JOURNAL OF ENGLISH STUDIES
Islamic Azad University, Science & Research Branch, 1(2), 3-19**,** Spring 2011

Examining the Features of an English Language Test: Reliability-related Issues

Esmaeel Abdollahzadeh
Iran University of Science and Technology

ABSTRACT: Most universities across Iran tend to develop English tests for placement, exit, achievement, and other purposes. Examining the various features of such tests is imperative for making informed decisions about learners' achievement level. The present study examined the features of a university-wide administered English language achievement test at Iran University of Science and Technology (IUST). The characteristics of the test including reliability, item facility/difficulty, and item discrimination were examined. The reliability estimates showed that the test had a relatively acceptable level of reliability. However, it was relatively easy, and the majority of the items had a rather low discrimination power. Moreover, all the subsections of the test were found to contribute significantly to the overall achievement of the participants. It was concluded that more attention needs to be paid to the difficulty and discrimination characteristics of such achievement tests assigned on a large scale across universities. Suggestions are made about how more accountable decisions can be made to have fair tests for language achievement purposes.

Keywords: achievement, language skills, language test, proficiency, reliability

There are different types of language tests like achievement, knowledge, and proficiency tests, each developed for a particular purpose. These tests are used to make decisions about the abilities of individuals, to place them in different educational settings based on their ability, or to check the success or failure of a program. Examination of the characteristics of a test can have both theoretical and applied consequences in our understanding of the nature of language proficiency, language teaching, and language learning (Bachman, 1990). Language achievement tests are designed to measure the body of knowledge the students are supposed to achieve through a course or courses of study (Farhady, Jafarpour, & Birjandi, 1995). They are used for making judgments about the success of an instructional program or for evaluation purposes. Language proficiency tests attempt to measure an individual's ability to understand and use the language. They are used to make placement into or exit from a program, or to judge a persons' language level for educational or other opportunities.

The present study, using classical testing theory measures, tries to examine the English language achievement test administered to the undergraduate engineering university students at Iran University of Science and Technology. A general English achievement test is administered at the end of a semester-long English course at this university, and all the entrants into this university must take this exam. To a large extent, performance on this exam will determine the success or failure in English for the students of science and technology.

The motivation behind this study was that since this test is used to make judgments about the success or failure of a large number of students at the end of an English course, an assessment of the characteristics of this test in terms of its reliability, item difficulty, item discrimination, and of the performance on its subsections seems warranted. The results of this analysis can enlighten the instructors and faculty members as to the quality of this test and the possible shortcomings and problems associated with it. This achievement test is a multiple-choice one measuring reading, grammar, vocabulary, and listening skills. In what follows, first a review of the relevant literature on these skills is presented. Then, the research questions, along with the methodology employed, are discussed, followed by conclusions that could be drawn from the findings.

Furthermore, recent research on the role of gender in reading comprehension shows differences between the two sexes. For instance, Brantmeier (2003) showed that subject matter familiarity interacted with gender and affected reading comprehension of intermediate-level EFL readers. Bugel and Buunk (1996) found that females achieved higher scores for a list of text topics that included a housewife's dilemma and males scored higher on text topics that included sports. However, the literature on male/female performance, especially in reading comprehension, is scarce in EFL contexts. Therefore, in this study, an attempt was made to discover any differences in performance due to gender differences.

Background

Reading

Reading comprehension has gained tremendous importance in SLA research and pedagogy both as a critical component of academic learning and as a primary means of learning on ones' own beyond the classroom (Carrell & Grabe, 2002). Reading in L2 seems to vary from the L1 reading processes. As such, L1 research findings are not directly applicable in L2 reading contexts. Grabe and Stoller (2002) categorize L1 versus L2 reading differences in three groups: (a) linguistic and processing differences, (b) individual and experiential differences, and (c) differing socio-cultural backgrounds of L2 readers. L2 readers use a bilingual mental lexicon of some type and possess a wide range of L2 proficiencies. For example, Nassaji's (2003) study of Persian ESL readers showed that all the high and low processing levels of reading (phonological, orthographic, and lexico-semantic) contributed significantly to the discrimination between these readers. In addition, individual and experiential differences, which refer to the amount of exposure to L2 print, motivations, kinds of texts in L2 settings, and language learning resources play a crucial role in reading. Differing socio-cultural backgrounds of L2 readers, different patterns of discourse and text organization, as well as expectations of educational institutions in L1 and L2 settings should also be considered.

