TEST TAKER CHARACTERISTICS AND PERFORMANCE ON A CLOZE TEST: AN INVESTIGATION USING DIF METHODOLOGY

by

LING HE

M.Ed. in Faculty of Education, Memorial University of Newfoundland, Canada, 2000

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF ARTS

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF EDUCATIONAL AND COUNSELLING PSYCHOLOGY, AND SPECIAL EDUCATION

With Specialization in

MEASUREMENT, EVALUATION, AND RESEARCH METHODOLOGY

We accept this thesis as confirming to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 2002

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational Counselling Psychology, and Special Education

The University of British Columbia
Vancouver, Canada

Date September 10, 2002

DE-6 (2/88)

ABSTRACT

This quantitative study aims to investigate the influences of test taker characteristics across two different broad language groups, Asian and Non-Asian, on cloze test performance. This study used differential item functioning (DIF) methodology. The specific purpose is to provide empirical evidence for factors that can bias items on cloze tests due to (a) examinees' familiarity with test format, and (b) their academic discipline. Data in this study are collected from 216 English as a second language (ESL) students (ages 17-56) from Carleton University. Each examinee completes two cloze tests in the same session. However, the focus of the current study is Test 1 whereas Test 2 is used in creating a composite score which would serve as the matching variable in the DIF analyses. The method used for DIF analysis is logistic regression approach with multiple matching variables and multiple grouping variables. Empirical results by logistic regression method showed that only familiarity with cloze test format was differently related to DIF. This result was also supported with some theoretical DIF hypotheses founded on prior knowledge pertaining to cognitive processes that could related to differential performance of test items.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I wish to express my gratitude and my indebtedness to my supervisor, Dr. Bruno Zumbo for generously giving me hundreds of hours of his precious time, for insightful guidance on language testing issues, for reviews on earlier drafts of my manuscript, and for enduring support from the beginning to the end of this thesis. I also want to acknowledge my committee member, Dr. Anita Hubley, who gave insightful suggestions throughout this thesis, especially in literature review part about multiple- choice. Finally, I would like to thank Dr. Janna D. Fox for being the external examiner and providing the data of this thesis.

As always, I am grateful to my parents for their assistance in many ways, too numerous to list here.

Introduction

Introduction to the Problem

Language use can hardly be optimistically seen as a fully uniform repertoire but depends on the context of use and the language user's characteristics (social and geographical origin, education, occupation, age, gender etc.). These factors are related to one group of characteristics that has been identified as test taker characteristics or background characteristics (Bachman, 1990). Test taker characteristics generally include personal characteristics (e.g., age, gender, culture), educational characteristics (e.g., previous instruction in English), and cognitive, psychological and social characteristics (e.g., learning strategies and styles, attitude and motivation). It has been realized that sometimes test-takers' incorrect responses may not indicate erroneous understanding but reflect perspectives not shared by them. The complexity of these uncontrolled factors differentially affects the test-taker's performance. The impact of test taker characteristics on their test performance has been shown by a casual survey of the measurement literature over the past two decades where it was clearly attested to the emergence of test bias in college administration scholastic aptitude scales as one of the most burning and debated issues in the field of testing and evaluation. Also, research on several of these characteristics from the perspectives of second language acquisition has shown that some of these influence language learning to differing degrees (Gardner, 1985, 1988).

It is clear that biased scores can systematically over- or underestimate the abilities of certain groups. Bleistein (1986) wrote, "It is widely recognized that test performance may be influenced by variables other than ability in the trait being measured and, to the extent that such influence occurs, the validity of the test may be compromised" (p. 4). In

this regard, the AERA/APA/ NCME *Standards* (1999) also remind us that the test needs to be equivalent for the examinees with different backgrounds.

> The term *equivalence* refers to the degree to which test scores can be used to make comparable inferences for different examinees... In general, linguistic and cultural characteristics of the intended examinee population should be reflected in examinee samples used throughout the processes of test design, validation, and norming. (pp. 92 - 93)

Given the potential test bias problems existing in the field of testing and evaluation, a recent concern among researchers in the field of language testing has been the identification and characterization of the individual characteristics that influence performance on tests of English as a foreign language (EFL) or English as a second language (ESL). An interesting context, cloze tests, is chosen in this study wherein we can investigate whether a test includes a potential bias against some particular groups of test takers' characteristics. In order to investigate whether there is equivalence for different groups formed by background variables, one needs to match on the variable of interest (i.e., the construct being measured) and then see if there is a group difference. This translates to the use of differential item functioning (DIF) methodology to investigate the equivalence.

To this end, the following section will briefly review differential item functioning, and the logistic regression approach for detection of DIF. In the next section, I will review the research in the areas of the application of cloze tests in Asian (i.e., China, Japan, and Korea) and non-Asian countries, and previous studies about language

background variables. This review aims to find the "gap" in the previous studies so that the rationale and implications to my study are provided. Then I will state the theoretical framework that has helped me frame my study. After discussing the purpose of the study, a research question and design for my study will be presented. In this part, I will describe areas such as (a) data collection, discussing the issues like subjects and instrument (i.e., the specific cloze test used in this study), and (b) analytical procedures applied to the data, using logistic regression approach, that is, how the data are formed for the DIF analysis. The implications for research and practice of the current study are considered.

