Is the test constructor a facet?

Abdoljavad Jafarpur *Shiraz University, Iran*

Most modern language testers believe that the writing of a successful instrument begins with specifications, and that instruments constructed devoid of specifications are likely to go astray. The purpose of this study was to explore the relative impact of the test-developer on the performance of test-takers using multiple-choice reading comprehension tests that had no specifications. Traditional reading comprehension tests often consist of short prose passages followed by sets of multiple-choice comprehension questions. The test constructor formulates the stem and the choices on the basis of his or her own renditions of the passage. The test-takers provide evidence of their comprehension in terms of their written responses to the items designed by the test constructor. Since the characteristics of the test method limit the responses the testees can provide, the expected response becomes part of the test method. Accordingly, it seems reasonable to suggest that there may be a facet associated with the test-developer. Six passages each with 3 different sets of multiple-choice items constructed by 3 (groups of) individuals were trialled on 335 Iranian EFL (English as a foreign language) learners. The results revealed differential performance on almost all sets, suggesting a test constructor effect.

I Introduction

Most modern language testers believe that the starting point for test construction is specifications and an agreed blueprint (Harrison, 1983; Weir, 1988; Hughes, 1989; Davies, 1990; Alderson *et al*., 1995; Alderson, 2000). Specifications provide an 'official statement about what the test tests and how it tests it' (Alderson *et al*., 1995: 9). Alderson (2000) believes that specifications are needed by a range of different people: item writers, test editors, test validators, test evaluators, test users, teachers, admission officials, and publishers. A blueprint 'describes how actual test tasks are to be constructed, and how these tasks are to be arranged to form the test' (Bachman and Palmer, 1996: 90). Accordingly, it is now a standard practice in test construction – whether for high-stakes or low-stakes tests – to base item writing on clearly written item specifications that give detailed guidance on the

Address for correspondence: A. Jafarpur, PO Box 71345–1354, Shiraz, Iran; email: ajafarpur@rose.shirazu.ac.ir

Language Testing 2003 20 (1) 57–87 10.1191/0265532203lt244oa

kind of items and tests to be prepared. In fact, specifications have become so indispensable to test-development that 'whether the item writing is done by one person or by a working party is incidental' (Davies, 1990: 12–13).

Nevertheless, despite the fact that specifications are now the *sine qua non* for test-development, their importance is not yet recognized in many language programs. In these programs, each individual teacher is responsible for constructing tests in the same manner that he or she shoulders teaching and class management. Even when 20 sections of one course are taught by 20 different instructors, each instructor constructs his or her test individually. The assumption is that if an individual knows how to teach, he or she also knows how to test (Spolsky, 1978). Hence, one danger of traditional multiple-choice reading comprehension tests, for instance, developed by such individual writers without any recourse to specifications is that the tests may reflect a personal idiosyncratic interpretation of a text rather than one derived from a consensus of opinions. Another danger relates to the fact that the scores obtained on these diverse tests may not be comparable.

A test of reading comprehension is a means of determining the extent to which the testee is capable of understanding written texts. Evidently, a text cannot be interpreted in isolation from the procedures humans use to produce and receive it. To be able to comprehend a text, the testee must interact with the text (Carrell and Eisterhold, 1983). With a formidable amount of information, ideas, attitudes, and beliefs (Clarke and Silberstein, 1977: 49), the testee maps the task against his or her own existing schemata. Each testee constructs a personal understanding of the text by interactively relating bottom-up messages from the text with top-down information from the schemata (Clapham, 1996). Obviously, since different test-takers bear different characteristics, they 'are likely to interact *individually* in different ways with different test tasks' (Bachman, 1990: 10; emphasis in the original).

Multiple-choice tests are by far the most common format used in the assessment of reading ability (Alderson, 2000). The input in such tests consists of a number of texts each followed by a set of multiple-choice comprehension items. Since there is no fixed set of procedures for formulating multiple-choice reading comprehension items, the decision as to what point to raise in the stem and how to word the choices is left to the discretion of the test-developer. The examinee provides evidence of successful reading by selecting the appropriate alternative from those provided by the test-developer. For this reason, many language testing researchers have long challenged multiple-choice reading comprehension tests as being too indirect a measure

since this type of test places the test writer too prominently between the stimulus materials and the testee (Harris, 1976).

By composing multiple-choice items on the basis of his or her own personal interpretation of the passage, the test constructor intervenes between the passage writer and the testees, requiring the latter – with their different individual characteristics, expectations, and perceptions – to respond to reading comprehension items that represent the test-developer's own personal discernment. That is to say, the test-developer specifies the expected response through test design and, as such, the expected response becomes part of the test method. This is tantamount to saying that the test-developer introduces 'a source of systematic variance *in addition* to the systematic variance associated with the ability we wish to measure' (Bachman, 1990: 224; emphasis in the original). Bachman further states that when the impact of test method is sizable, this clearly limits the validity of the test scores as an indicator of the individual's language ability (Bachman, 1990: 224).

II Background research

Numerous studies have explored the influence of test method on test performance. In this regard, many researchers have explored various aspects of test-takers' performance with reading instruments. Bensoussan *et al.* (1984), for example, demonstrate that text-level information (local/global) and changes in distractors (distractors with [obvious] wrong paraphrases/simpler language or distractors requiring finer distinction/more general information) may directly affect the difficulty level of the questions. Perkins (1992) shows that the topical structure of a passage affects performance. Alderson (1983), Klein-Braley (1983), and Bachman (1985) have explored the effect of different methods of deletion in cloze tests of reading. Several others have investigated the effect of the method of assessing reading comprehension. Shohamy (1984), for instance, studied the impact of multiple-choice and open-ended questions with first language (L1) and second language (L2) learners of different abilities. Davey and Lasasso (1984) studied selected response items against constructed response items. Lee (1987) investigated cloze, open-ended, and free recall tasks with L1 and L2 learners. Wolf (1993) studied multiple-choice, open-ended questions, and rationally-deleted cloze items with L2 readers. Riley and Lee (1996) examined the effect of summary and recall protocol on the reading performance of L2 readers. In these investigations, the researchers found significant effects of different task types on performance.

The effect of test-takers' background and cultural knowledge on

reading performance has been confirmed by Johnson (1982), Alderson and Urquhart (1985), Bernhardt (1985), Chihara *et al.* (1989), Gipps and Murphy (1994), and Sasaki (2000). Concluding an extensive study with IELTS (International English Language Testing System) on various factors affecting performance with reading tests, Clapham (1996: 205) states that:

> when the modules included 'general' passages, the level of language proficiency had markedly more effect on students' scores than did background knowledge. However, once the modules contained only 'specific' passages, background knowledge became proportionately more important. It might be hypothesized that, if all the subtests had been 'highly specific', background knowledge might have been made an equal or greater contribution to comprehension than language ability.

The influence of the topic of the text on sex differences in reading scores has also been confirmed by Hudson (1982), Afflerbach (1986), Lee (1986), Tan (1990), Hammadou (1991), Valencia *et al.* (1991), and Bügel and Buunk (1996), to name a few.

