Topic Variable in Narrow-Scope EAP Short-Context Reading Tests

Hasan Ansary & Esmat Babaii

Shiraz University

Abstract

The concern for selecting topics that may tap test takers' optimal performance has been a fundamental consideration in reading assessment. To date, several studies have documented the fact that the topic of input has a substantial effect on test takers' performance. Much in the same line, the present study was an attempt to explore (a) the extent to which EAP students' reading performance on a discipline-specific short-context reading test is influenced by their knowledge of subject-matter of the test tasks, (b) the impact(s) that narrow-scope discipline-specific test tasks may have on the psychometric characteristics of the test, and (c) the suitability of short-context method of testing reading comprehension in an EAP context. To this end, two separate 20-item multiple-choice short-context reading tests comprising 20 items on statistics and 20 items on discipline-neutral topics were developed and administered to 53 Iranian college-level students of (1) statistics, (2) chemistry, and (3) history. Results demonstrated that students of statistics significantly outperformed students of history and chemistry. Furthermore, topic specificity of the test tasks did not seem to improve the psychometric properties of the discipline-related test. And short-context testing appeared to function equally well in an EAP context.

Key words: ESP testing, topic variable, short-context testing, EAP reading test, test qualities.

Introduction

> We contend that all topic-based tests must investigate the possibility of a topic effect as part of the ongoing process of test validation.
>
> (Jennings *et al.*, 1999:427)

Recently there has been an increasing consensus of opinion among many language test developers that tests should, where possible, be

related to candidates' future language needs (cf. Clapham, 1993; Alderson & Clapham, 1993, among others). For example, future lawyers might be tested on the legal English needed in the court. Much in the same line of argument, many test writers now consider that a language test for university students should, therefore, contain samples of the kinds of language tasks required of them in their academic work.

Perhaps, no one questions the sensibility of such an approach to English for Specific Purposes (ESP) testing. However, it is not clear to what extent the 'subject matter' or 'topic' should vary according to the discipline of the test takers. Whether students in different academic disciplines should take separate discipline-specific tests or a single discipline-neutral test and whether discipline-related tests produce superior psychometric characteristics are questions open to investigation. Some evidence (Weir, 1983) shows that the language tasks in different academic disciplines are significantly similar, but it is not clear whether the topic of the test should be similar as well.

Seen from a different perspective, performance on language tests varies as a function both of an individual's language ability and of the facets of test method (Bachman, 1990). Test performance is also affected by individual attributes that are not part of test takers' language ability. These may include test takers' prior *topical knowledge*, their age, sex, L1, cognitive and affective characteristics, etc. Language test writers consider factors such as these as potential sources of *test bias*. Relating topical knowledge to language test performance, Bachman (1990) has suggested a framework of test method facets which includes facets of the (a) testing environment, (b) test rubric, (c) input, (d) expected response, and (e) relationship between the input and response. It essentially "provides much of the context of language tests, [and] affects performance on language tests" (p. 119).

In considering the specific test qualities that determine the overall usefulness of a given language test, Bachman & Palmer (1996: 25) define "interactiveness" as a quality which refers to "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task." The degree of interactiveness of a given

test can hence be characterized in terms of the ways in which test taker's areas of language knowledge, *topical knowledge*, cognitive and affective attributes, communication strategies, etc. are engaged by the test task.

Therefore, individuals' topical knowledge as an index of the test quality of interactiveness and as a source of test bias needs to be considered in a description of test performance. For, it seems that certain test tasks that presuppose topical knowledge may be easier for those who have that knowledge, and more difficult for those who do not. For instance, a reading passage that includes a great deal of subject matter information specific to a particular discipline might be more difficult for individuals who do not possess the relevant topical knowledge than those who do. This *topic effect* in language tests constitutes what Messick (1989, 1996) refers to as "a source of construct-irrelevant variance." Therefore, investigating the possibility of a topic effect appears to be a critical step in establishing the validity of language tests.

Background

Several studies have documented the fact that the topic of input has a substantial effect on test takers' performance. In fact, topic variables (i.e., subject matter, text structure, mode of discourse, rhetorical organization,...) have been thoroughly investigated from different angles (cf. Hoetker, 1982; Carrell, 1984, 1985; Tan, 1990; Clapham, 1993, among many others). However, a clear understanding of the relationship of different aspects of topic variable to reading performance has not yet been reached as such.