There are many other factors involved in reading. For example, reading rate studies and its effect on comprehension, level of L2 language proficiency required to perform effectively in reading comprehension, the role of L1 reading strategies in L2 reading comprehension, the role of background knowledge or schemata in reading comprehension, the complex interactions between background knowledge and other variables like interest and text structure, knowledge of metacognitive strategies and how to use those strategies in conjunction with other strategies as an index of successful reading (Anderson, 1991), as well as the role of extensive reading in L2 reader progress.

The IUST English course book for engineering students essentially includes content-based reading passages on issues in science and technology. The reading practice is expected to provide the students with the orientation content towards science and technology issues in general, followed later by their more specific discipline-based ESP courses assigned nearly at the end of their undergraduate years at IUST.

Vocabulary

The IUST coursework devotes a significant portion of its practice to teaching the vocabulary items. This involves word definition and exemplification sections in the textbook, followed by fill in the blanks exercises based on list of given items, as well as matching exercises. Vocabulary tests can be used as a means of proficiency measure, short-term/long-term achievement measure, or as diagnostic measure (Bensoussan & Laufer, 1984). The role of context in guessing the meaning of words is an important line of research on vocabulary. Laufer's (2006) study on incidental vocabulary learning from reading showed that explicit memorization and focus on vocabulary forms can increase learners' performance to a significant degree. Bensoussan and Laufer (1984) argue that word guessibilty is less a function of context than some preconceived notions about the new vocabulary and those more proficient learners, despite their greater vocabulary knowledge, are not any more effective than less proficient students in the use of context. Moreover, what is guessable context for native speakers may be an incomprehensible context for non-native speakers (Haynes, 1984). The effect of dictionary use in L2 vocabulary learning is also yet to be examined. Research results have shown that, like guessing from context, it may not be effective for more proficient students, because it is not an automatic skill but needs to be practiced to be used effectively (Carrell & Grabe, 2002). Carrell and Grabe argue that fluency across the four language skills can be enhanced through vocabulary acquisition drawing on knowledge already acquired. This can be done either through repeated practice with the previously learned items, or using those items in a wide variety of contexts and situations and thus making many connections and associations with a known item. Moreover, very close relations have been found between the vocabulary size and language proficiency.

Grammar

Grammar is generally defined as a set of dos and don' ts that tell us what to say and what not to say. Models of grammar can be formal or functional. Celce-Murcia and Larsen-Freeman (1999) argue that grammar models should take account of not only form and function but meaning as well. Thus, their model which is more comprehensive to the present researcher, includes form (morpho-syntax), meaning (semantics), and use/function (pragmatics). The kind of grammatical knowledge presented and promoted in the engineering students' English textbook is essentially pedagogical grammar or grammar for academic purposes designed for the needs of L2 learners. Pedagogical grammar is eclectic and draws on the findings of formal and functional grammars and on corpus linguistics, discourse analysis, and pragmatics (Deccarico & Larsen-Freeman, 2002). The presentation of grammatical points in the IUST English textbook, nonetheless, is a traditionally-oriented one. First, a grammatical point common in academic discourse is presented and exemplified. Then, the students are required to change the structure of the sentence prompts based on the grammar point discussed.

Teaching grammar can be done both explicitly and implicitly through noticing, consciousness-raising, and focus on form techniques. Of course, grammar instruction must involve what Larsen-Freeman (2002) calls 'grammaring'; that is, asking students to use particular structures in meaningful communicative tasks. All in all, there should be a match between teachability and learnability. As such, traditional linear-fashion grammar instruction has little room in the recent trends of grammar instruction.