The effect of other aspects of the testing method on performance has been investigated by Katz *et al.* (1990), Royer (1990), Anderson *et al.* (1991), Freedle and Kostin (1993; 1999), Perkins and Brutten (1993), Perkins *et al.* (1995), Bachman *et al.* (1996), Fortus *et al.* (1998), and Lutje Spelberg *et al.* (2000). These researchers, except for Lutje Spelberg *et al.* (2000), found significant relationships between item difficulty and item characteristics. Some of the item characteristics that Freedle and Kostin (1999) investigated were: level of vocabulary, negativeness, concreteness, and topic specificity of information, syntactic complexity and cognitive demand, or amount of processing required. Freedle and Kostin postulate that, ideally, a language comprehension test 'should' assess primarily the difficulty of the text itself: the item structure should only be an incidental device for assessing text difficulty (Freedle and Kostin, 1999: 3).

Should this seemingly rational assertion be accepted, a relevant question is whether test-developers of different background (working without specifications) differ in the items they construct. In case any difference is noted among the items constructed by different item writers, one assumes that the impact of the test-developer has surpassed the difficulty of the text itself. In language programs in which test-development is each individual teacher's responsibility, the role of the test-developer appears to be more crucial than in programs in which test-development is carried out by a working party. In the former, each teacher develops his or her test according to his or her own idiosyncratic characteristics; in the latter, on the other hand, the working party prepares tests for all teachers or provides specifications

and a blueprint to be followed by individual teachers. To my knowledge, nobody has to date explored the test-developer as a facet of test variance.

III Method

1 Objectives

The aim of this article is to investigate the impact of the test constructor on test-taker performance in multiple-choice tests of reading comprehension produced without specifications. Specifically, the objective of this study is to explore how reading comprehension items constructed by different test-developers on the same text with limited moderation but without recourse to specifications compare with one another.

2 Item writers

The item writers were 6 participants in the researcher's course in language testing. They (3 male and 3 female) were Ph.D. candidates in Teaching English as a foreign language (TEFL) and with an average of 7 years in teaching English as a foreign language to Iranians of various ages and levels. They were at the time of the study teaching EFL and English for specific purposes (ESP) to undergraduates of diverse disciplines and content courses to undergraduate English majors at various universities. Moreover, each was also responsible for his or her own test-development in the same manner that he or she was for teaching and class management.

3 Test-takers

The test-takers were 335 Iranians studying at two universities in Shiraz: Shiraz University and Azad Islamic University. They had the same native language and cultural background. English majors comprised 220 of them and 105 were taking ESP courses in their major fields. The test-takers were in their twenties and of both sexes, 125 males and 210 females.

4 Materials and procedure

The materials for the study consisted of 6 passages each followed by 3 sets of reading comprehension items, 99 in all. The passages and 33 of the items were selected from the published literature (Harris, 1969: 65–66; Shohamy, 1984: 161 and 163–64; Anderson *et al.*,

1991: 65; Lynch, 1992: 353; Educational Testing Service, 1995: 34–35 and 82–84). See Appendix 1 for the passages including the original items that followed each passage. The passages were general enough to ensure that discipline-specific knowledge was not the primary factor affecting performance. They were of different lengths (41 to 278 words) and of different difficulty (86 to 23 on the Flesch scale). The passages were followed by 3, 6, 6, 5, 8, and 5 items, respectively.

To prepare the two experimental sets with a total of 66 items, each item writer was asked to prepare independently 11 items on a pair of passages they were assigned at random. The passages had been paired on the basis of the number of items that accompanied each in the original source. Table 1 summarizes the scheme used to assign the tasks to each item writer. They were not informed of the sources of the passages or of the intentions of the study. Neither were they provided with any table of specifications or any specific assistance on how to assess (any specific) reading skill. They were just asked to construct their items in accordance with the very general guidelines provided by Farhady *et al*. (1994: 247–55). This book is an introduction to language testing especially written for Iranian universities offering undergraduate programs in English. It is very much similar to Harris (1969) in terms of scope and coverage so far as assessing reading comprehension is concerned.

Once the items were complete, the researcher commented on each item with probes and prompts to stimulate the item writers themselves to improve the quality of their own items. No alterations were made by the researcher himself. The probes and prompts varied for each item writer and for each item relative to the nature of the item and the problem exhibited by the item. The problems in the draft items were numerous, diverse, and mostly serious but they gradually diminished as the item writers acted on the feedback from the researcher. All item writers were willing to a great extent to be advised and to act on information. In all, they discarded 4 items and revised most

Table 1 Scheme for preparing 66 experimental items

Item writer	Passages	Number of items	Set
A	1 and 5	3 and 8	
B	2 and 4	6 and 5	2
C	3 and 6	6 and 5	
D	1 and 5	3 and 8	
E	2 and 4	6 and 5	3
F	3 and 6	6 and 5	

others in various ways. Two of the draft items and the researcher's comments are given below by way of illustration:

Item 1: King Alfred was ____.
diligent
gregarious
bellicose
oblivious

Comment: 1) What kind of information should a 'stem' offer to the test-taker?
2) If a test-taker fails to mark the correct choice to this item, is it legitimate to conclude that he/she is unable to interpret the passage correctly?

Item 2: The achievement of King Alfred the Great was that he ____.
a. provided his people with some books
b. learned Latin as a mature adult
c. placed England in the leading position in Europe
d. provided opportunities for his people to learn

Comment: Didn't King Alfred accomplish all of the above? What about singling out one of his achievements?

The item writer who had written Item 1 discarded it completely and wrote a new one, but the writer of Item 2 simply amended the lead as 'The significant achievement . . .'

The participants' revised items along with their relevant passages were put into two tests, each with a total of 33 items. These experimental sets were trialled on145 Iranian EFL learners and the results item analyzed on the basis of sample separation. The highest scoring (top 28%) and the lowest scoring (bottom 28%) test-takers according to their total scores on each test were identified. For each item, then, the facility index, the discrimination index, and the frequency of subjects who chose each alternative were obtained. Subsequently, the participants further modified their items at their own discretion. No item was discarded at this stage, and the alterations were rather minimal and involved a few items only. As an example, the first choice of Item 2 cited above was found to be a very strong distracter; hence, the item writer merely added the word 'Latin' to it. The task resulted in two experimental sets of reading comprehension items for each passage.

Finally, to facilitate test administration for the purposes of the study, the original 33 items plus 66 experimental ones were included in 3 subtests. Each subtest contained 2 passages, and each passage was followed by 3 sets of items, 11 original and 22 experimental items. Set 1 was original, Set 2 consisted of items developed by 3 writers (A, B, and C) and Set 3 by the other 3 (D, E, and F). To avoid clues in items of one set giving away the correct choice in another, each set was printed on a separate page together with the relevant passage and the test-takers were not allowed to go back. To

control for set order effect, half of the test-takers were given the original questions first and then the others, whilst the second half were given the tests in reverse order. (No analysis was carried out afterwards to ascertain that this goal was indeed achieved.) The tests, treated as part of their course chores, were randomly distributed among the 335 test-takers in groups in their regular classes. The test-takers were allowed ample time to complete the tests.