In short, as an overview, this thesis is a study of the differential item functioning of cloze tests based on (a) country of origin and (b) faculty of study or academic discipline. An additional contribution is the introduction and application of an extended DIF model that includes multiple matching variables, a measure of effect size, and a group contrast that is more complex than typically found in DIF analyses. It should be noted that "country of origin" as grouping variable represents a variety of factors and social and educational forces. In this study, the country of origin will be primarily interpreted to represent differences in (a) degree of familiarity with the testing format, and (b) educational and curricular differences. Like most DIF studies it will be difficult to disentangle these various factors of "country of origin".

Differential Item Functioning

The DIF procedure is a statistical technique to identify the differential item response pattern across different groups. The DIF procedure assumes that if test takers have approximately the same knowledge (as measured by total test scores), then they

should perform in statistically similar ways on individual test questions or tasks regardless of, for example, their sex, race or ethnicity. In other words, the DIF procedure matches the test-takers on the knowledge and skills included in the test, and identifies differential item functioning, which may unfairly influence examinees' scores across different language groups. In educational and psychological fields, a variety of methods have been developed for detecting DIF. The most recent DIF research by Zumbo and Hubley (in press) has presented valuable insights to the current and future directions and focus of DIF research, which they summarize into two broad classes of DIF detection methods: Mantel-Haenszel and logistic regression approaches. Zumbo and Hubley write,

> "Today, in addition to matters of bias, DIF technology is used to help answer a variety of basic research and applied measurement questions wherein one wants to compare item performance between or among groups when taking into account the ability distribution. At this point, applications of DIF have more in common with the uses of [analysis of covariance] (ANCOVA) or [attribute-by-treatment interaction](ATI) than test bias per se." (p. 10)

Logistic regression approach. Zumbo and Hubley (in press) pointed out that the logistic regress approach allows conducting a regression analysis for each item wherein each tests the statistical effect of the grouping variable(s) and interaction of the grouping variable and the total score after controlling on the total scores. Considering the purpose of the current study is to investigate DIF across two different ethic groups, logistic regression approach will be used because it allows for not only uniform DIF (i.e., the

main effect of groups), but also nonuniform DIF (i.e., the interaction of group by ability) while Mental-Haenszel assumes no interaction.

Review of the Literature

The History and Application of Cloze Tests

In this section, I will review the history, properties, and development of cloze tests. The reader should note that cloze tests are widely used in Asian countries such as China, Korea, and Japan whereas cloze tests are used less frequently in Canada or the U.S. A. By this review, I hope the reader has an understanding of this history so as to better understand the current study. The other important part reviewed in this section is about previous studies on the effect of language background of test-takers on test performances. Both test characteristics and test taker characteristics are discussed in relation to construct validation.

The history of cloze tests. The cloze procedure as a method is a "fill-in-the-blanks" activity which involves deleting every nth word from a passage. It was invented by Wilson Taylor in 1953 first as means of assessing both the readability of a variety of written materials and the reading ability of a variety of subjects for native English speakers. The passages in Taylor's first experiment were taken from Flesch's book *How to Test Reliability* (1951). In the process of this work, Taylor compared various methodological procedures. He found that deletion systems did not affect the ranking of a passage but a greater number of blanks (i.e., 35 rather than 16) tended to discriminate among subjects more efficiently. This result was confirmed in his second experiment.

Taylor's work laid the foundation for a great deal of research. He, himself, extended the study in his doctoral dissertation (Taylor, 1954).

Cloze tests are considered to be derived from the Gestalt psychology notion and the notion of redundancy provided by information theory. Gestalt theory is a broadly interdisciplinary general theory which provides a framework for a wide variety of psychological phenomena, processes, and applications; Human beings are viewed as open systems in active interaction with their environment. Similarly, the notion of redundancy states human ability to restore the missing words to reconstruct textual coherence in the incomplete pattern constructed by deleting words from a running text. Both notions indicate the human ability to complete an incomplete pattern (closure).

Later, the cloze procedure was applied to testing English as a second language (ESL) and/or English as a foreign language (EFL) learners' reading proficiency (for an overview, see Oller, 1979 and Hinofotis, 1987). Now the cloze procedure has become one of the accepted integrative means of assessing the learner's overall language proficiency (Bachman, 1982; Breland, 1970; Diaz, 1983; Brown, 1983; Hale, Stansfield, Rock, Hicks, Bulter & Oller, 1989; Olmedo, 1977, Padilla, 1979), especially widely adopted as a measure of overall second and /or foreign language proficiency (Alderson, 1978a; Jones, 1977). Alderson (1978) described how:

> "The last decade, in particular, has seen a growing use of the cloze procedure with non-native speakers of English to measure not only their reading comprehension abilities but also their general linguistic proficiency in English as a Foreign language." (p. 2)

Studies (e.g., Alderson, 1979, 1980; Bachman, 1982, 1985; Brown, 1983) on the cloze test have offered differing evidence regarding pros and cons of cloze testing. Several researchers (Bachman, 1982; Brown, 1983; Chavez-Oller, Chihara, Weaver, & Oller, 1985; Jonz, 1990) have argued that cloze tests are a stable, reliable, and sensitive measure of the inter-sentential components of language. Brown (1992) states that the cloze items assess a wide range of language points from morphemic and clausal level grammar rules to discourse and pragmatic rules of cohesion and coherence. Researchers generally agree that cloze tests combine the advantages of integrative testing, which measures overall language proficiency, and objective scoring, which tests the discrete elements of the language (e.g., grammar, vocabulary or pronunciation). Also, a number of researchers have studied reliability, validity, mean item facility and discrimination, and usability of various types of scoring methods (Bachman, 1985; Brown, 1980, 1983, 1984, 1988, 1992; Chapelle & Abraham, 1990; Darnel, 1970; Feldmann & Stemmer, 1987; Klein-Braley, 1985; Markham, 1985; Shin, 1983, 1990). Despite the extensive research carried out on the cloze, there are striking disagreements on the issue of what cloze is testing. There also exists a theoretical problem --- that of construct validity affected by students' major field, text content, and some other related variables like cognitive sex difference.