Listening

The listening coursework in IUST is essentially a task-based listening course. It involves the students' listening to some theme-based conversations by native speakers and answering relevant questions based on each conversation. This practice is a semester-long, one-hour session run on a weekly basis.

Listening involves perception of prosodic patterns, recognition of sounds, and interpretation of what is said in terms of topic and in relation to ones' language and world background (Lynch & Mendlesohn, 2002). It is assumed to be the most difficult task amongst native and non-native learners because the listener must draw upon a wide variety of sources, linguistic and non-linguistic, to interpret the rapidly incoming data (Graham, 2006). Listening involves the application of the linguistic knowledge (bottom-up processing), and of the topical knowledge and world knowledge interacting with the linguistic knowledge to get to an interpretation of the text (Buck, 2001; Rost, 2002).

Rubin (1994) highlights some important factors affecting listening comprehension:

> (1) text characteristics, such as temporal/acoustic variables like speech rate, pause, and hesitation; (2) interlocutor characteristics like interlocutors' gender, culture, place of origin; (3) task characteristics, such as multiple-choice, Wh-questions, open-ended questions; (4) listener characteristics, which refer to language proficiency level, memory, attention, gender, affective variables, background knowledge, and learning disabilities, and (5) process characteristics, which refer to how listeners interpret input in terms of what they know or identify what they do not know. It can also refer to the way in which listeners use different kinds of signals to interpret what is said. (p. 210)

These processes can be top-down, bottom-up, and parallel processing. Listening is assumed to share many important characteristics with reading in terms of comprehension processes involved (Bae & Bachman, 1998; Vandergrift, 2006). Listeners, however, need to comprehend spoken language which is more demanding, cognitively speaking (Buck, 2001).

Vandergrift (2006) found that both L1 listening ability and L2 proficiency were significant predictors of L2 listening comprehension ability with only L2 proficiency accounting for a significant amount of the variance.

Research Questions

1. To what extent are the IUST English Achievement Test (EAT) and its subparts reliable?
2. To what extent can knowledge of the different subsections of the IUSTEAT exam contribute to the overall performance of the participants?
3. Do the IUSTEAT items have acceptable levels of difficulty and discrimination?
4. Is there any significant difference between male and female engineering undergraduates on the IUSTEAT and its subparts?

Method

Participants

One hundred fifty four college level students studying engineering sciences from IUST took part in this research. They were selected from six classes of 35 pupils, 154 of whom had complete participation in all the study's test administration sessions. All the participants were taking the English for the Students of Engineering course, a three-credit course offered by the English Department that lasts for 32 sessions of 90 minutes for a whole semester. The students take this preliminary obligatory course to get prepared for the ESP course they will take later on. Students whose English scores on the nationwide entrance exam are low are required to take a remedial pre-university English course before taking this 3-credit course.

Instruments

Three instruments were employed in this study:

1. The Nelson language proficiency test (Version 300): This test consists of reading, grammar, vocabulary, and pronunciation sections. It includes 33 cloze-test reading items based on two reading passages, 12 grammar items, and five pronunciation items. All the test items are in multiple choice format. This test was used to examine the concurrent validity of the IUST English achievement test.

2. The IUST English Achievement Test (EAT): The EAT is a 50-item test developed by the Department of English to examine the language achievement of the students. The test is administered at the end of the term and all the undergraduate freshmen taking the English course should sit for it, and their performance on the test will constitute a large portion of their final exam score. This 50-item test consists of 22 reading items based on three passages (15 items are based on two reading passages and the other 7 on a cloze reading test). The vocabulary section of this test consisted of 15 multiple choice vocabulary items. Each item offered a prompting sentence with a missing item or with a highlighted vocabulary item, followed by four vocabulary choices. The students were asked to infer the meaning

from the sentence context and then select the appropriate choice. The test also included 13 grammar items on different grammatical structures and parts of speech.