IV Results and Discussion

Raw score descriptive statistics and reliability coefficients for all 6 passages appear in Table 2. Only the scores of the test-takers who completed all items in the 3 sets are reported. The least skewness is shown by the scores from Passage 1 and the highest by Passages 2

Table 2 Basic descriptive statistics for the 3 sets on all passages

Set	Mean (percentages = parentheses)	Median	SD	Skewness	Discrimination index	KR-21	F
Passage 1 ($k = 3$; $n = 96$)							
1	1.64 (55)	1.54	1.04	–.02	.74	.47	6.03
2	1.26 (42)	1.24	0.79	+.04	.52	–.26	$p < .003$
3	1.55 (52)	1.05	0.99	+.02	.70	.35	
Passage 2 ($k = 6$; $n = 114$)							
1	4.45 (74)	4.62	1.31	–.65	.52	.40	12.95
2	4.34 (72)	4.53	1.29	–.83	.49	.33	$p < .001$
3	3.82 (64)	3.91	1.30	–.28	.53	.21	
Passage 3 ($k = 6$; $n = 114$)							
1	4.33 (72)	6.62	1.32	–.71	.50	.37	1.04
2	4.27 (71)	4.54	1.18	–.61	.42	.14	NS
3	4.14 (69)	4.26	1.30	–.44	.48	.29	
Passage 4 ($k = 5$; $n = 105$)							
1	1.21 (24)	1.09	1.02	+.52	.50	.15	60.41
2	2.10 (42)	2.01	1.29	+.40	.63	.34	$p < .001$
3	2.73 (55)	2.61	1.35	+.07	.66	.40	
Passage 5 ($k = 8$; $n = 105$)							
1	3.17 (40)	3.09	1.63	+.27	.48	.32	4.13
2	3.66 (46)	3.34	1.65	+.53	.48	.31	$p < .02$
3	3.42 (43)	3.47	1.80	+.21	.54	.45	
Passage 6 ($k = 5$; $n = 111$)							
1	2.50 (50)	2.57	1.29	+.05	.64	.31	26.38
2	2.15 (43)	2.57	1.39	–.09	.65	.46	$p < .001$
3	1.68 (34)	1.64	1.04	+.29	.49	–.04	

Notes: Set 1 contains original items; Set 2 and Set 3 contain experimental items.

and 3. The scores representing Passages 2 and 3 are negatively skewed whereas those from Passages 4 and 5 are positively skewed. The scores from Passages 1 and 6 show both positive and negative skewness. Nonetheless, since these values are mostly less than .50, the distribution of the scores from all passages should be regarded as almost normal. However, the performance of the test-takers on the 3 sets of each passage appear different, and F statistics for ANOVA among mean scores from the 3 sets on all passages, except for Passage 3, revealed a significant difference among all sets ($p < .02$). That is, the performance of the test-takers on the comprehension questions constructed by different item writers appear to differ considerably.

Mean discrimination indices prepared on the basis of the scores obtained for each passage from each set are also shown in Table 2. These means vary from .42 to .74. Fourteen items (out of 99 cases = 33 items × 3 sets) have indices below the acceptable discriminability .30 (Harris, 1969). These 14 cases represent the most difficult items in both original and experimental sets, 6 in the original and 4 in each experimental version. It is believed that the acceptable discrimination indices of the experimental sets are mainly due to the fact that their prototype versions (discussed above under Materials and Procedure) had been trialled on sample subjects, and the items had been further improved on the basis of item analysis. Evidently, item analysis helps individual item writers to generate (new) test items with similar/acceptable psychometric qualities; nonetheless, it does not provide item writers with any means to control the unintended variation among their newly modified test items.

Examination of the test items in each of the 3 sets, particularly when carried out with multiple methods such as think aloud, introspection, content analysis, and categorization (Pearson and Johnson, 1978) can shed some light on the variability among the items constructed by different writers. Such an undertaking necessitates a large-scale study in itself, but suffice it to briefly examine the test items on Passages 3 and 4, respectively. The items on Passage 3 are the easiest and those from Passage 4 are the hardest. In addition, the original items on Passage 4 (Set 1) differ from the other 2 sets more clearly than in the other passages. Accordingly, it is important and illustrative to have a look at what exactly the items in Set 1 seem to test as compared with those in Sets 2 and 3.

Passage 3 involves facts and figures about 'tornadoes.' The comprehension items based on it are also local questions (Bensoussan *et al.*, 1984) and require no personal judgment or reading between the lines. In other words, the 3 sets on Passage 3 basically require microprocesses (Kintsch and Yarbrough, 1982), which ask for local, phrase-by-phrase comprehension. Specifically, 5 items in Set 1, 3

items in Set 2, and 4 items in Set 3 contain textually explicit questions that require the test-takers to look for both the question information and the correct answer in the same sentence. On the contrary, the 3 sets on Passage 4 involve more macro-level questions which require inferencing and a synthesis of information from various sentences. Among other subskills being tapped, Set 1 contains 3 inferencing items, Set 2 one, and Set 3 two. Set 1 also contains one scriptally implicit item (Pearson and Johnson, 1978) that requires integrating text information with background knowledge. The presence of inferencing items on Passage 4 coupled with the fact that Set 1 contains more such items than the other 2 sets on this passage support Davey (1988, cited in Alderson, 2000) who claims that difficulty can be partially accounted for by the degree of inferential processing. There is no doubt that Set 1 (the original test items taken from the published materials) contains better questions. Its items are shorter, for instance. Each item in Set 1 contains 17 words on average; the items in the other sets contain about 36 and 24 words, respectively. Set 1 contains more plausible distracters, requiring examinees to make finer distinctions between what the text conveys and the wording of the alternatives. Sets 2 and 3, on the other hand, contain more directly quoted parts of the text and distracters that the examinees could easily disregard. Table 3 summarizes the skills that each set on Passages 3 and 4 appears to tap. The categorization of the items is based solely on the researcher's own judgment. All three sets of questions on Passage 4 and the results of an item analysis for all items on Passage 4 are presented in Appendix 2.

To recapitulate, this cursory examination has demonstrated that there are inadvertent differences among the comprehension items in the original and experimental sets. These differences are believed to have been caused by item writers whose work varied due to the fact that their test items were not based on clear specifications and/or were

Table 3 Reading skills assessed by Passages 3 and 4

Reading Skill	Passage 3			Passage 4		
	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
1) Inductive reasoning to get the topic/gist	1	1	1	–	1	1
2) Understanding the syntactic structures	–	1	–	–	1	–
3) Understanding meaning from context	–	1	1	–	–	–
4) Understanding textually explicit information	5	3	4	1	2	2
5) Deducing textually implicit information	–	–	–	3	1	2
6) Deducing scriptally implicit information	–	–	–	1	–	–

Notes: Set 1 contains original items; Set 2 and Set 3 contain experimental items.

not given adequate specific editorial feedback by another individual or individuals. Apparently, each passage lends itself to tapping certain skills: Passage 3 to textually explicit questions and Passage 4 to textually implicit questions. Nonetheless, it appears that the differences in the items written over the same passage probably vary systematically from one test writer to another. In other words, the difficulty of each set is probably representative of the difficulty of the items that each specific writer writes.

To investigate the dissimilarity of the results obtained for each set, a multiple-sample analysis was carried out. For each passage, the scores from the 3 sets were divided into high (H) and low (L) ability groups. Ability level was defined by each test-taker's total score from the 3 sets. The H and L groups each contained 28% of the sample.