Studies of the impact on test performance of topic (Bachman & Palmer, 1983; Erickson & Molloy, 1983; Alderson & Urquhart, 1983, 1985, 1988; Koh, 1985; Hale, 1988; Tan, 1990; Clapham, 1993; to name but a few) have so far focused on the effect(s) of separate aspects of the topic. For example, in the oral interview test of communicative proficiency developed by Bachman & Palmer (1983), the test maker and test taker go through a list of topics to determine

which is best suited to the interviewee's interests and knowledge in order to elicit candidate's optimal performance.

Erickson and Molloy's (1983) paper was a report on a group effort to develop two discipline-specific ESP language tests which involved six graduate students in UCLA's M.A. TESL program working in pairs. Each pair developed one listening comprehension test based on a classroom lecture and one reading comprehension test based on a course textbook. All teams used materials from undergraduate engineering courses. This study basically dealt with the performance of engineers and nonengineers (including native and nonnative speakers of English in both categories) on test items aimed at engineering content or at English itself. Their results, inter alia, showed that engineers outperformed nonengineers on both the specialized and the English language questions. The contrast, however, was greater on the items that dealt with engineering content. Also the natives outscored the nonnatives on all tests. They concluded that "language does play an important role in performance on the test, despite its being a discipline-specific test in an area not typically considered language-dependent" (p.287).

Alderson and Urquhart (1983, 1985, 1988) described three studies on the effect of background topical knowledge on test results carried out with overseas students attending English classes in Britain in preparation for going to British universities to pursue various fields of study. In each study, they compared students' scores on reading tests related to their own field of study with scores on tests in other subject areas. Subjects (Ss) took the Social Studies and Technology Modules of the English Language Testing Service (ELTS) test. The Ss' scores on the modules were found to be contradictory. They concluded that background topical knowledge had an inconsistent effect on test scores.

It appears that Alderson and Urquhart's investigations aroused much interest and led to several follow-up studies. For example, Koh (1985) administered a set of cloze tests to three groups of students, two in science and one in business studies at Singapore University. Examining the effect of background topical knowledge on test scores, she found that students did not always do best in their own subject

area. Besides, students with the highest language proficiency did consistently better than the other two on all the texts. Koh concluded that prior subject knowledge did affect test scores but that students with high linguistic knowledge could compensate for their unfamiliarity with the subject matter.

Hale (1988) looked at candidates taking the Test of English as a Foreign Language (TOEFL) over four administrations to investigate the interaction between students' major field of study and text content. The texts were aimed at the general reader but were based on a range of topics in the arts and sciences. Ss were divided into two groups: (1) humanities/social sciences and (2) biological/physical sciences. Hale found that for three of the four test forms the effect of subject area was significant at .001. He wrote that Ss' reading performance was affected by a combination of their major field area and the nature of the texts.

Tan (1990) explored whether familiarity with test content or level of language proficiency was the best predictor of ability in reading comprehension. She administered three tests to undergraduates at the University of Malaysia: (1) a "prior knowledge" test developed by the subject teachers, (2) a discipline-related cloze reading test, and (3) a "general" proficiency test. Tan studied three subject areas— medicine, law, and economics. She found that both knowledge of the subject matter and language level could predict comprehension of a discipline-related text. But language level was found to be a better predictor.

Clapham (1993) described a pilot study in which performances on two different subject modules of IELTS test, one within the Ss' own subject area and one outside it, were compared. Using a repeated measures ANOVA design, in her whole sample she found that the mean of advantage— Ss supposedly being advantaged because they took tests in their own subject area— was higher but not significantly so. Subject area, she wrote, therefore did not significantly affect test results. In her Indonesian sample, however, the results were different: Ss did seem to do better in their own subject module than in the other. The subject area effect was significant at .01. As such, the results were

contradictory and difficult to credit. Clapham (1993: 267) finally suggested that there is "no need for three academic subject modules in the IELTS test battery."

In all, it appears that findings in this area are still contradictory and inconclusive. And, we know of no study investigating the possible effect(s) of topic variables on test psychometric characteristics. Therefore, further research is well motivated.

The Current Study

The purpose of this study is threefold. First, an attempt is made to investigate the extent to which EAP students' reading performance on a discipline-specific short-context reading test is influenced by their knowledge of subject-matter of the test task. Secondly, within norm-referenced testing every attempt is made to maximize variance. As such, the specificity of topic in discipline-specific tests may influence the psychometric properties of the test, resulting in differences in reliability indices, the item discrimination values, etc. Therefore, this study also attempts to shed some illuminating light on the possible impact(s) that narrow-scope discipline-specific test tasks may have on the psychometric characteristics of the tests.