The essential property of the cloze procedure is its deletion frequency, the subjects then being required to replace those words. Two common ways the cloze procedure used in cloze tests are fixed-ration deletion, that is "the systematic deletion of words from text" (e.g., a deletion of every fifth word from text), and a rational random deletion of words from text. The more words a subject replaces exactly, the greater his or

her reading ability. Researchers (e.g. Alderson, 1978) showed that changes in deletion frequency sometimes results in significant differences between tests. In this regard, the ways cloze procedures (i.e., fixed-ratio or rational-deletion) used in cloze tests impact on their equivalence as measures of the test-takers' language proficiency. This also indicates another important property of cloze tests: Cloze items are contextually interdependent on one another. In addition, systematic differences in the materials that are chosen for the cloze tests, and the differences in material itself may have a substantial impact on whether the results of two cloze tests are equally valid indicators of language proficiency of the test-takers. These properties indicate that prior skill or knowledge, and background variables may impact on examinees' test performance.

The application of cloze tests. According to Chinese test history, the cloze procedure was used in Chinese public service examinations as early as the Qing dynasty (1644 -1911 A. D.) when language arts as one of the testing subjects was given in the government official exam. Even today cloze tests are still used in many national standardized tests like the College English Test (CET, Band 4 and Band 6) and Waiyu Shuiping Kaoshi (WSK), both of which are national tests to measure examinees' language proficiency. Cloze tests are also regularly used as classroom exercises and quizzes when students are learning English or their native language.

Literature shows that the fact that cloze tests are not so popularly used in Western countries as in China may result from the communicative movement in the late 1970s and early 1980s. The approach of "communicative" language testing stands or falls by the degree of real life or authenticity; that is, naturally occurring texts ("used language", to use Brazil's term, 1995) should be used in language teaching and testing. Communicative

approach maintains that in real-life situations the learner will meet authentic (non-simplified and non-made-up) texts and will have to solve authentic tasks by using authentic language.

Since the early 1980s, therefore, the focus of foreign language instruction has moved away from the mastery of discrete language skills, such as grammar, vocabulary, and pronunciation, to the development of communicative proficiency--that is, the ability to communicate about real-world topics with native speakers of the target language. Widely termed the "proficiency movement," this change has developed in tandem with changes in how students' foreign language skills are assessed.

It is considered that the traditional assessment tools of earlier decades--usually discrete-point tests that focused on individual skills, such as knowledge of vocabulary and grammatical accuracy--evaluated students' knowledge *about* the language, not what they could *do with* the language. Although discrete-point tests are still used in many circumstances, particularly for large-scale standardized assessments, many of the newer assessment measures and techniques are performance based; that is, they require students to demonstrate knowledge and skills by carrying out challenging tasks. This enables teachers to measure what the students can actually do in various communicative contexts using the target language.

The key targets to attack for the new "communicative" language testers were the multiple-choice item as embodied in the Test of English as a Foreign Language (TOEFL; Spolsky, 1995). There is a considerable doubt about the validity of the multiple-choice format as measure of language ability. Answering multiple-choice items is considered an unreal task, as in real life one is rarely presented with four or three alternatives from

which to make a choice to signal understanding. Normally, when required, an understanding of what has been read or heard can be communicated through speech or writing.

Yet cloze tests are not necessarily communicative although they are authentic. These might be part of the reasons to have led to cloze tests, which is made of multiple choices in its response, less popular in Western countries. However, multiple choice test continues to serve the needs of education, business, and government, probably because schools are not the only consumers of multiple choice and there has been no large movement within the various credentialing systems of business and government to replace their present multiple-choice tests. Therefore it is reasonable to argue that students in the Western countries are familiar with multiple-choice format.

What is worth noting here is that cloze and multiple choice tests have different validity in measuring language ability although they have similar superficial features in terms of multiple choice formats in both tests. In comparing cloze and multiple-choice tests, Engineer (1977) concluded that the two techniques are measuring different aspects of reading activities; namely, a timed cloze test measures the *process* of reading (i.e., the reader's ability to understand the text while he or she is actually reading it) whereas multiple-choice tests, on the other hand, measure the *product* of reading (i.e., the reader's ability to interpret the abstracted information for its meaning value). Thereby the Western students' familiarity with multiple-choice tests does not mean that they are good at cloze tests.