Table 4 presents the means, standard deviations (SDs), and standard errors (SEs) for the scores of H and L groups. Except for Set 3 of Passage 1 and Set 1 of Passage 4, the scores from Passages 1, 2, 3, and, possibly, 4 have higher SDs and SEs for L groups. The higher SEs indicate decreases in the true scores for L groups. This indicates that the comprehension questions on these 4 passages are not suitable for low ability test-takers despite their texts being easier. That is, the comprehension questions developed for Passages 1 to 4 are suitable for high-ability test-takers. On the other hand, except for Set 1 of Passage 6, the scores from Passages 5 and, possibly, 6 have higher SDs and SEs for H groups. These higher SEs delineate, in turn, a decrease in their true scores and, hence, their suitability for L ability test-takers. In other words, despite the fact that these passages are more difficult, their comprehension questions are less appropriate for more proficient test-takers. The passages had been numbered (1 through 6) on the basis of both the Flesch readability ease and the appraisal of 3 experienced EFL instructors, according to which the beginning passages (1, 2, and 3) are more appropriate for low ability and the others (4, 5, and 6) for high ability test-takers. To summarize, these results indicate that an easy text does not automatically make it a good test for the beginners, and that the difficulty of the stimulus material depends very much on what the reader/test-taker is asked to do with the stimulus. Accordingly, one may conclude that the suitability of the test items on a passage (for low-ability or high-ability test-takers) is related to the interaction between the passage and the test items.

These results support the findings of Drum *et al.* (1981), Katz *et al.* (1990), Royer (1990), and Lutje Spelberg *et al.* (2000) in that the difficulty of reading comprehension items relate primarily to item variables. The present results also partly concur with those of Freedle and Kostin (1999) in that the difficulty variance is also related to the

Table 4 Statistics for High and Low ability levels

Set	Mean		SD		SE	
	H	L	H	L	H	L
Passage 1						
1	2.74	.74	.53	.66	.10	.13
2	1.93	.67	.62	.68	.12	.13
3	2.37	.59	.74	.57	.10	.11
Passage 2						
1	5.66	3.25	.65	1.19	.12	.21
2	5.50	3.06	.51	1.16	.09	.21
3	4.97	2.75	.74	1.08	.13	.19
Passage 3						
1	5.44	2.97	.56	1.26	.10	.22
2	5.38	3.12	.71	1.10	.13	.19
3	5.00	2.94	.80	1.11	.14	.20
Passage 4						
1	2.17	0.41	.80	.57	.15	.11
2	3.38	1.14	1.12	1.13	.21	.21
3	4.10	1.69	.82	.89	.15	.17
Passage 5						
1	4.72	1.69	1.39	1.00	.26	.19
2	5.38	2.45	1.18	1.12	.22	.21
3	5.45	1.52	1.38	.91	.26	.17
Passage 6						
1	3.63	1.37	.79	.81	.14	.15
2	3.90	1.07	.80	.78	.15	.14
3	2.53	.80	.86	.55	.16	.10

Notes: H = High; L = Low. Set 1 contains original items; Set 2 and Set 3 contain experimental items.

content and structure of the passages selected for reading comprehension tests.

Sets 1 and 2 of Passage 1 show high SDs and SEs for L groups, whereas Set 3 on this passage shows a higher SD for the H group. Set 1 of Passage 4 shows higher SD and SE values for the H group, whereas the other 2 sets on this passage exhibit higher values for L groups, though those for Set 2 would not suggest a significant difference. On the contrary, Set 1 of Passage 6 shows higher SD and SE values for the L group; Sets 2 and 3 represent higher values for H groups, though the differences in Sets 1 and 2 are not likely to be significant. Taken together, these results speak of a difference between the original and the experimental items. This difference has

been caused by the variations in the items that different item writers wrote without specifications and adequate vetting.

Table 5 shows the Pearson product–moment correlation coefficients among the scores from the 3 sets on each passage. These coefficients vary from a low of .23 for Passages 1 and 6 and to a moderate of .51 for Passage 5. Whilst these low correlations may be a feature of the low number of items involved, the results seem to suggest that the items may be tapping into different underlying constructs of reading. In other words, the items constructed by different item writers without recourse to specifications and blueprints exhibit variation in the skill(s) they intend to measure.

It was surmised that the absence of uniformity in the performance of the test-takers might have been due to the difficulty of the reading tasks for some students. With multiple-choice comprehension items, it is common to assert, in both readability research and in classroom practice, that a test-taker must answer correctly at least 75% of the items over a passage before the passage is said to be suitable for his or her use. When his or her score falls between 75% and 90%, the material is said to be suitable for use in supervised instruction. A

Table 5 Correlation coefficients among the scores from 3 Sets on Each Passage

Set	1	2
Passage 1		
2	.33	–
3	.37	.23
Passage 2		
2	.49	–
3	.32	.37
Passage 3		
2	.32	–
3	.45	.24
Passage 4		
2	.32	–
3	.36	.31
Passage 5		
2	.44	–
3	.48	.51
Passage 6		
2	.50	
3	.23	.36

Notes: All significant ($p < .02$). Set 1 contains original items; Set 2 and Set 3 contain experimental items.

score below this range indicates that the material is too difficult for ordinary instructional purposes (Bormuth, 1967: 292). In order to investigate this surmise, the scores of the test-takers who had correctly answered 75% of the items on each passage were further examined. The examination of these sample means again revealed significant differences among the means for each passage ($p < .05$).

In order to see if the performance of the test-takers under investigation was in any way related to sex, independent *t*-tests were carried out between the performance of males and females on each set. Out of 18 comparisons, only 2 were found to be significantly different ($p < .05$). The scores of male and female test-takers from the third set on Passages 5 and 6 were different ($t = 4.47$ and 5.35, respectively). Generally speaking, then, these results contradict those of Hudson (1982), Afflerbach (1986), Lee (1986), Tan (1990), Hammadou (1991), Valencia *et al*. (1991), and Bügel and Buunk (1996). Since investigating the effect of sex was not the main objective of the present study and the content of the passages were neutral with respect to sex-related issues, the findings cited need to be regarded as tentative.

An investigation into the effect of background knowledge revealed no meaningful difference between the performance of English majors and that of ESP students. When respondents were divided into 3 ability groups according to their total scores on each of the 3 sets, the numbers of English majors and ESP students for all 6 passages in each group were roughly equal. Although the results were not given the acid test of significance because the comparisons involved very few students for certain passages, the trend observed appears to suggest that the scores were affected by the level of language proficiency rather than background knowledge. This finding is consistent with those of Clapham (1996) but not with Johnson (1982), Alderson and Urquhart (1985), Bernhardt (1985), Chihara *et al*. (1989), Gipps and Murphy (1994), and Sasaki (2000). Compared with these studies, the texts utilized in the present investigation were very general and were not expected to be biased for or against one sex or major field of study. That is, the results of the present study do not show any sex differences in foreign language text comprehension, most probably because the texts utilized were not about topics in the female or in the male areas of expertise/interest. Neither do the results show any effect for background knowledge in that the passages were free of content bias.

V Conclusions

The results of this investigation point to inadvertent variation in the kinds of items constructed by different test-developers working without test specifications. There is, consequently, unintended variation

in the test-taker's scores when they take different sets of items designed by different item writers. Thus, the results suggest that there is probably a facet associated with the test constructor that explains some of the variation in the test-takers' performance. Clearly, this conclusion is only suggestive.

The experimental test items used for this investigation were constructed by individual EFL teachers who had the same native tongue and cultural background. Moreover, they had undergone the same orientation and had to abide by the same very general guidelines. However, they were not provided with any specifications or blueprints and the moderation that they received was minimal. In the absence of an agreed-upon framework for formulating multiple-choice reading comprehension items and a criterion as to what point to raise in the stem and how to word the choices, this variation seems, therefore, quite natural.