3. A listening test: This test is administered at the end of the term to all the participants. It is a task-based listening test consisting of 15 items. All the students taking the 3-credit English course should also sit for this exam, and their performance on this test will constitute 25 percent of their English achievement score. The test is administered in a language lab where the participants listen to a tape recorder and then respond to the tasks based on the requirements of the test.

Procedure

Initially, the participants took the Nelson test (Version 300) in 50 minutes. They were notified that the test was for research purposes and that they would gain a bonus if they took part in all phases of the research, and this would be considered as part of their academic record for the course.

Thereafter, the participants took the listening test at the end of semester. This task-based listening test measured the participants' listening proficiency. The participants who sat for Nelson test were required to take this exam too. The items on this test included both recognition and open-ended questions. The listening test items were selected from a package of listening test banks including conversations and episodes performed by native speakers. The participants took this test in a well-equipped language lab where they listened through headphones to this test and responded to the questions. Time allotment for this test was 20 minutes.

The final test administration was the final English achievement test which lasted for 70 minutes. This test was a written exam of 50 multiple-choice items of vocabulary, grammar, and reading as explained above. It was an achievement test administered at the end of the term.

The collected data from each test were scored. Each correct response was given a score of one, and incorrect responses were given a score of zero, with no penalty for the wrong responses. Then, the data were inserted into SPSS and Minitab software for analyses and comparisons.

Results and Discussion

Data Analysis for Question 1

Reliability. Reliability, a necessary condition for validity, is a function of variation. Variation can be systematic due to students' ability, or unsystematic due to other factors like speediness of a test (Bachman, 1990). As Bachman argues, the aim in language testing is to measure the systematic variation; therefore, the higher the systematic variation in test scores, the more reliable the test is.

Reliability of a test can be measured through different means. The Alpha (Cronbach's) models the internal consistency based on average correlation among items. It is the most common form of reliability coefficient. Alpha equals zero when the true score is not measured at all, and there is only an error component. Alpha equals 1 when all items measure only the true score and there is no error component. More importantly, because reliability is influenced by the group tested, the test content, and testing conditions, a single method for estimating reliability is not recommended (Ebel & Frisbie, 1986). As such, reliability was estimated through Alpha, split-half, and KR-21 (Table 1 below). The Overall Achievement score in this table refers to the overall score on the EAT plus the listening score.

Table1. *Reliability Co-efficient of the Test*

	No. of items	Reliability (Alpha)	Reliability (Split-half)	Reliability ((KR-21)
Nelson	50	.87	.86	.87
English Achievement Test (EAT)	50	.79	.78	.79
Reading	22	.84	.84	.82
Vocabulary	15	.84	.83	.84
Grammar	13	.85	.85	.85
Listening	15	.87	.87	.87
Overall Achievement	65	.79	.77	.79

The highest reliability level was found for the listening section. Guerrero (2000) and Davis (1990) believe a test should have a consistency of at least .90. Though rather close to that level, none of the subtests of the IUSTEAT reached that reliability level. One probable reason is the ease or difficulty of the items. In norm-referenced testing, too easy or too difficult items result in a restricted range of scores or very little variance (Bachman, 1990). That is, the greater the size of the score variance, the more reliable the test is. Item analysis results (see data for Question 3) show that a sizeable proportion of the items (68%) are very easy or relatively easy. Thus, the resultant decreased variance is most likely the prime cause of the less than acceptable or ideal reliability estimates.

The correlation of the EAT with the Nelson test was shown to be moderate (r =.50). Low to moderate correlations were also found between the subsections of the EAT and the Nelson test (i.e. .32, .42, .36, .46 for listening, reading, vocabulary, and grammar, respectively).

Data Analysis for Question 2

To examine the degree of contribution of the predictor variables (i.e. listening, reading, vocabulary, and grammar) to the dependent variable (the students' overall achievement), a multiple regression analysis was run. In this type of regression, the naturally occurring scores on a number of predictor variables are measured to see which set of observed variables leads to the best prediction of the dependent variable (Brace, Kemp, & Snelger, 2000).