Language Classrooms in Asian and Non-Asian Countries

Although our understanding of language proficiency has been considerably broadened in the past few years by the notion of communicative competence, which has always laid great stress on authenticity, Curriculum with high emphasis on grammar-translation methods in learning English still dominates language classrooms in China. This has found its support. For example, Chu (1990) claims that educators should not be so ready to dismiss traditional grammar-translation methods, but be willing to modify their approaches. He maintains that grammar & translation may be made more effective in language instruction through the incorporation of semantics & discourse. Cortazzi and Jin's study (1996) has provided a clear picture in that area wherein they state that the Chinese culture of learning English generally has four main foci of attention: (a) teacher, (b) textbook, (c) grammar, and (d) vocabulary-centeredness; in contrast, the Western language learning culture is characterized as learner and problem-centered, with focus on functions, uses, and interaction. The higher requirements of authentic tasks, that is, "the simulation of real-life texts or tasks and the interaction between the characteristics of such texts and tasks and the language ability of the test takers" (Douglas, 2000, p. 90) makes grammar-translation method more practical and efficient in testing in Asian countries such as China, Japan and Korea wherein testing is important at all levels of the school system, in terms of high population and fierce competition.

Changes in foreign language assessment in recent years (i.e., from discrete-point tests to integrative tests) can be divided into two main categories based on their catalysts. National assessment initiatives have widely influenced classroom instruction in a "top-

down" approach; local assessment initiatives, which have appeared in response to curricular and instructional changes, may be seen as "bottom-up" initiatives.

Such primary literature has encouraged me to hypothesize that Asian students from China, Japan, and Korea are more familiar with test cloze test format than non-Asian students; the fact that Asian students special training in grammar are in favour of their answers to cloze tests because cloze test score has been found to be the best predictor of the number of grammar and word-choice errors for the L2 students (Hu and Hsian, 2000). I hypothesize, therefore, that Asian students may show differential item functioning as compared to non-Asian students on cloze tests. In essence, this hypothesis is that test format familiarity (as indicated by Asian vs. non-Asian) will be a relevant background variables effect test performance on the cloze.

Previous Studies on the Effect of Language Background Variables

Previous studies on the effect of language background variables have been mainly concerned two areas: One is related test-characteristics in scale level with focuses on construct validity (Ackerman et al., 2000; Brown, 1999; Ginther and Stevens, 1998; Hale et al., 1989; Kunnan, 1994; Oltman et al., 1988; Swinton and Powers, 1980). That is, whether a test measures the same constructs for various language groups. The other area is test-takers' characteristics in terms of examinees' gender, test-taking skills or test wiseness, inherent individual differences, interest, and educational fields (Chen and Henning, 1985; Curley and Schmitt, 1993; Fox, Pychyl, and Zumbo, 1997; O'Neil and Mcpeek, 1993; Ryan and Bachman, 1992; Sasaki, 1991; Stricker, 1981). In other words, whether the difference in test performance is impacted on by examinees' personal

difference variables, which reflects a complex interaction of biological, psychological, and social factors.

As stated in the very beginning in this thesis, test taker characteristics generally involve examinee's personal attributes, qualitative topic knowledge, and affective schemata (Bachman and Palmer, 1996). In contrast, test characteristics are usually related to the surface features or content characteristics of the questions (Osterlind, 1985). For example, whether the construct of the test is consistent or not has directly affected the validity of tests. Test characteristics and test taker characteristics are highly interrelated because to judge whether the construct measured by the test is consistent or not (i.e., test-characteristics) depends on the language groups (i.e. test taker characteristics) in the test. Therefore, there is a value to review these two areas in the following.

Test characteristics and construct validation. A number of studies on the construct validity in language testing were based on the sample of the Test of English as a Foreign Language (TOEFL) by using correlation and factor analysis. Both Swinton and Power (1980) and Kunnan (1994) have separately found bias items in terms of inconsistent constructs across Indo-European (IE) and non-Indo-European (NIE) language groups on the TOEFL. One finding was that the subsection of "Reading Comprehension and Vocabulary" tapped different concepts or behavior in two groups because of the examinees' inherent situations (i.e., the similarities and differences of their native languages and the target language). That is, the test conveyed one dimension for the NIE group but different dimensions for the IE groups whose languages are obviously different from English. This indicates that some DIF in the test could be caused by improper score

interpretations to the two language groups whose native languages are different from each other and English.

However, other studies (Olman et al., 1988; Hale *et al.*, 1989; Brown, 1999; Ackerman et al., 2000) have shown that one single dimension was generally present for the different language groups regardless of the language background. Olman et al (1988) pointed out that it was the proficiency level that led to structural relationships among test construct components in the TOEFL. Hale et. al.'s (1989) study confirmed these results across four groups: Semitic, Sino-Tibetan, Altaic and Indo-European languages, and more recently, Ackerman et al (2000) also showed this result across three groups: Arabic, French and Korean on the TOEFL as did Brown (1999) across 10 different language groups.

The above inconsistent results of various studies on the impact of language background indicates that DIF analysis should not stop at the scale level but should go further to investigate how DIF items affect the total test scores based on the possible item composites (Bolt and Stout, 1996). That is what are the influences of the test-taker characteristics on test performance across language groups?

Test taker characteristics and construct validation. Language background variables relating to test taker characteristics have been mainly concerned two areas in the literature: (a) achievement or knowledge tests, and (b) language proficiency tests.