Finally, short-context reading comprehension tests are often claimed to be reasonably reliable, authentic, and valid measures of reading ability which do not suffer from the flaws of the traditional reading tests on both practical and theoretical grounds (see Corrigan *et al.*, 1978; Jafarpur, 1987). However, the question remains as to whether this technique functions well in the construction of EAP discipline-specific tests too. With this in mind, this study aims to offer some insights into whether or not the short-context technique of testing reading comprehension works well in an EAP context. As such, three research questions are explored in the present study:

1. ***Does discipline-specific topical knowledge have any effect(s) on EAP students' reading comprehension scores?***
2. ***Does discipline-specific topical knowledge have any effect(s) on psychometric characteristics of reading tests as well?***

3. *Is the short-context technique an authentic test method for testing reading comprehension in an EAP context?*

On account of the non-availability of conclusive findings, three null hypotheses are formulated: First, it is hypothesized here that in tests of EAP reading comprehension, in cases where subject-matter of the input is familiar to some students but not others, they do not significantly outperform those to whom subject-matter is unfamiliar. Secondly, by the same token, it is speculated that psychometric characteristics of discipline-specific reading tests are not influenced by test takers' prior topic familiarity. And it is finally postulated that discipline-specific short-context tests of reading ability are not appropriate in an EAP context.

Method

Subjects

A group of 53 senior Iranian undergraduate students of statistics (N = 23), chemistry (N = 16), and history (N = 14) at Shiraz University participated in this study. They were divided into three groups. Group One (students of statistics) were assumed to be familiar with their discipline-related (statistics) test tasks. Group Two (students of Chemistry) were presumed to have some rudimentary knowledge of statistics in L1, but they could not be as competent in discipline-specific (statistics) test tasks as Group One. Group Three (students of history), in contrast, were supposed to lack such a subject-matter knowledge. Subjects were all speakers of the Persian language whose ages ranged from 22 to 32. Sex was not a variable in this study.

Instrumentation

Three instruments were used in the study: (1) a 35-item reading subtest, Module C, Test No. 1, of the International English Language Testing Service (IELTS, 1995) test, (2) a newly-developed 20-item Discipline-Specific Short-context Reading Test (DST) comprising a collection of discipline-specific (statistics) excerpts taken from a statistics textbook (Howell, 1989), and (3) a newly-developed 20-item

Discipline-Neutral Short-context Reading Test (DNT) incorporating a variety of texts taken from a TOEFL workbook (Davy & Davy, 1984). DST (see Appendix 1) and DNT (see Appendix 2) were constructed according to the guidelines given by Jafarpur (1987: 197-8). The reading section of IELTS test comprising four long passages and a map and containing a total of 35 completion test items was, in fact, administered as a criterion to establish the concurrent validity of DST/DNT.

Procedure

The data collection procedure for this study was carried out in two separate phases in which repeated reading ability scores were obtained from Ss. To reduce unwanted error variance produced by, for example, fatigue; the IELTS test was administered on one occasion which corresponded to the first phase. Ss were given 50 minutes to complete this test. The results of this phase served well for validation purposes. In the second phase, however, which took place one week later, the Ss responded to the prompts on both DST and DNT tests during a 60-minute testing session. To shun order effect(s), a counterbalanced design was utilized. And, as expected, there arose the problem of subjects dropout in data gathering. That is, some subjects were lost due to the low show-up rate on the second occasion of testing.

Results

Using Pearson Product-Moment correlation formula, the IELTS test scores as a criterion were correlated with those of DST and DNT tests in order to establish the empirical validity of the newly-developed tests. Table 1 below presents the results. As it is evident, moderate but highly significant correlation coefficients indicate a satisfactory concurrent validity of the instruments used to elicit the data.

Other psychometric properties of the DST and DNT tests including mean Item Facility (IF), mean Item Discrimination (ID), and reliability indices (KR-21) were also estimated. Table 2 displays the

obtained values. Overall, it appears that the newly-developed tests enjoy acceptable indices, hence serving well the research purposes.