In many modern testing systems, there are fixed specifications for developing test items. The specifications might differ from one another in certain respects but they do exist. Since writing multiple-choice comprehension items is not a mechanical and objective process, one could speculate that if the item constructors involved in the present investigation had been given clear specifications as well as adequate feedback on their work, the results would have shown more consensus. When specifications are not available, the item writers may circumvent what is expected of them but engage in producing tests that lead to items/interpretations that deviate from those of others. That is, in the absence of specifications and a free hand to write items, the items produced may reveal what the item writers actually do, based on their own background, education, personality, etc. In other words, item writers may produce good items in the technical sense even without the help of specifications, but they will produce different items in terms of what is asked in each question and what kind of cognitive and textual processing each question requires. On the other hand, when item writing begins with specifications, guidance, and perhaps even training, the items produced can show what the item writers can do. In other words, with adequate training and experience in test-development under the condition of being given specifications and adequate moderation, different item writers can construct test items that are more homogenous in terms of which subskills the items measure. A textbook may give the item writer some advice; it is not, however, the same as test specifications. Specifications clarify the skill(s) intended to be taught and tested. Specifications enable item writers to balance the number as well as the type of skill(s) included in the objectives. As such, specifications help minimize variability, which may be brought about as a result of

the item writers' different backgrounds and idiosyncratic interpretations.

Another possible cause of the variation might be that the test-developers had different notions of reading (Alderson, 2000). These notions might have resulted in some developers testing certain skills such as inferencing, while others focused on eliciting explicit information. Obviously, any attempt at preparing teachers to assess language performance must first acquaint them with the nature of the construct to be assessed. The truth of the matter is that the majority of practitioners who develop and use language tests, both in language classrooms and as part of applied linguistics research, still do so with little or no professional training (Bachman, 2000: 19–20).

Of course, whilst it may be possible to train item writers to produce more homogenous tests – Bachman *et al.* (1996) show this to be true for judges identifying item content – it is still possible that the test constructor's interpretation may be different from that of the testee's. If different readers do have legitimately different interpretations (Alderson, 2000), then testers need to find a way of assessing reading objectively, which at the same time caters to the rights of the examinees. In so doing, testers must also guard against the danger of instigating cloning (Alderson, 2000). That is, the outcome of training and forcing item writers to abide by fixed guidelines and specifications may simply indicate the success of the cloning process. This is in contrast with the vivacious spirit of language testing that encourages ingenuity in test construction.

All in all, the seriousness of the findings of this study must prompt administrators in such language programs to produce changes in how language tests are constructed. Administrators need to provide adequate training for teachers – particularly those responsible for test-development – in testing theory and experience in item writing. In addition, administrators need to replace reference to guidelines in testing books with the compulsory use of detailed specifications, blueprints, and moderation. As Alderson (2000) says, simple advice on item writing needs to be replaced by guidelines that recognize the complexity of the test taking process (Alderson, 2000: 91–92). Although the use of detailed specifications for classroom testing may not be feasible in general, it is definitely imperative in programs in which more than one section of a course is offered. Moreover, administrators need to attempt to establish testing groups so that a more coordinated team effort is utilized to produce tests and to resolve individual deviations to some extent. Obviously, 'the development of standards of practice and mechanisms for their implementation and enforcement' (Bachman, 2000: 19) is also in order. Such endeavors

are worthwhile given the context and importance of testing in all programs.

Teachers and test-users need to keep in mind that the test constructor could be a facet. In constructing multiple-choice items and in using test results, they must exercise every care so that the ability intended to be measured is not influenced by the test constructor facet; otherwise, this potential source of error decreases 'both the reliability of the scores and the validity of their interpretations' (Bachman, 1990: 226). One possible way to minimize this source of error is by developing a number of different items for each passage. With more items, the item writer is likely to develop items that are varied in nature and test a wider range of different skills. Test takers are often better at some skills/tasks than others. If the item writer happens to include few skills that the test-taker is good at, the performance will not be representative of the test-taker's ability at all times. The feasibility of more items per passage surely depends on what is in the test specifications as well as the extent that different subskills are possible to differentiate and specify. The specifications might define only a very limited set of subskills for certain purposes. Even when the specifications define very many subskills and a number of different texts are used in the test, asking very many questions may not be helpful: it is quite possible that some questions tap rather trivial things or too many different questions tap rather similar things or even no question taps some subskills. It is recommended that item writers make use of peer debriefings and group member checks as a way of ensuring that their reading comprehension items represent the consensus of a working party rather than the interpretation of individual teachers. It is hoped that this good testing practice will become more common in all testing situations, including the very low-stakes settings where specifications may not have to be very detailed or formalized. Moreover, a more practical approach is to safeguard against possible format effect by spreading the base of a test more widely, employing a variety of valid practical and reliable formats for testing each skill (Weir, 1988: 45).

Although the current study needs to be replicated using different test constructors, other passages, other test types, and test-takers from other language and ethnic backgrounds, a relevant question worthy of further investigation is whether test-development on the basis of specifications and moderation is helpful in minimizing diversity among tests produced by different test constructors.

Acknowledgements

Grateful thanks are extended to the six participants in the researcher's course in language testing. Test administration was made possible by

the assistance of the following colleagues to whom I remain indebted: Mortaza Yamini, Abdolhossein Parsi, Siyavash Sadravi, Aspet Minasian, Mahbube Saadat, Mohammad Rahimi and Ehya Amal-Saleh. The researcher is also immensely grateful to three anonymous *Language Testing* reviewers for their very valuable comments and stimulating suggestions, which contributed a great deal to the improvement of the earlier drafts of this article. The author is solely responsible for the remaining errors.

VI References

Afflerbach, P. 1986: *The influence of prior knowledge on expert readers' main idea construction processes*. Newark, DE: International Reading Association (ED 284 193).

Alderson, J.C. 1983: The cloze procedure and proficiency in English as a foreign language. In Oller, John W., Jr., editor, *Issues in language testing research*. Rowley, MA: Newbury House, 205–17.

—— 2000: *Assessing reading*. Cambridge: Cambridge University Press.

Alderson, J.C., Clapham, C. and **Wall, D.** 1995: *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J.C. and **Urquhart, A.H.** 1985: The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2, 192–204.

Anderson, N.J., Bachman, L., Perkins, K. and **Cohen, A.** 1991: An exploratory study into the construct validity of a reading comprehension test: triangulation of data sources. *Language Testing* 8, 41–66.

Bachman, L.F. 1985: Performance on the cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly* 19, 535–56.

—— 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

—— 2000: Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing* 17, 1–42.

Bachman, L.F., Davidson, F. and **Milanovic, M.** 1996: The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing* 13, 125–50.

Bachman, L.F. and **Palmer, A.S.** 1996: *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Bensoussan, M., Goldenblatt, L. and **Kreindler, I.** 1984: Changing the difficulty level of multiple-choice EFL reading comprehension questions. *Language Testing* 1, 105–9.

Bernhardt, E.B. 1985: Reconstructions of literary texts by learners of German. In Heid, M., editor, *New Yorker Werkstattgesprach 1984: Literarische Texte im Frendsprachenunterricht*. München: Kemmler and Hoch, 254–89.

Bormuth, J.R. 1967: Comparable cloze and multiple-choice test scores. *Journal of Reading* 10, 291–99.