Using the "enter method", a significant model (F4, 150= 4336.23, p<.005, Adjusted R square =.99) was found. All the predictor variables were found to be significant as shown below:

Table 2. *Regression Analysis on the Variables of the Study*

Predictor Variable	**Beta**	**P**
Reading	.377	p<.0005
Vocabulary	.363	p<.0005
Grammar	.300	p<.0005
Listening	.281	p<.0005

Reading had the highest contribution and listening the lowest. Vocabulary rated the second most important predictor of overall foreign language achievement. A similar regression analysis was conducted with the same predictor variables but with language proficiency (based on the performance on the Nelson test) as the dependent variable. The results, however, were slightly different; that is, vocabulary was not a significant predictor in this analysis (F4, 150= 1056.75, p<.0005, Adjusted R square=.96).

Table 3. *Regression Analysis with Language Proficiency as Dependent Variable*

Predictor Variable	**Beta**	**P**
Reading	.332	p<.0005
Vocabulary	.363	p=.828
Grammar	.015	p= .006
Listening	.275	p<.0005

One reason for the significant predictor role of vocabulary in the achievement score in the first regression analysis may be the considerable attention paid in language curricula to vocabulary (like guessing unfamiliar word, predicting, matching words, etc.) than to other skills, like grammar or listening. This insistence on vocabulary, as Barnet (1986) argues, is justified. It may however be an overreaction to the former stress on grammar, and, if so, further research needs to prove that. The insignificant predictor role of vocabulary in language proficiency in the latter regression analysis is supported by Alavi (2001). He discovered that, contrary to the common assumption, vocabulary was the least related variable to the TOEFL score of Iranian test takers. These findings, of course, do not undermine the important role of vocabulary in improving performance on other skills like listening comprehension (Vandergrift, 2006).

Data Analysis for Question 3

Item facility/difficulty and item discrimination.. One of the main purposes of this study was to examine the difficulty and discrimination value of the test items. To this end, item facility, item difficulty, and item discrimination indices of the test items were analyzed through classical item analysis. Therefore, performance of two misfit statistics was examined: P-value (item difficulty), and Biserial correlation (item discrimination).The justification for examining the ease or difficulty of an item was that a test item that is too easy or too difficult can tell us very little about the cognitive and achievement level of the participants and thus must be removed from a test (Farhady, et al., 1995; Fulcher, 1997). Item difficulty index can range from zero to 1. Table 4 summarizes the range of difficulty for an item and what it actually means.

Table 4. *Item Difficulty Indices*

Index	Difficulty level
P Low < .37	Difficult test
P moderate .37 - .63	Suitable/moderately difficult
P high >.63	Easy test

Items having p-values less than .37 or greater than .63 are considered to be misfitting for analysis (Reynolds, Perkins, & Brutten, 1994). In norm-referenced testing, deciding on how difficult an item should be is mainly a function of the purpose of the test and is not necessarily absolute. Item discrimination shows how well a test item discriminates between less knowledgeable and more knowledgeable examinees in the ability being tested. An item with a too high or too low facility index is not likely to have a discrimination power (Osterlind, 1998). The discrimination index (Pbis) can range from -1 to +1. A discrimination index of zero shows an item has no power to differentiate between the strong and weak students. Therefore, the closer the Pbis to +1, the more powerful an item is in distinguishing strong and weak students. Nonetheless, an item with a Pbis value between .30-1 can be considered an appropriate item. Items with

Pbis-correlation of less than .30 are considered to be misfitting (Fulcher, 1997). An index below .30 is considered a low Pbis (Heward, 1994). A low Pbis index can be attributed to either test preparation problems or to the lack of correlation between the items in a test. An item with a minus Pbis value shows that even the weakest students could answer that item (Fulcher, 1997).

The findings of difficulty and discrimination analyses are summarized in the appendix. It should be pointed out that these indices were not examined for the listening subsection, because it was a ready-made test the items of which were not developed by the English Department staff. Thus, the final achievement test which had 50 items was targeted for item analysis. The results showed that 22% of the items in the final achievement exam were acceptable items with appropriate levels of difficulty. On the other hand, 68% of the items were easy or relatively easy, while 10% of the items were very difficult. Taking the above norm-referenced indices of difficulty, it can be claimed that the IUSTEAT is a relatively easy test for the population under investigation.