Achievement tests. DIF on ability and achievements tests used in large-scale testing programs, such as Graduate Record Examination (GRE) General Test, and Scholastic Assessment Test (SAT), has been extensively examined in the last two decades. Although a great deal of inconsistency exists in the findings, some similarities in

the content of verbal and quantitative items displaying DIF have been identified in review of research on this topic (Curley & Schmitt, 1993; O'Neill & Mcpeek, 1993; Stricker, 1981). With regard to verbal items, for example, the O'Neill and Mcpeek (1993) review reported that items with science content favored males whereas those with social science, human relationships, aesthetics/ philosophy, or humanities content favored females; items with minority content favored minority examinees whereas those with homographs (i.e., words that are spelled and pronounced alike but have multiple meanings) favored majority examinees.

Language proficiency tests. Only a few studies (Chen & Henning, 1985; Ryan & Bachman, 1992; Sasaki, 1991) have explored DIF on language proficiency tests used in large-scale testing programs, such as the TOEFL and the English as a Second Language Placement Examination (ESLPE). Form research (Ercikson & Molloy, 1983; Peretz, 1986) has investigated DIF for faculty of study (i.e., the program examinees are studying). With regards to item difficulty, for example, Chen and Henning's (1985) study was the first one to identify DIF items across different native language backgrounds in second/foreign language test. It was explained in the study that the flagged DIF item resulted from the cognate words, meaning that similarity between the native language and the target language lexicon influenced test performance.

Summary of the Literature Review

As mentioned above, various ways of identifying DIF items in language tests are available. However, the underlying substantive reasons for DIF are still largely unknown (Roussos & Stout, 1996). The most common and most widely discussed explanation is

examinees' familiarity with the content of the items, variously referred to as exposure, experience, or cultural loading. What is unclear, however, is the role, if any, of familiarity of test formats, on language test performance. Also, from the perspective of language testing, the influence of these test-taker characteristics has not been given sufficient attention. As shown above, there are already several papers reporting the results of "faculty of study" as a variable that affects language test performance. I will also investigate this topic focusing on the cloze tests. That is, what role does the faculty in which a student is studying in have on cloze test performance. This is particularly important because language tests are often used to ascertain language proficiency in English for academic purposes. The above literature review has made it clear that the investigation of DIF is crucial in language proficiency tests in which test takers with diverse backgrounds are involved, because DIF items pose a considerable threat to the validity of tests.

Research Question

This quantitative study investigated the influences of test taker characteristics across two different broad groups (a) country of origin, Asian and non-Asian, and (b) faculty of study on cloze test performance by using differential item functioning (DIF) analysis. The above literature review has motivated me to interpret the sources of "country of origin in terms of test taker characteristics such as (a) examinees' familiarity with test format, and (b) educational and curricular differences. The grouping variable which represents the examinees' faculty of study or academic discipline is clearer to interpret because it represents scholastic and professional orientation and its impact of the

student's schema for language use. Specifically, I wish to determine what empirical evidence exists in this study to clarify the following question:

> Does item DIF attributable to background variables exist in a cloze test? In particular, do cloze test items display DIF based on the test-takers' characteristics such as country of origin or the choice of field of study?

Methodology

I start this section by discussing the following aspects: (a) data collection, discussing the issues like participants and instrument; that is, the specific cloze test used in this study, and (b) analytical procedures to the data, using logistic regression approach.

Participants

Cloze test data are provided by Dr. Janna Fox of Carleton University. The subjects in this study were 215 ESL graduate and undergraduate students who took the Canadian Academic English Language (CAEL) Assessment at Carleton University between January and August 2000 (see Table 1 for a description of the countries of origin for these examinees). An average age of these examinees was 25.7 years, male (N=116, Mean age = 25.8), and female (N= 99, Mean age =25.6). Examinees' information was provided on country of origin and faculty of study. Their statuses in Canada were student visa (48.6%), permanent resident (40.7%), Canadian citizens (5.1%), and others (5.6%). The length of their stay in Canada ranged from 0.03 to 128 months. They came from four faculties of study: Engineering (24.5%), Faculty of Arts/Social Science (FASS, 7.6%),

Public Affairs and Management (PAM, 30.1%), and Science/Computer Science (25.5%). All of them had took the Canadian Academic English Language Assessment (CAEL).

In our sample slightly over two thirds of the examinees were from Asia, so Asian examinees were grouped together and contrasted to the non-Asian group. The Asian group consisted of 112 Mandarin native speakers from Mainland China and three from Taiwan, three from Korea, and seven Japanese native speakers from Japan. The non-Asian group comprised 93 native speakers of 35 languages. The non-Asian group is admittedly rather heterogeneous – see Table 1.

The results are based on 215 examinees, among whom are 116 males and 99 females ranging in age from17 to 56 years.

Instruments

A Multiple-choice, Rational Cloze Test was used for this study. The rational deletion was based on the guideline stated by Fox (2000). As Fox describes, the guideline for deciding which words or phrases would be deleted from the passage embodied the rationale from Bachman (1985) and Bensoussan and Ramraz (1984): (a) micro-level deletions which focus on lexical choices of words and "their interaction with other words" (p. 231), (b) pragmatic-level deletions which focus on "extra-textual" or "general knowledge" (p. 231), and (c) macro-level deletions which focus on "the function of sentences and the structure of the text as a whole" (p. 231). These three levels of deletion were considered by Fox in developing the current rational cloze test. Figure 1 and 2 list the two cloze tests used in their study. Note that Cloze 1 is the test that is the focus of study whereas Cloze 2 was used as an additional matching variable for the DIF analysis.