Bügel, K. and **Buunk, B.P.** 1996: Sex differences in foreign language text comprehension: the role of interests and prior knowledge. *Modern Language Journal* 80, 15–31.

Carrell, P.L. and **Eisterhold, J.C.** 1983: Schema theory and ESL reading pedagogy. *TESOL Quarterly* 17, 553–73.

Chihara, T., Sakurai, T. and **Oller, J.W., Jr.** 1989: Background and culture as factors in EFL reading comprehension. *Language Testing* 6, 143–51.

Clapham, Caroline 1996: *Studies in language testing 4: the development of IELTS: a study of the effect of background knowledge on reading comprehension.* Cambridge: Cambridge University Press.

Clarke, M.A. and **Silberstein, S.** 1977: Toward a realization of psycholinguistic principles in the ESL reading class. *Language Learning* 27, 135–54.

Davey, B. 1988: Factors affecting the difficulty of reading comprehension items for successful and unsuccessful readers. *Experimental Education* 56, 67–76.

Davey, B. and **Lasasso, C.** 1984: The interaction of reader and task factors in the assessment of reading comprehension. *Experimental Education* 52, 199–206.

Davies, A. 1990: *Principles of language testing.* Oxford: Blackwell.

Drum, P.A., Calfee, R.C. and **Cook, L.K.** 1981: The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly* 16, 486–514.

Educational Testing Service 1995: *TOEFL practice tests*. Princeton, NJ: Educational Testing Service.

Farhady, H., Jafarpur, A. and **Birjandi, P.** 1994: *Testing language skills: from theory to practice.* Tehran: SAMT Publications.

Fortus, R., Corriat, R. and **Fund, S.** 1998: Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test. In Kunnan, A.J., editor, *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum, 61–87.

Freedle, R. and **Kostin, I.** 1993: The prediction of TOEFL reading item difficulty: implications for construct validity. *Language Testing* 10, 133–70.

—— 1999: Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing* 16, 2–32.

Gipps, C. and **Murphy, P.** 1994: *A fair test? Assessment, achievement and equity?* Buckingham: Open University Press.

Hammadou, J. 1991: Interrelationships among prior knowledge, inference and language proficiency in foreign language reading. *Modern Language Journal* 75, 27–38

Harris, D.P. 1969: *Testing English as a second language*. New York: McGraw-Hill.

—— 1976: Testing reading comprehension in ESL: background and current

state of the art. In Morely, J., editor. *Papers in ESL: selected conference papers of the Association of Teachers of English as a Second Language*. Washington, DC, NAFSA, 25–30.

Harrison, A. 1983: *A language testing handbook*. Bethleham, PA: ELTS.

Hudson, T. 1982: The effects of induced schemata on the 'short circuit' in L2 reading: non-decoding factors in L2 reading performance. *Language Learning* 32, 2–31.

Hughes, A. 1989: *Testing for language teachers*. Cambridge: Cambridge University Press.

Johnson, P. 1982: Effects of reading comprehension on building background knowledge. *TESOL Quarterly* 16, 503–16.

Katz, S., Lautenschlager, G., Blackburn, A. and **Harris, F.** 1990: Answering reading comprehension items without passages on the SAT. *Psychological Science* 1, 122–27.

Kintsch, W. and **Yarbrough, J.C.** 1982: Role of rhetorical structure in text comprehension. *Educational Psychology* 74, 828–34.

Klein-Braley, C. 1983: A cloze is a cloze is a question. In Oller, J.W., Jr., editor, *Issues in language testing research*. Rowley, MA: Newbury House, 218–28.

Lee, J.F. 1986: Background knowledge in L2 reading. *Modern Language Journal* 70, 350–54.

—— 1987: Comprehending the Spanish subjunctive: an information processing perspective. *Modern Language Journal* 71, 50–7.

Lutje Spelberg, H., De Boer, P. and **Van den Bos, K.P.** 2000: Item type comparisons of language comprehension tests. *Language Testing* 17, 311–22.

Lynch, B. 1992: Evaluating a program inside and out. In Alderson, J.C. and Beretta, A., editors, *Evaluating second language education*. Cambridge: Cambridge University Press, 61–99 and 350–66.

Pearson, P.D. and **Johnson, D.D.** 1978: *Teaching reading comprehension*. New York: Holt, Rinehart and Winston.

Perkins, K. 1992: The effect of passage topical structure types on ESL reading comprehension difficulty. *Language Testing* 9, 163–72

Perkins, K. and **Brutten, S.R.** 1993: A model of ESL reading comprehension difficulty. In Huhta, A., Sajavaara, K. and Takala, S., editors, *Language testing: new openings*. Jyväskylä: University of Jyväskylä, 205–18.

Perkins, K., Gupta, L. and **Tammana, R.** 1995: Predicting item difficulty in a reading comprehension test with an artificial neutral network. *Language Testing* 12, 34–53.

Riley, G.L. and **James, F.L.** 1996: A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing* 13, 173–89.

Royer, J. 1990: The sentence verification technique: a new direction in the assessment of reading comprehension. In Legg, S. and Algina, J., editors. *Cognitive assessment of language and math outcomes*. Norwood, NJ: Ablex, 144–91.

Sasaki, M. 2000: Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing* 17, 85–114.

Shohamy, E. 1984: Does the testing method make a difference? The case of reading comprehension. *Language Testing* 1, 147–70.

Spolsky, B. 1978: Introduction: linguists and language testers. In Spolsky, B., editor, *Advances in language testing series: 2: approaches to language testing*. Arlington, VA: Center for Applied Linguistics, pp. v–x.

Tan, S.H. 1990: The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In de Jong, J.H.A.L. and Stevenson, D.K., editors. *Individualizing the assessment of language abilities*. Clevedon: Multilingual Matters, 214–44.

Valencia, S.W., Stallman, A.C., Commeryas, M. and **Hartman, D.K.** 1991: Four measures of topical knowledge: a study of construct validity. *Reading Research Quarterly* 16, 204–33.

Weir, C. 1988: *Communicative language testing*. Exeter: University of Exeter.

Wolf, D. 1993: A comparison of assessment tasks used to measure foreign language reading comprehension. *Modern Language Journal* 77, 473–89.

Appendix 1 Six passages and original items

Passage 1: Anderson *et al*., 1991: 65 (with permission)
During the 1950s, each TV season offered 39 weeks of new shows, and 13 weeks of repeats. Slowly, the ratio has reversed. The ultimate goal may be a one-week season, 50 weeks of repeats, and one week off for good behavior.

1. The main point the writer is making is that ____.
   a) television shows are being repeated more often than ever
   b) shows must be repeated to allow to prepare new shows
   c) repeated shows are used to gain good ideas for new shows
   d) repeating shows cuts down the costs
2. When did the change in television that the passage describes take place?
   a) During the past year
   b) Only very recently
   c) Over a period of time
   d) Several years ago

3. What does the writer most probably think of the situation in television that he or she is telling us about?
   a) It is better than it was before.
   b) It cannot be helped.
   c) It may soon improve.
   d) It is becoming ridiculous.