Table 1
Correlation between the Three Tests
(N = 30)

	DNT	IELTS	DST
DNT	—		
IELTS	.73*	—	
DST	.67*	.70*	—

P < .001

Table 2
Summary Statistics for Total Sample on the Three Tests
(N = 53)

	K	Mean	SD	KR-21	Mean IF	Mean ID
DNT	20	6.74	3.59	.69	.34	.72
DST	20	9.36	4.07	.73	.47	.73
IELTS	35	16.3	7.52	.87	—	—

Table 3, however, displays summary test statistics only for the scores of Group One (the students of statistics) on the three tests. The reason for this particular analysis is to examine the (de)merits of using discipline-specific tests for measuring language ability of the students from a similar discipline background. It clearly shows that DST, in comparison to DNT, does not enjoy a high reliability index. Besides, a close look at IF and ID indices leads to two conclusions: (a) DST test, in contrast to DNT test, is easier for Group One (the students of statistics), and (b) it has less discriminatory power.

Finally, in order to determine the possible main effect and interactions between disciplines consisting of three levels (statistics, chemistry, and history) and tests having two levels (discipline-specific vs. discipline-neutral), a 3 × 2 Multivariate Analysis of Variance

(MANOVA) was performed (see Table 4 below). It produced statistically very significant F values at the $p < .001$. Consequently, a Scheffe test was run to locate the source(s) of differences (see Table 5 below).

Table 3
Summary Statistics for the Students of Statistics on the Three Tests (N =23)

	K	Mean	SD	KR-21	Mean IF	Mean ID
DNT	20	7.83	4.17	.77	.40	.78
DST	20	12.35	3.17	.55	.61	.65
IELTS	35	20	8.74	.91	—	—

Table 4
Results of MANOVA by Discipline and Test

Source	df	SS	MS	F
Between Ss				
Field (A)	2	385.27	192.64	11.77*
Ss within Groups	50	817.99	16.36	
Within Ss				
Test (B)	1	124.51	124.51	28.62*
Test by Field (AB)	2	109.73	54.87	12.61*
B × Ss within Groups	50	217.49	4.35	

P < .001

Table 5
Results of Scheffe's Post-Hoc Multiple Comparison Procedure on Scores

Mean		1	2	3	4	5
12.35	S_{DST}					
7.83	S_{DNT}	12.2*				
8.62	CH_{DST}	2.17*	-.72			
6.00	CH_{DNT}	5.77*	1.06	4.94*		
5.29	H_{DST}	3.79*	2.11*	1.52	.51	
5.79	H_{DNT}	5.46*	1.09	2.02*	.096	.82

P< .05

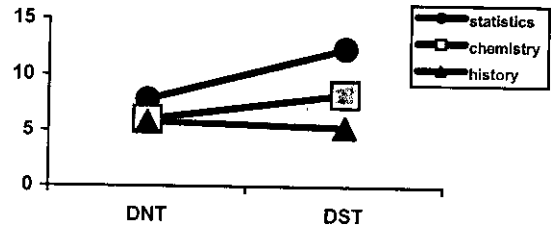
As can be seen in Table 5, with respect to within-group comparisons, Group One (students of statistics) scored considerably higher on the topic-familiar (DST) test than on the topic-neutral (DNT) test (S_{DST} vs. S_{DNT}). Group Two (students of Chemistry) also did so (CH_{DST} vs. CH_{DNT}). Topical knowledge, therefore, emerges as a powerful factor in reading comprehension test performance for both the statistics and the chemistry students. However, as expected, Group Three (students of history) appeared to perform similarly on both tests (H_{DST} vs. H_{DNT}).

As regards between-group comparisons, Table 5 shows that no statistically significant difference exists between the groups on the DNT test (S_{DNT} vs. CH_{DNT} , S_{DNT} vs. H_{DNT} , CH_{DNT} vs. H_{DNT}), whereas Group One's performance was significantly different from both Groups Two and Three on the DST test (S_{DST} vs. CH_{DST} , S_{DST} vs. H_{DST}). Furthermore, the test performance of both Group Two and Group Three on the DST test was not significantly different (CH_{DST} vs. H_{DST}).

Fig. 1 below, nevertheless, seems to be a more straightforward visual representation of the results of Scheffe's comparison procedure.

Figure 1

Mean Performance of the Three Groups



Discussion

The overall results of this study give rise to a number of points. First, as to the first research question, the main concern was whether

students would be advantaged if they took reading tests based on subjects within their own areas of study. Clearly, data (cf. Tables 4 & 5) suggested that the statistics students did better on their own discipline-related test tasks. This is, in fact, referred to as 'test bias' to the disadvantage of those test takers who lack relevant topical knowledge. It is what test writers often try to avoid as a source of construct-irrelevant variance. In all, these results are analogous to the findings of Erickson & Molloy (1983) and Hale (1988). And they are also congruent with Clapham's (1993) observations on her culturally-homogeneous Indonesian sample.