Two cloze tests used in the study at Carlton are supposed to include 24 blanks in Test One, *The Kitchen and Beyond* (see Figure 1), the yellow version, and 22 blanks in Test Two, *Future Watch: The Internet* (see Figure 2), the white version, with the several sentences intact at the beginning and end of the passage to provide context (Hinofotis, 1987). The acceptable- word scoring method (see Brown, 1980), in which only the word given in the original text is considered correct, will be used.

Figure 1

The Kitchen and Beyond

By Allison Gore

There are amazing changes taking place in the applications of Internet technology. While most of us are familiar with new phrases such as e-commerce, a term applied to Internet shopping and buying, there's much more going on with the Internet than you may ever have imagined.

For example, a new web browser named "Leonardo" allows a person with a cell phone to send a message 1 _____ the Internet, to start the dishwasher, change

1. A. for B. at C. through D. in

a program on the washing 2_____, make the fridge

2. A. dryer B. machine C. time D. temperature

colder, or have the oven 3 _____ the dinner more

3. A. in B. cook C. to D. hot

slowly 4 _____ the dinner are caught in traffic.

4. A. in B. so C. because D. and

While such ideas sound far fetched, Mr. Xiao, Ormer CEO of Omnitel SpA, said the technology 5 _____ already available. Although much of the focus on today's Internet is on 6_____ fast it can send files and the bandwidth required to transmit

5. A. is B. he C. would D. had

6. A. how B. so C. the D. its

video, "the ability 7 _____ always be connected and be connected everywhere is

7. A. it B. to C. that D. will

8_____ as important as speed."

8. A. so B. the C. much D. just

Professor Nielson believes that the global costs of using the Internet will dramatically fall over the next few years, based 9_____

9. A. in B. on C. to D. at

Either a flat monthly 10 _____ for a connection

10. A cost B. or C. pays D. make

11_____ on volume, a model currently applied to electricity and water consumption.

11. A. about B. put C. based D. turning

See **INTERNET** on page 2

INTERNET *continued from page 1*

He said Internet traffic increases annually 12 _____ a factor of 10 --- a faster growth. rate than television, the VCR and cell phones. And the Internet 13 _____ expected to have one billion people online via computers, personal digital assistants, phones, kiosks, 14 _____ other information appliances early in the next millennium.

By composition, recent figures 15 _____ by the International Internet survey firm, Nua Ltd. Showed slightly more than 200 million 16 _____ are now 17 _____ to the Net.

On the technology front, according to Professor Nielson, viedo-conferencing will soon be as 18 _____ as sending a fax.

And 19 _____ will be satellite and wireless links everywhere to facilitate connections to the Internet, he 20 _____ .

Still, in terms of 21 _____ own evolution, "the Internet has undergone less than three per cent 22 _____ the technological development required to make it truly as 23 _____ as electricity or as multi-functional as it can be," Professor Nileson said. "Look at the speed of technological development over the past ten years. Think of what's possible --- and probable --- for the next ten. All I'm saying is you haven't seen anything yet".

Professor Nielson has his eyes squarely on the future. That future 24 _____ filled with possibilities for the Internet. If scientists like Professor Nielson have their way, almost every aspect of future life will be enhanced by use of the Internet.

With files from Jaune Y. Ellow
of the Bloomsbergy Press

12. A. by B. in C. as D. of

13. A. has B. will C. should D. is

14. A. many B. and C. some D. with

15. A. invested B. invented C. contrasted D. compiled

16. A. you B. information C. people D. factors

17. A. use B. getting C. prepared D. connected

18. A. powerful B. same C. soon D. common

19. A. connection B. there C. it D. that

20. A. add B. reads C. said D. be

21. A. our B. the C. its D. computer

22. A. of B. and C. quickly D. by

23. A. rare B. much C. common D. many

24. A. to B. is C. will D. can

Figure 2

FUTURE WATCH: THE INTERNET WILL SOON BE AS COMMON AS ELECTRICITY

By Allison Gore

The Internet will soon become a part of our daily lives, according to research scientists at Carleton University in Ottawa, Canada. Imagine pushing a few buttons on your in-car computer on the way home form work. With one push you tell the oven 1 _____ start cooking the roast.

1. A. on B. to C. when D. how

Another push and the computer tells you which roads to avoid because of heavy traffic or how the 2 _____ are doing in their playground. Yet another

2. A. others B. they C. kids D. players

3. _____ and the lights in your house go on, your garage door opens and the fridge prepares

3. A. push B. computer C. electricity D. thing

4. _____ favourite drink.

4. A. for B. with C. his D. your

Does this sound like science fiction? Not so, according to Professor Mark Nielson, Chair of the Innovation-Internet Research Group (IIRG). Professor Nielson 5. _____ an

5. A. said B. told C. answered D. was

audience at the National Press Club yesterday that tomorrow's Internet will be as common 6. _____ electricity. "If you're at the office

6. A. in B. about C. on D. as

and 7. _____ you've got to work late.

7. A. house B. lab C. if D. to

You'll be 8. _____ to log on, quickly send a message to your VCR to record the football

8. A. required B. asked C. told D. able

game 9. _____ you can watch it when you get home," he said.

9. A. also B. that C. so D. which

"When you're sitting 10. _____ bed, you'll be able to roll over and say, "Get breakfast!"