Passage 2: Shohamy 1984; 161–64 (with permission)

READERS' LETTERS, Sir, I visited my old school yesterday. It hasn't changed in thirty years. The pupils were sitting in the same desks and reading the same books. When are schools going to move into the modern world? Books belong to the past. In our homes, radio and television bring us knowledge of the world. We can see and hear the truth for ourselves. If we want entertainment, most of us prefer a modern film to a classical novel. In the business world, computers store information so that we no longer need encyclopedias and dictionaries. But in the schools, teachers and pupils still use books. There should be a radio and television set in every classroom and a library of tapes and records in every school. The children of today will rarely open a book when they leave school. The children of tomorrow won't need to read and write at all. M.P. Miller, London

1. The general idea of the passage is that ____.
   a) the best books were written in the past
   b) schools use the same books instead of modern ones
   c) people no longer need to get information through books only
   d) children don't need to go to school
2. This letter was written to ____.
   a) a television program
   b) the writer's old school
   c) a friend from school
   d) the newspaper
3. Which of the following best describes the writer's feelings? He is ____.
   a) pleased
   b) confused
   c) dissatisfied
   d) confident
4. According to the writer, how does his school look today?
   a) Every room has a computer.
   b) Children learn only through books.
   c) Children learn with the aid of television.

d) Tapes are used instead of books.

5. According to the writer, why do 'Books belong to the past'?
   a) Because information can be gotten through other sources.
   b) Because they are more expensive than in the past.
   c) Because tapes are easier to carry than books.
   d) Because information in books is more correct.
6. According to the author, the best place to get information from is ____.
   a) a computer
   b) an encyclopedia
   c) a novel
   d) a film

Passage 3: Lynch, 1992; 353 (with permission)

The smallest but most intense of all known storms is the tornado. It seems to be a typically American storm since it is most frequent and violent in the United States. Tornadoes also occur in Australia in substantial numbers and happen occasionally in other places in midlatitudes.

The tornado is a small, intense cyclone in which the air is spiralling at tremendous speed. It appears as a dark funnel cloud hanging from a cumulonimbus cloud. At its lower end, the funnel may be 300 to 1500 feet (90 to 460 meters) in diameter. The funnel appears dark because of the density of condensing moisture, dust, and debris swept up by the wind.

Wind speeds in a tornado exceed anything known in other storms. Estimates of wind speed run as high as 250 miles (400 km) per hour. As the tornado moves across the country, the funnel writhes and twists. The end of the funnel cloud may alternately sweep the ground, causing complete destruction of anything in its path, and rise in the air to leave the ground below unharmed.

1. An appropriate title for this passage is ____.
   a) Varieties of Storms
   b) Tornadoes
   c) Wind Speed
   d) Cloud Formations
2. Among storms, the tornado is said to be the ____.
   a) largest
   b) smallest
   c) most widespread
   d) least intense
3. Tornadoes occur primarily in ____.

a) Australia
b) higher latitudes
c) lower latitudes
d) America

4. Tornadoes resemble a dark funnel proceeding from ____.
   a) an intense cyclone
   b) writhing and twisting
   c) a cloudy sky
   d) dust and debris
5. The darkness of a tornado's appearance is due to ____.
   a) cloud type
   b) wind speed
   c) debris density
   d) cloud cover
6. Wind speeds of a tornado are known to reach ____.
   a) 400 miles per hour
   b) 250 miles per hour
   c) 460 meters per second
   d) 1500 feet per minute

Passage 4: Harris, 1969; 65–66 (with permission)

In the development of literature, prose generally comes late. Verse is more effective for oral delivery and more easily retained in the memory. It is therefore a rather remarkable fact that English possessed a considerable body of prose literature in the ninth century, at a time when most other modern languages in Europe had barely developed a literature in verse. This unusual accomplishment was due to the inspiration of one man, King Alfred the Great, who ruled from 871 to 899. When he came to the throne, Alfred found that the learning which in the previous century had placed England in the forefront of Europe had greatly decayed. In an effort to restore his country to something like its former state, he undertook to provide for his people certain books in English, books which he deemed most essential to their welfare. In preparation for this task, he set about in mature life to learn Latin.

1. According to the information given in the paragraph, King Alfred may most probably be regarded as the father of English ____.
   a) poetry
   b) learning
   c) prose
   d) literature
2. The writer suggests that the earliest English poetry was ____.

a) written in very difficult language
b) not intended to be read silently
c) never really popular with the public
d) less original than later poetry

3. According to the paragraph, England's learning had brought it to the 'forefront of Europe' (lines 9 and 10) in the ____.
   a) seventh century
   b) eighth century
   c) ninth century
   d) tenth century
4. The writer suggests that at the time of King Alfred most of the other modern languages of Europe had ____.
   a) both a verse and a prose literature
   b) a literature in prose but not in verse
   c) neither a prose nor a verse literature
   d) a literature in verse but not in prose
5. We may conclude that the books which Alfred 'deemed most essential' were ____.
   a) already available in another language
   b) written largely in verse
   c) later translated into Latin
   d) original with Alfred himself

Passage 5: Educational Testing Service, 1995; 82–84 (with permission)

Are organically grown foods the best choices? The advantages claimed for such foods over conventionally grown and marketed food products are now being debated. Advocates of organic foods – a term whose meaning varies greatly – frequently proclaim that such products are safer and more nutritious than others.

The growing interest of consumers in the safety and nutritional quality of the typical North American diet is a welcome development. However, much of this interest has been sparked by sweeping claims that the food supply is unsafe or inadequate in meeting nutritional needs. Although most of these claims are not supported by scientific evidence, the preponderance of written material advancing such claims makes it difficult for the general public to separate fact from fiction. As a result, claims that eating a diet consisting entirely of organically grown foods prevents or cures disease or provides other benefits to health have become widely publicized and form the basis for folklore.

Almost daily the public is besieged by claims for 'no-aging' diet, new vitamins, and other wonder foods. There are numerous unsubstantiated reports that natural vitamins are superior to synthetic ones,

that fertilized eggs are nutritionally superior to unfertilized eggs, that untreated grains are better than fumigated grains, and the like.

One thing that most organically grown food products seem to have in common is that they cost more than conventionally grown foods. But in many cases consumers are misled if they believe organic foods can maintain health and provide better nutritional quality than conventionally grown foods. So there is really cause for concern if consumers, particularly those with limited income, distrust the regular food supply and buy only expensive organic foods instead.

1. In line 5, the work 'others' refers to ____.
   a) advantages
   b) advocates
   c) organic foods
   d) products
2. The phrase 'welcome development' mentioned in line 7 is an increase in ____.
   a) interest in food safety and nutrition among North Americans
   b) the nutritional quality of the typical North American diet
   c) the amount of healthy food grown in North America
   d) the number of consumers in North America
3. According to the first paragraph, which of the following is true about the term 'organic foods'?
   a) It is accepted by most nutritionists.
   b) It has been used only in recent years.
   c) It has no fixed meaning.
   d) It is seldom used by consumers.
4. The word 'unsubstantiated' in lines 17 and 18 is closest in meaning to ____.
   a) unbelievable
   b) uncontested
   c) unpopular
   d) unverified
5. The word 'maintain' in line 24 is closest in meaning to ____.
   a) improve
   b) monitor
   c) preserve
   d) restore
6. The author implies that there is cause for concern if consumers with limited incomes buy organic foods instead of conventionally grown foods because ____.
   a) organic foods cost more but are often no better than conventionally grown foods

b) many organic foods are less nutritious than similar conventionally grown foods
c) conventionally grown foods are more readily available than organic foods
d) too many farmers will stop using conventional methods to grow food crops