Item discrimination analysis showed that 26% of the items very well discriminate between high and low level participants, while 72% of the items had low discrimination power, and 2% had no discrimination power or zero discrimination. As can be seen, item 1, which is a display reading question checking knowledge of specific information in the text, had zero difficulty indexes and thus zero discrimination power as well. No malfunctioning items with negative discrimination power were identified. Accordingly, it can be concluded that the majority of the items were not suitable functioning items, i.e., they had low discrimination value. The results are somewhat similar to Kiany and Haghighi’s (2005) study where only 13% of the items of a high-stakes test like TOEFL had discrimination values above .50.

Data Analysis for Question 4

The performance of male and female engineering students was compared in terms of their performance on the Nelson test and on the different sections of the EAT.

Table 5. *Independent Samples t-test Comparing Males and Females*

	F	d.f.	Mean difference	Sig.
Nelson	.282	152	-1.71	.112
EAT	.065	152	-.30	.796
Reading	.335	152	.12	.813
Vocabulary	1.821	152	-.12	.805
Grammar	.041	152	-.13	.729
Listening	.173	152	-.22	.563
Overall Achievement	.043	152	-.52	.688

No significant differences were found between males and females in their language proficiency, their achievement score, and in the different subsections of the test. The results are consistent with previous findings on gender role in language proficiency (Ryan & Bachman, 1992; Vandergrift, 2006; Wu & Lin, 2003). Nevertheless, it should be pointed out that the two groups were compared only in terms of receptive skills, and gender differences in productive skills like speaking and writing should be investigated (Bermúdez & Prater, 1994).

Conclusion

This study examined the characteristics of the IUST English achievement test. The reliability estimates showed that the listening section has the highest degree of reliability than the other parts. This implies a higher variance for the subjects' scores in this section. Nonetheless, the reliability estimates of the test and its subsections were only relatively acceptable. Given that this it is a university-wide test which all the engineering students of the university must take, it is necessary for test designers and the IUST English Department to be more concerned about the high quality of the test. Based on the researcher's knowledge of the English departments across Iran, almost all departments follow more or less the same testing procedure used in the IUST English Department. Therefore, it is vital that the test developers in these departments be cautious about the estimate of error introduced into their testing procedure and into the validity of their interpretation. Triangulation and calibration measures in this regard are very helpful to increase test reliability and fairness. Such a triangulation effort can help increase reliability by reducing systematic error, and it helps test designers employ multiple methods of measurement. Examination of data from the alternative methods gives insight into how individual scores may be adjusted to come closer to reflecting true scores, thus increasing reliability (Tenko, 1998). Pre-testing, item banking, and trialing can be very helpful in reducing unsystematic variation. Pre-testing even on a smaller scale and holding some calibration meetings to discuss some particular items can significantly improve the quality and fairness of a test like the IUSTEAT.

Item difficulty and discrimination measures in this study can help language instructors realize the important role of checking the items for their difficulty or discrimination index. This process may not be possible to follow on every occasion for norm-referenced tests, however. Nonetheless, checking for acceptable levels of difficulty and discrimination for high-stakes local tests developed by some universities for selection and entry purposes is essential. These tests are used to make decisions about the future educational career of many graduate students. This implies that test designers and more importantly policy and decision makers need to be more careful about the quality and usefulness of test items and the possible faults with them so that they can make more intelligent, informed, and fair decisions. Furthermore, a point to bear in mind is that in deciding about whether to include or exclude an item, the theoretical assumptions and purposes behind the interpretation of the scores must be taken into account (Farhady et al., 1995). For example, if a test item has high item facility, this may mean that the instruction has been successful and thus need not be excluded.