10. A. in B. on C. with D. from

to some device that's hooked up to the Internet. The device 11. _____ translate your

11. A. to B. will C. on D. had

See **INTERNET** on page 2

Internet *continued from page 1*

command into a computer than will tell your toaster to get the breakfast ready." Professor Nielson is not alone in 12. _____.

12. A. lab B. the C. an D. as

optimistic view of the role that the Internet will 13. _____ in our everyday lives.

13. A. work B. be C. get D. play

Yesterday in Paris, for 14. _____

14. A. the B. example C. tomorrow D. Nielson

Francesco Xiao, former chief executive of Italian Mobile phone company Omnitel SpA and now chief executive 15 _____ appliance maker

15. A. to B. as C. of D. who

Merloni Eltrodomestici SpA, unveiled the first washing machine 16 _____ can be controlled remotely by a cell phone or the Internet.

16. A it B. that C. probably D. which

It goes on sale December 9th in Italy, as the first component of what is to 17 _____ .

17. A. do B. make C. be D. use

an entirely interactive kitchen Mr. Xiao said he expects to sell 50,000 Carmelita2000.com machines 18 _____ year, out of the seven

18. A. next B. first C. in D. a

million machines Merloni sells annually. Within three 19 _____, he expects about a third of Merloni's appliances will have some sort of digital connection.

19. A. of B. years C. appliances D. time

Mr. Xiao 20 _____ showed off a prototype touch-screen Web browser named "Leonardo"

20. A. has B. also C. had D. did

which will allow the owner to send cooking instructions 21 _____ a web site straight into

21. A. from B. to C. at D. on

art owner 22 _____ also allows the owner, to do many other amazing things.

22. A. He B. But C. It D. And

With files form Jeremy Whyte London of the Bloomsberg Press

Analysis

In our sample, slightly over half of the examinees were from Asia (i.e., Chinese, Japanese, and Korean) so I am able to group them together and contrast them to the non-Asian examinees. Likewise, the four faculties of studies are Engineering, Arts/Social Science, Business (Public Affairs and Management), and Science. In addition, the Carleton testing has each examinee completing two cloze tests in the same session (with some people taking Test 1 first and others taking Test 2 first, randomly). The focus of my study is Test 1 whereas Test 2 is used in creating a composite score which will serve as an additional matching variable. That is, we created composite a "language" score which is the composite of Cloze 1 and 2. I am, in essence, using a multiple matching variable design.

Unlike nearly all DIF studies reported in the literature, this study used logistic regression approach with multiple variable matching and multiple group variables. The logistic model can be expressed in words as

Y = (multiple matching composite language) + (faculty of study) +
(Asian / non-Asian)

Y = (matching) + (faculty) + (Asian/ non-Asian)

Where Y is the natural logarithm of the ratio of probabilities, and the first term is the matching variable. Faculty is three dummy coded variables, and the Asian/ non-Asian components one dummy coded variable. Note that although modeling an interaction is conceptually possible, we do not have such an interaction in our model because of sample

size limitations. Therefore, we are doing an ANCOVA because a more complex model would require a larger sample size.

The DIF analysis using logistic regression approach will be calculated on each of the 24 items, at a Type I error rate of .01 for each. The odds-ratio, Exp. (B) labeled in the Variables in the Equations of SPSS output, is used as the measure of effect size for the DIF analysis. The analyses follow in two steps. First, each item is studied for DIF for the faculty, and then the Asian/ non-Asian variables. Once these DIF items are flagged, it is crucial to investigate the potential sources of DIF across the groups. The flagged items are not necessarily biased items. A DIF item simply indicates that matched groups of examinees perform differently on an item without offering any information about the cause(s) of the difference in performance (Allen & Holland, 1993).

Once the results are reported, the study will examine the DIF items because evaluation of items with DIF can provide insight into resources or distracters that might be related to DIF and knowledge about which distracters differentially attract a specific subgroup may help us understand the respondents' cognitive processes (Schmitt, Hollan, and Dorans, 1993).

Results and Interpretation

I will discuss those items for which a statistically significant effect was found for either of the DIF variables: faculty or Asian/non-Asian. Items 5, 10 and 24 showed some statistically significant DIF.

For item 5, the Asian/non-Asian DIF variable was statistically significant, Wald =8.03, df =1, p = . 005, odds-ratio of 3.6. The odds-ratio should be interpreted to mean

that after matching on language proficiency, item 5 was 3.6 times easier for the Asian group than the non-Asian group. The faculty of study variable was not statistically significant for this item.

For item 10, the Wald =16.38, df =1, p = .0001, odds-ratio = 4.2 for the Asian/non-Asian variable, therefore it demonstrates statistically significant DIF. As in the previous item, item 10 was over four times easier for the Asian group than the equally matched non-Asian group. Likewise, the faculty of study variables was statistically not significant.

DIF for the Asian/non-Asian variable was statistically significant for item 24 (Wald = 10.87, df =1, p = .001). The odd-ratio is 3.4 in favor of the non-Asian group. Unlike the previous two items, this item is easier for the non-Asian group that is equally matched on the ability being tested. Again, faculty of study did not have an effect.

Now let me turn to each of these three items for a closer study of sources of DIF. It should be noted that the overall conclusions do not support a familiarity/non-familiarity to the test format explanation. Therefore, I will focus my interpretation of the DIF results on the educational and curricular explanations.