7. According to the last paragraph, consumers who believe that organic foods are better than conventionally grown foods are often ____.
   a) careless
   b) mistaken
   c) thrifty
   d) wealthy
8. What is the author's attitude toward the claims made by advocates of health foods?
   a) Very enthusiastic
   b) Somewhat favorable
   c) Neutral
   d) Skeptical

Passage 6: Educational Testing Service, 1995: 34–35 (with permission)

A distinctively American architecture began with Frank Wright, who had taken to heart the admonition that form should follow function, and who thought of buildings not as separate architectural entities but as parts of an organic whole that included the land, the community, and the society. In a very real way, the houses of colonial New England and some of the southern plantations had been functional, but Wright was the first architect to make functionalism the authoritative principle for public as well as for domestic buildings. As early as 1906, he built the Unity Temple in Oak Park, Illinois, the first of those churches that did so much to revolutionize ecclesiastical architecture in the United States. Therefore, he turned his genius to such miscellaneous structures as houses, schools, office buildings, and factories, among them the famous Larkin Building in Buffalo, New York, and the Johnson Wax Company building in Racine, Wisconsin.

1. The phrase 'taken to heart' in line 2 is closest in meaning to which of the following?
   a) Taken seriously.
   b) Criticized.
   c) Memorized.
   d) Taken offence.

2. In what way did Wright's public buildings differ from most of those built by earlier architects?
   a) Their designs were based on how they would be used.
   b) Their materials came from the southern United States.
   c) They looked more like private homes.
   d) They were built on a larger scale.
3. The author mentions the Unity Temple because it ____.
   a) was Wright's first building
   b) influenced the architecture of subsequent churches
   c) demonstrated traditional eccelesiastical architecture
   d) was the largest church Wright ever designed
4. The passage mentions that all of the following structures were built by Wright EXCEPT ____.
   a) factories
   b) public buildings
   c) offices
   d) southern plantations
5. Which of the following statements best reflects one of Frank Wright's architectural principles?
   a) Beautiful design is more important than utility.
   b) Ecclesiastical architecture should be derived from traditional designs.
   c) A building should fit into its surroundings.
   d) The architecture of public buildings does not need to be revolutionary.

Appendix 2: Three sets of questions

Set 1: Original questions (Harris, 1969: 65–66 [with permission]) The items tap subskills 5, 4, 6, 5, and 5 (listed in Table 3), respectively.

1. According to the information given in the paragraph, King Alfred may most probably be regarded as the father of English ____.
   a) poetry
   b) learning
   c) prose
   d) literature
2. The writer suggests that the earliest English poetry was ____.
   a) written in very difficult language
   b) not intended to be read silently
   c) never really popular with the public
   d) less original than later poetry
3. According to the paragraph, England's learning had brought it to the 'forefront of Europe' in the ____.
   a) seventh century

b) eighth century
c) ninth century
d) tenth century

4. The writer suggests that at the time of King Alfred most of the other modern languages of Europe had ____.
   a) both a verse and a prose literature
   b) a literature in prose but not in verse
   c) neither a prose nor a verse literature
   d) a literature in verse but not in prose
5. We may conclude from the paragraph that the books which Alfred 'deemed most essential' were ____.
   a) already available in another language
   b) written largely in verse
   c) later translated into Latin
   d) original with Alfred himself

Set 2: Experimental items

The items tap subskills 1, 4, 5, 4, and 2 (listed in Table 3), respectively.

1. Which of the following is an appropriate topic for the paragraph?
   a) The Unusual Development of Prose in England
   b) Comparison of the Development of Prose and Verse in Europe
   c) How Verse develops Memory Retention
   d) Talents of Alfred the Great, the King of England
2. A majority of European nations did not possess even ____ literature in the ninth century.
   a) oral
   b) prose
   c) written
   d) verse
3. The significant achievement of King Alfred the Great was that he ____.
   a) provided his people with books in Latin
   b) learned Latin as a mature adult
   c) placed England again in the leading position in Europe
   d) provided opportunities for his people to learn Latin
4. What was the situation in England immediately before the king's rule?
   a) English had a great body of literature in prose.
   b) Learning was a central issue in the society.
   c) The majority of people paid no attention to learning.

d) England was a leading European country.

5. 'This task', in line 13, 'means 'providing people with ____.
   a) welfare
   b) books in English
   c) formal education
   d) books in Latin

Set 3: Experimental items

The items tap subskills 5, 4, 5, 1, and 4 (listed in Table 3), respectively.

1. England had an important position in Europe because of ____.
   a) paying attention to verse before prose
   b) possessing a considerable amount of prose
   c) encouraging foreign language education
   d) giving more importance to verse
2. The inspiration of King Alfred caused ____.
   a) the development of literature in prose
   b) people to become interested in verse
   c) people to study other languages
   d) an unusual achievement in the seventh century
3. We can infer from the passage that in most European languages, prose developed ____.
   a) before verse
   b) after the 9th century
   c) when Alfred came to the throne
   d) between 871 and 899
4. The main idea of the passage is that ____.
   a) learning Latin helped people to develop a literature in prose
   b) English literature has been the richest since the 10th century
   c) King Alfred the Great ruled England from 871 to 899
   d) English Literature owes a great deal to King Alfred
5. To make his country famous again, King Alfred started to ____.
   a) teach Latin
   b) read more books
   c) provide certain books for the people
   d) improve people's well being

Item Analysis for all items on Passage 4: sample separation based on total scores on each set [N=105{H(igh)=29, M(iddle)=47, L(ow)=29}]. Bold figures indicate correct options.

Set:	Item	Choice A			Choice B			Choice C			Choice D			Omit	Facility	Discrimination
		H	M	L	H	M	L	H	M	L	H	M	L			
Set 1:	1	1	4	3	5	10	13	**22**	**14**	**1**	1	18	10	3	.35	.72
	2	5	14	9	**7**	**5**	–	9	10	6	7	11	9	13	.11	.24
	3	2	2	1	**18**	**19**	**1**	8	23	25	1	3	2	–	.36	.59
	4	4	6	7	3	9	4	**8**	**5**	–	14	26	17	2	.12	.28
	5	**19**	**11**	–	3	10	12	5	13	8	2	11	8	3	.28	.66
Set 2:	1	**25**	**21**	**5**	–	15	9	2	3	5	2	7	8	3	.49	.69
	2	1	7	8	6	17	13	3	5	6	**19**	**16**	**2**	2	.35	.66
	3	2	6	7	3	5	5	**20**	**19**	**3**	4	16	13	2	.40	.59
	4	9	18	15	2	2	2	**18**	**18**	**4**	–	6	8	3	.38	.48
	5	1	9	11	**25**	**20**	**4**	2	7	7	1	11	7	–	.47	.72
Set 3:	1	–	9	3	**25**	**27**	**8**	2	4	6	2	6	12	1	.58	.59
	2	**28**	**23**	**6**	–	16	7	1	4	8	–	3	5	4	.54	.76
	3	4	17	8	**22**	**16**	**2**	3	10	14	–	4	4	1	.38	.69
	4	2	13	13	1	8	4	–	3	4	**26**	**22**	**3**	5	.49	.79
	5	–	9	7	–	1	2	**28**	**35**	**15**	–	2	3	3	.74	.45

Notes: H = High; M = Middle; L = Low