That reading comprehension was found to be a significant predictor of performance on both the language proficiency test and the achievement test lends further support to the fundamental role of reading comprehension in Iranian students' performance on English tests (Alavi, 2001; Nassaji, 2003). The implication is that language instructors and curriculum developers need to develop a more integrated reading instruction which focuses on developing the knowledge of large recognition vocabulary, of metacognitive awareness to become strategic readers, of practice in reading fluency to develop automaticity, and other techniques for a more comprehensive reading instruction. Taking all these steps into account is essential to move readers from 'learning to read' towards 'reading to learn' in English.

We did not find any differences between males and females on any of the tests and their subsections. However, it is possible that different results may come up comparing male and female students from other disciplines (e.g., humanities) and in productive skills like writing or speaking. Future research would need to examine the quality of achievement tests like the EAT in terms of their validity and fairness for the particular groups of students. Such a study should investigate fairness of both the test and the testing practice (Kunnan, 2004).

Further research would of course be necessary to examine which strategies learners use to deal with the different task types and test questions like the IUSTEAT and similar tests of this caliber. Moreover, examining whether proficiency in L1 can contribute to English language performance, and, if so, to what degree, is another significant area of research. The other important issue relates to the level of L2 language proficiency (threshold level) required to perform effectively in different language skills like reading, listening, etc.

References

Alavi, S. M. (2001). On the relationship between grammar knowledge and FCE-TOEFL reading comprehension tests. *Proceedings of the First IELTI Conference in Iran.* Faculty of foreign languages, University of Tehran.

Anderson, A. (1991). Individual differences in strategy use in second language reading and testing. *Modern language Journal*, *74*, 460-472.

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

Bae, J., & Bachman, L. (1998). A latent variable approach to listening and reading: Testing factorial invariance across two groups of children in the Korean/English two-way immersion program. *Language Testing*, *15,* 380-414.

Barnet, M. A. (1986). Syntactic and lexical/semantic skill in foreign language reading: Importance and interaction. *The Modern Language Journal*, *70*(4), 343-349.

Bensoussan, M., & Laufer, B. (1984). Lexical setting in context in EFL reading comprehension. *Journal of Research in Reading*, *7*, 15-32.

Bermúdez, A. B., & Prater, D. L. (1994). Examining the effects of gender and second language proficiency on Hispanic writers' persuasive discourse. *Bilingual Research Journal*, *18*, 47-62.

Brace, N., Kepm, R., & Snelger, R. (2000). *SPSS for psychologists*. New York: Palgrave.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. *Reading in a Foreign Language*, *15*, 1, 1-27.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Bugel, K., & 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., & Grabe, W. (2002). Reading. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 233-250). London: Oxford University Press.

Celce-Murcia, M., & Larsen- Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course*. Newbury House: Rowley M.A.

Davis, A. (1990). *Principles of language testing*. Oxford: Basil Blackwell Ltd.

De Carrico, J., & Larsen-Freeman, D. (2002). Grammar. In N. Schmitt (Ed.), *An introduction to applied linguistics* (pp. 19-35). London: Oxford University Press.

Ebel, R. L., & Friesbie, D. A. (1986). *Essentials of educational measurement*. N.J.: Prentice Hall.

Farhady, H., Jafarpour, A., & Birjandi, P. (1995). *Testing language skills: From theory to practice*. Tehran: The Centre for Publishing and Compiling University Books in Humanities.

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing*, *14*(2), 113-138.

Grabe, W., & Stoller, F. L. (2002). *Teaching and researching reading*. Harlow: Longman.

Graham, S. (2006). Listening comprehension: The learners' perspective. *System, 34*, 165–182

Guerrero, M. D. (2000). The unified validity of the four skills exam. *Language testing, 17*(4), 397-421.

Haynes, M. (1984). Patterns and perils in second language reading. In J. Ahndscombe, R. Orem, & B. Taylor (Eds.), *On TESOL '83: The question of control* (pp. 163-176). Washington DC: TESOL.

Heward, W. L. (1994). Three low-tech strategies for increasing the frequency of active student response during group instruction. In R. Gardner, D. Sainato, J. O. Cooper, T. Heron, W. L. Heward, J. Eshleman, & T. A. Grossi, (Eds.), *Behaviour analysis in education: Focus on measurable superior instruction* (pp. 283-320). Pacific Grove, CA: Brooks/Cole.

