Harris, F. (1990). Answering reading comprehension items without passages on the SAT. Psychological Science, 1, 122-127.

Klare, G.R. (1984). Readability. In P. D. Pearson (Ed.), Handbook of Reading Research (pp. 681-744). New York: Longman.

McNamara, D. S., Kintsch, I., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition and Instruction, 14, 1-43.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (pp. 13-95). New York: Macmillan.

Meyer, B. (1985). Prose analysis: purposes, procedures and problems. In B. Britton & J. Black (Eds.), Understanding Expository Text (pp. 269-304). Hillsdale, NJ: Lawrence Erlbaum Associates.

Perkins, K., Gupta, L, & Tammana, R. (1995). Predicting item difficulty in a reading comprehension test with an artificial neural network. Language Testing, 12, 34-53.

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

Scheuneman, J., Gerritz, K., & Embretson, S. (1989). Effects of prose complexity on achievement test item difficulty. Paper presented at the meeting of the American Educational Research Association, San Fransisco, CA, March.

Stansfield, C.W., & Kenyon, D.M. (1995). In A. Cumming & R. Berwick (Eds.), Validation in Language Testing (pp. 124-153). Clevedon: Multilingual Matters.

Voss, J. F., & Silfies, L. N. (1996). Learning from history text: the interaction of knowledge and comprehension skill with text structure. Cognition and Instruction, 14, 45-68.

Appendix A

The following are examples of the types of questions used in the English subtest of the PET.

Sentence Completions

1. We cannot be certain, but the branch of philosophy known as logical reasoning ____ originated in ancient Greece.

(1) recently (2) closely (3) probably (4) normally

---

2. Atomic clocks are extremely ____, gaining or losing only a few seconds every 100,000 years.

(1) expensive (2) particular (3) dangerous (4) accurate

---

3. Many Native Americans subsisted primarily on maize and other crops, occasionally their diet with meat and fish.

(1) generating (2) supplementing (3) dismissing (4) reclaiming

---

Restatements

1. Despite our familiarity with the effects of alcohol on people's behavior, we know surprisingly little about the reasons for these effects.

(1) It is not surprising that in order to understand the effects alcohol has on people's behavior, we must also understand why it has these effects

(2) Several of the things we know about alcohol are surprising, such as how it affects people's behavior and why it has these effects

(3) If we knew more about why alcohol has certain effects on people's behavior, the fact that it has these effects would not be so surprising

(4) It is surprising how little we know about why alcohol affects people's behavior, considering how much we know about the effects themselves

---

2. In his play *Saint Joan*, as in no other, George Bernard Shaw created characters who express their emotions freely, thus facilitating the audience's identification with them.

(1) Audiences found it harder to identify with the characters in Shaw's play *Saint Joan* than with those in plays of other writers because the characters do not speak freely about their emotions

(2) Shaw felt that the characters in his play *Saint Joan* reflected his emotions, as well as the audience's, better than the characters in any of the other plays he had written

(3) It is easy for the audience to identify with the characters in Shaw's play *Saint Joan* because the characters, more than those in his other plays, express their emotions freely

(4) In all of the plays Shaw wrote, except for *Saint Joan*, the fact that the characters express strong emotions makes it easier for the audience to understand them

---

Reading Comprehension

(1) The Indians of the Andes call it "Camanchaca" – the wetting fog. When the cold morning air above Chungungo, a small village located 600 kilometres north of Santiago, collides with the warm air coming in from the Chilean coast, a ribbon of mist forms around the high mountain ridge above Chungungo. But the Camanchaca
(5) fog never turns into rain. By afternoon, the sun burns it away and the landscape remains as dry as any desert. In the past, the lack of rainfall made agriculture in this area practically impossible; vegetables had to be brought from a market located 80 kilometres away. The villagers found it difficult to obtain water even for basic necessities such as drinking and cleaning; showers were almost unheard of.

(10) Recently, however, the situation in Chungungo has changed. The village still receives only about 40 centimetres of rain a year, so the new situation is not the result of a sudden climatic shift. Rather, it is due to human inventiveness. High on the mountains overlooking the village, giant nets have been erected. Like great spider webs, these nets catch the fog, trapping drops of water in their fine mesh.
(15) The drops slowly trickle down the mesh into containers, and the water is then carried by tubes to an underground storage tank, where it is filtered.

On a good day, the "fog harvest" provides 10,000 litres of fresh water to the village. This amount is sufficient to supply a population of 1,000 with drinking water, and there is even enough left over for bathing and gardening. Flower and
(20) vegetable gardens have appeared where once there was only dust. The water costs less than it did when it was carried over the mountains by truck – the only way of supplying the village with water in the past.

Officials from Asian, African and Latin American countries have visited the area and have begun investigating the possibility of establishing their own fog-harvesting (25) systems. The Chungungo project is attractive to them because it is a cheap, practical and environmentally friendly way to bring water to poor communities.

Questions

1. The main purpose of the text is to describe -

(1) the climatic conditions in one region of Chile
(2) an original and effective method of collecting water
(3) recent developments in the water systems used by various countries
(4) the changing lifestyles of people in small villages

---

2. Why is the area around Chungungo so dry?

(1) It receives very little rainfall
(2) There was a sudden climatic shift in the area
(3) The nearby mountains block the fog
(4) It is impossible to bring water to it

---

3. The author compares the nets to giant spider webs in order to -

(1) explain how the nets were erected on the mountains
(2) show how environmentally friendly the nets are
(3) explain where the idea for the nets came from
(4) provide an image of how the fine nets trap the water

---

4. The purpose of the third paragraph is to describe the -

(1) method by which water is brought to the village
(2) improvements which have resulted from harvesting fog
(3) problems involved in supplying the village with water
(4) way in which the water system was developed

---

5. "The Chungungo project" (line 25) refers to the -

(1) visits made by officials from different countries
(2) investigation of water systems
(3) harvesting of fog
(4) plan to carry water by truck

Appendix B

The following are examples of the types of item stems used in the English subtest of the PET.

1. Main idea: "The main purpose of the text is to -"
2. Title: "An appropriate title for this text would be -"
3. Inference: "It can be inferred that scientists think that the existence of the monster -"
4. Sentence processing: "The author uses Thomas Mann (line 13) as an example of -"
5. Reference: "In line 8, 'them' refers to -"
6. Paragraph purpose: "The main idea presented in the third paragraph is that -"
7. Which of the following: "Which of the following is not a common treatment for migraine?"
8. Comparative: "The opinions presented in the second paragraph are _____ those presented in the first paragraph."
9. Author's tone: "It can be understood that the author is _____ the events he discusses in lines 20-25."
10. Vocabulary: "The words 'susceptible to' (line 4) can be replaced by -"
11. Continuation: "A continuation of the text would probably discuss -"