Much in the same line, a result of interest is that although the students of history did not perform significantly different on both the discipline-specific and discipline-neutral tests (see Table 5), they mostly questioned the fairness of a language test containing specialized content not related to their own area of specialization. Therefore, it can be concluded that for language assessment purposes test developers may, chameleon-like, either (a) administer a discipline-neutral test to test-takers who are heterogeneous in terms of topical knowledge in order to, to the extent possible, eliminate test bias, or (2) administer a discipline-related test only to the test-takers who are homogeneous in terms of their subject matter topical knowledge.

A second related point (cf. second hypothesis) is that this study yielded almost no evidence that relevance of test content to the test takers' background topical knowledge contributes much, if any, to the psychometric properties of the discipline-specific test used to tap reading comprehension ability of test takers from similar subject-matter knowledge background (cf. Table1 and Table 3). Specifically, the DST test, in comparison to DNT test, obtained a reliability index of .55. Low reliability of such tests is, in fact, expected. Because a topic-related test administered to a homogenous group does not produce much variance.

This finding, however, should not lead test makers to question the usefulness of ESP field-specific tests altogether. Because, within the framework of test usefulness proposed by Bachman & Palmer (1996), reliability is only one quality out of the six necessary qualities of a

useful test. Furthermore, test usefulness must be determined by striking an appropriate balance among the qualities of a test for each specific testing situation. In other words, an individual test quality must not be evaluated independently. Rather, the combined effect of all qualities on the overall usefulness of a test must be established . As such, in cases where there is a discipline-specific test enjoying the following characteristics (1) the relevance of the test tasks to the test takers' background topical knowledge (authenticity), (2) the potential to involve test takers' linguistic and cognitive capacities in the process of test taking (interactiveness), and (3) the psychologically positive responses of the test takers (face validity), test writers are just advised to weigh up these qualities and strike an appropriate balance among these and other less satisfactory psychometric qualities. Such decisions can justify the degree of usefulness of a particular test in a particular testing situation.

There seems to exist a paradox here. While, discipline-specific test tasks were not found to be an improvement over discipline-neutral test tasks, they appeared to be highly face valid in the eyes of the students from similar discipline background. Therefore, if testing is more than just the scores (see Jennings *et al.,* 1999: 451) and if theories of motivation suggest that the correspondence between content and students' (topical) background knowledge is beneficial in that it enhances language learning and in that it helps test makers tap test takers' optimal language test performance, the more subtle psychological effects of such tests must not be overlooked, though there is evidence to the contrary.

A final point of interest, addressing the third research question, is that consistent with the previous findings (Jafarpur, 1987, among others), this study provided further evidence in support of the conviction that short-context technique of testing reading comprehension works equally well in an EAP context (cf. Table 1, Table 2 & Appendix 3).

One final word of caution, however, seems in order here. The findings of present study should be judged judiciously. One should be, in fact, cautious about the interpretation of the results of this study.

Because, the scope of the study was narrow and the number of subjects small. Besides, this study did not take account of the subjects' differing language proficiency levels. There can exist evidence that the effect of subject matter knowledge on test scores may not be consistent over the whole range of proficiency levels. Further research may, therefore, offer a clearer idea of the effect of subject area on reading comprehension test performance. Perhaps, another study examining larger samples of students at different stages of language proficiency to investigate the interaction of second language proficiency level and subject area can shed more light on this area. A further study may look more closely at the performance of test takers on individual reading texts and/or on individual test tasks to examine whether some or all sections of the tests are biased for or against students in differing disciplines. IRT item analyses can be used for this purpose.

Acknowledgment

Throughout the travail of writing this manuscript, we have always paid careful attention to the advice, or head-bashing, if you will, of an individual who has often been of great help to us: Dr. A. J. Jafarpur, Professor of language testing in the Department of Linguistics and Foreign Languages at Shiraz University whose careful reviews of the tests that we developed improved them immensely. He must therefore share the credit for what we have offered here. However, we assume responsibility for the probable error(s).

References

Alderson, J. C. and Clapham, C. (eds). (1993). ***Examining the ELTS test: An account of the first stage of the ELTS revision project.*** London & Camberra: University of Cambridge Local Examination Syndicate and International Development Program.