Kiany, G. R., & Haghighi, M. (2005). The investigation of TMU English language proficiency test: Reliability-related issues. *Journal of Humanities of Alzahra University, 16*(58), 55-73.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *Europe language testing in a global context: Selected papers from the ALTE conference in Barcelona* (pp. 27–48). Cambridge: Cambridge University Press.

Larsen-Freeman, D. (2002). *Grammar dimensions*. Boston, MA: Heinle and Heinle.

Laufer, B. (2006). Comparing focus on form and focus on forms in second-language vocabulary learning. *The Canadian Modern Language Review, 63*(1), 149-166.

Lynch, T., & Mendelsohn, D. (2002). Listening. In N. Schmitt (Ed.), *An introduction to applied linguistics*. London: Oxford University Press.

Nassaji, H. (2003). Higher-level and lower-level text processing skills in advanced ESL reading comprehension. *The Modern Language Journal, 87*(2), 261-267.

Osterlind, S. J. (1998). *Constructing test items*. Boston: Kluwer Academic Publishers.

Reynolds, T., Perkins, K., & Brutten, S. (1994). Comparative item analysis: Study of a language placement test. *Language Testing, 1*(1), 1-13.

Rost, M. (2002). *Teaching and researching listening*. England: Longman, Harlow.

Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of language proficiency. *Language Testing, 9*(1), 12-29.

Rubin, J. (1994). A review of second language listening comprehension research. *The Modern Language Journal, 78*(2), 199-221.

Tenko, R. (1998). Coefficient alpha and composite reliability with interrelated non-homogeneous items. *Applied Psychological Measurement*, *22*(4), 375-385.

Vandergrift, L. (2006). Second language listening: Listening ability or language proficiency? *The Modern Language Journal*, *90*(1), 6-18.

Wu, F., & Lin, J., (2003). Differential performance by gender in foreign language testing. *Poster for the 2003 Annual Meeting of NCME in Chicago.*

AUTHOR

Esmaeel Abdollahzadeh has a Ph.D. in applied linguistics and TEFL. He is an associate professor teaching applied linguistics, ESP, and advanced writing courses at the Department of Foreign Languages of Iran University of Science and Technology. He has presented and published nationally and internationally on issues in second language academic reading and writing, discourse and ESP, as well as language learning strategies.

Email Address: s_abdolah@iust.ac.ir

Appendix

Item facility/difficulty and discrimination indices

Item	Item facility	Item difficulty	Item discrimination
1	1	0	.0
2	.57	.42	.18
3	.90	.09	.13
4	.86	.13	.17
5	.79	.20	.17
6	.47	.52	.25
7	.91	.08	.14
8	.88	.11	.13
9	.92	.07	.08
10	.44	.55	.44
11	.97	.02	.05
12	.67	.32	.33
13	.10	.89	.04
14	.82	.17	.32
15	.89	.10	.08
16	.64	.35	.29
17	.67	.32	.36
18	.76	.23	.21
19	.65	.34	.28
20	.36	.63	.22
21	.74	.25	.25
22	.36	.63	.40
23	.63	.36	.41
24	.80	.19	.26
25	.92	.07	.11
26	.88	.11	.18

27	.88	.11	.10
28	.09	.90	.008
29	.60	.39	.32
30	.55	.44	.42
31	.82	.17	.21
32	.73	.26	.29
33	.92	.07	.09
34	.90	.09	.13
35	.79	.20	.23
36	.39	.60	.43
37	.83	.16	.21
38	.83	.16	.17
39	.91	.08	.09
40	.62	.37	.42
41	.76	.23	.08
42	.60	.39	.11
43	.94	.05	.10
44	.62	.37	.30
45	.88	.11	.01
46	.91	.08	.11
47	.79	.20	.26
48	.91	.08	.12
49	.70	.29	.30
50	.26	.73	.37