Alderson, J. C. and Urquhart, A. H. (1983). Effect of student background discipline on comprehension: A pilot study. In Hughes, A. and Porter, D. (eds) ***Current developments in language testing,*** pp. 121-127. London: Academic 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.

Alderson, J. C. and Urquhart, A. H. (1988). This test is unfair: I'm not an economist. In Carrell, P. L., Devine, J., and Eskey, D. E. (eds) ***Interactive approaches to second language reading.*** Cambridge: Cambridge University Press.

Bachman, L. F. (1990). ***Fundamental considerations in language testing.*** Oxford: Oxford University Press.

Bachman, L. F., and Palmer, A. S. (1983). ***Oral interview test of communicative proficiency in English***. Urbana, ILL.: Photo-offset.

Bachman, L. F., and Palmer, A. S. (1996). ***Language testing in practice.*** Oxford: Oxford University Press.

Carrell, P. (1984). Evidence of a formal schemata in second language comprehension. ***Language Learning.*** 34: 87-112.

Carrell, P. L. (1985). Facilitating ESL reading by teaching text structure. ***TESOL Quarterly.*** 19: 727-752.

Clapham, C. (1993). Is ESP testing justified? In Douglas, D. and Chapelle, C. (eds) ***A new decade of language testing research,*** pp. 257-272. Virginia: TESOL Inc.

Corrigan, *et al.* (1978). ***English placement tests, Form A, B, and C.*** Ann Arbor: Michigan University Press.

Davy, E., and Davy, K. (1984). ***TOEFL reading comprehension and vocabulary workbook.*** New York: Prentice Hall Press.

Erickson, M., and Molloy, J. (1983). ESP test development for engineering students. In Oller, J. W. (ed.) ***Issues in language testing research***, pp. 280-288. USA, Rowley: Newbury House Publishers, Inc.

Hale, G. (1988). Student major field and text content: Interaction effects on reading comprehension in the Test of English as a Foreign Language. ***Language Testing.*** 5: 49-61.

Hoetker, J. (1982). Essay examination topics and students' writing. ***College Composition and Communication.*** 33: 337-392.

Howell, D. C. (1989). ***Fundamental statistics for the behavioral sciences* (2nd ed.).** USA, Boston: PWS-KENT Publishing Co.

Jafarpur, A. (1987). The short-context technique: An alternative for testing reading comprehension. ***Language Testing.*** 4/2:195-212.

Jennings, M., Fox, J., and Graves, B. (1999). The test-takers' choice: An investigation of the effect of topic on language-test performance. ***Language Testing.*** 16/4: 426-456.

Koh, M. (1985). The role of prior knowledge in reading comprehension. ***Reading in a Foreign Language.*** 3: 375-380.

Messick, S. (1989). Validity. In Linn, R. L. (ed) ***Educational measurement.*** New York: American Council on Education/ Macmillan.

Messick, S. (1996). Validity and washback in language testing. ***Language Testing.*** 13: 241-256.

Tan, S. (1990). The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In de Jong, J., and Stevenson, D. (eds) ***Individualizing the assessment of language abilities***, pp. 214-224. Clevedon and Philadelphia: Multilingual Matters.

Weir, C. (1983). ***Identifying the language problems of overseas students in tertiary education in the UK.***. Unpublished doctoral dissertation. University of London.

Appendix 1

Sample Items of Discipline- Specific Test (DST)

To convert to a decimal, divide the numerator by the denominator.

1. Which one is an elaborate example of the above rule?

A. $1/5 = 0.20$

B. $\frac{8}{1/3} = 8\left(\frac{3}{1}\right) = 24$

C. $\frac{16}{4} = \frac{8}{2}$

D. $\frac{2 \cdot 8}{4} = \frac{2}{8} \cdot 8 = 4$

I have a bag of caramels hidden in the drawer of my desk which contains 85 of the light caramels and 15 of the dark ones. I reach into the bag and grab a caramel at random.

2. The probability of the occurrence of an event (A) is A/(A+B). What is the probability that I will pull out a light-colored caramel?

A. $15 \div (85 + 15) = .15$
C. $(85 \div 85) + (85 \div 15) = 6.6$
B. $85 \div (85 + 15) = .85$
D. $(15 \div 15) + (15 \div 85) = 1.1$

A percentile is the point on a scale at or below which a given percentage of the scores fall. The percentile rank of a score, on the other hand, is the percentage of cases falling at or below that score. Assume that 83% of people taking a test had a score less than or equal to 74.

3. In a case like this, 74 is called the....

A. percentile rank
C. 83rd percentile
B. percentage
D. 74th percentile

For a contingency table, the expected frequency for a given cell is obtained by multiplying together the totals for the row (R) and the column (C) in which the cell is located and dividing by the total sample size (N). These totals are known as marginal totals. Now study the following table.

(Cell 1) **30**	(Cell 2) **20**	**50**
(Cell 3) **10**	(Cell 4) **40**	**50**
40	**60**	**100**

4. What are the marginal totals for cell 2 ?

A. 20 and 60
C. 60 and 50
B. 50 and 100
D. 20 and 40

5. The expected value for cell 4 is ....

A. 60
C. 25
B. 30
D. 100

Appendix 2

Sample Items of Discipline-Neutral Test (DNT)

Nat Turner was born in Africa in 1800. As a young man, he organized a group of fellow slaves in a violent uprising in which eighty-five whites were killed.

1. Nat Turner was famous as a person who ....

A. became a politician in his young age
B. fought for freedom for slaves
C. killed many white people
D. traded his fellow slaves

The government of China has announced that consumers may soon purchase TV sets and other expensive items on the installment plan. No interest will be charged when the plan is initiated. However, should the necessity arise, interest payments will be added later.

2. What is implied in the passage about the installment plan in China?
A. It is managed by the government.
B. It is of no interest to people.
C. It will decrease the prices.
D. It is a new plan.

3. Installment payments will be interest free….
A. forever
B. for the moment
C. for the next year
D. for a decade

Joan is fourteen years old, a bright student, and suffering from self-starvation. She has anorexia nervosa. *Anorexia* means "without appetite," and *nervosa* means "of nervous origin."

4. We can conclude that the root of anorexia nervosa is probably….
A. lack of appetite
B. adolescence
C. psychological problems
D. physical deficiencies

The veterinarian and the psychologist have joined forces to redress the behavioral ills of dogs. Subject to the same emotional problems as their owners, dogs have increasingly developed neuroses formerly attributable to humans.

5. Dogs resemble people in their inability to….
A. work together
B. develop neuroses
C. live in restricted spaces
D. cope with emotional change

Appendix 3
Item analysis
The whole sample (n = 53)

DST

	1	2	3	4	5	6	7	8	9	10
IF	0.74	0.72	0.81	0.57	0.43	0.3	0.19	0.6	0.43	0.53
ID	0.61	0.71	0.54	0.67	0.8	0.29	0.33	0.67	0.86	0.67
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
IF	0.42	0.53	0.38	0.21	0.36	0.42	0.38	0.34	0.49	0.47
ID	0.71	0.76	1	0.8	0.83	0.75	0.92	1	0.87	0.81

DNT

	1	2	3	4	5	6	7	8	9	10
IF	0.42	0.38	0.34	0.34	0.47	0.43	0.26	0.17	0.28	0.33
ID	0.64	0.71	0.91	0.5	0.69	0.8	0.67	0.5	0.78	0.63
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
IF	0.26	0.6	0.26	0.4	0.26	0.28	0.42	0.43	0.28	0.23
ID	0.8	0.76	0.73	0.75	0.83	1	0.81	0.64	0.57	0.75

Item Analysis
Group One, Students of Statistics (n = 23)

DST

	1	2	3	4	5	6	7	8	9	10
IF	0.91	0.91	0.91	0.65	0.61	0.22	0.13	0.78	0.57	0.61
ID	0.54	0.54	0.54	0.7	0.67	0.33	0.67	0.5	0.75	0.86
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
IF	0.52	0.68	0.48	0.35	0.52	0.48	0.57	0.74	0.78	0.7
ID	0.63	0.58	1	0.6	0.78	0.67	0.75	0.7	0.45	0.78

DNT

	1	2	3	4	5	6	7	8	9	10
IF	0.55	0.48	0.52	0.35	0.65	0.52	0.26	0.22	0.26	0.32
ID	0.67	0.86	0.57	1	0.6	0.86	0.83	0.25	1	0.4
	11	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
IF	0.26	0.57	0.3	0.35	0.39	0.3	0.61	0.43	0.35	0.26
ID	0.75	0.71	1	0.83	1	1	0.67	1	0.5	1