Item 5. While such ideas sound far fetched, Mr. Xiao, former CEO of Omnitel, said the technology 5 _____ already available

* A. is B. he

C. would D. had

In this regard, it is argued here that the grammar orientation used in learning English by Asian/Chinese ESL students may be in favor of their test performance in Item 5 because the Item 5 deletion is strictly grammatically-based. To infer the corresponding tense(s) and/or voice(s) of English grammar from clue words has become one of the favorite strategies applied by most Chinese ESL and EFL learners. To meet recursive syntax, Chinese examinees are generally more sensitive to discrete nature of English (e.g., a part of the speech). Although the clue word "already" in the stem of Item 5 is an obvious sign indicating perfect tense to Chinese examinees, yet their sensitiveness to the part of speech of the adjective word "available" in the latter part of the stem could make them give up perfect tense but use a be-predicative pattern. If the examinees from non-Asian countries viewed only communicative competence as authenticity, they might solely skim the general meaning of Item 5 with less attention to sentence structures. In fact, in oral English, either A or B or C or D in Item 5 does not make a big difference; people can understand what is said. Therefore, Item 5 with micro-level deletions which focus on lexical choices of words and their interaction with other (see rationales from Bachman, 1985 and Bensoussan and Ramraz, 1984 above in this thesis) provided an applying context for Asian/Chinese group who had been nurtured in "grammar rules".

Similarly to Item 5, Item 10, one of the three items (Item 9, Item 10,and Item 11) in the same stem, is another big DIF item in favor of the Asian/Chinese group. For the convenience of analysis, Item 9's and Item 11's answers have been filled in.

Item 10: Professor Nielson believes that the global costs of using the Internet will dramatically fall over the next few years, based 9 on either a flat monthly 10.____for a connection 11 based on volume, a model currently applied to electricity and water consumption.

* A. cost	B. or
C. pays	D. make

Grammar approach makes Asian students highly emphasize sentence formation and structural properties. This could lead to DIF in Item 10. Linguistic research in the past two decades has made Asian educators (e.g., Chinese language teachers) aware of the complexity of patterning in English. It is an impossible task for them to list all of this complexity in a skill checklist; therefore, Chinese ESL instructors have opted to select salient patterns that might prove problematic to Chinese ESL learners whose task is to obtain the meaning from given sentences. Sentence formation was made into an important position in this situation. This is consistent to the assumption posed by some early theoretical linguistic views backed by psycholinguistic research (Clark and Clark, 1977; Fodor, Bever, & Garrett, 1974). According to this assumption, derivation of the meaning of sentences is based on the ability to recognize the logical subject(s), verb(s), and object(s) encoded in sentence constituents, as well as ability to recognize semantic relationships linking together the information found in sentence constituents.

In this Confucian-influenced society, memorization is considered an important way to obtain knowledge; Chinese examinees usually memorize the common coordinating conjunctions like *either ... or*, *neither... nor*, *both...and, not only... but also* and the similar or the idiomatic verb phrase like *pay for*, *look at*, *listen to* etc. when they

start to learn English. However, it is not proper to associate Item 10 blank with coordinating conjunction, either… or, which requires to connect two equivalent parts. The clues, the indefinite article "a" and the adjective word "monthly", here indicated that the right answer for Item 10 must be a noun. Both the option A "cost" and C "pays" can be nouns, but "pay" is an uncountable noun. Therefore, "C" is the right answer.

Item 10 suggests that it is precisely the presence of such a coordinating conjunction that facilitates comprehension of sentences, while, on the other hand, it could easily mislead to a wrong answer if examinees only look at the superficial features.

In opposite to Item 5 and 10, Item 24 was a more difficult item for Asian/Chinese group in terms of its odds-ration of Exp.(B) was 3.4 in favour of Non-Asian group.

Item 24 Professor Nielson has his eyes squarely on the future.
That future 24 _____ filled with possibilities for the Internet.
If scientists like Professor Nielson have their way, almost every aspect of future life will be enhanced by use of the Internet.

A. to	*B. is
C. will	D. can

As stated early in this thesis, cloze items are context-based because answering one item correctly may aid a student in answering other items because examinees will then have more complete context. This means more context information, the more probability of getting the correct answer. However, context could also contaminate examinees'

judgment, especially in the limited examination time. This could be the possible situation for Item 24, the last item in Test 1.

The word "future" in the Item 24 stem subconsciously activated Chinese examinees' background knowledge relating to the future tense, "will /shall + original v." The careless examinees could be misled by this context information especially when they had time constrain problem to complete this last item in the test. The context after Item 24 (see the italicized sentence in Item 24), could also impact the examinees' response because of the information, "...every aspect of future life will be enhanced...."

Item 24 showed that the test-takers' previous academic experience (e.g. familiarity with test format) is not always in favor of their test performance. This result has added something new to the previous studies which only claimed the positive impact of examinees' familiarity with test formats on their test scores. Until now the best way to assess the cognitive ability of individuals whose primary language is not English is yet unknown. To identify the right answer to Item 24 actually mirrors the examinees' overall language competency, especially pertaining to the intensity of thinking and problem solving activity. Item 24 is an application item which not only measure understanding, but typically at high level than that of comprehension.

Conclusion and Significance of the Study

Conclusion of the Study

The purpose of this study was to investigate the potential differential item functioning (DIF) on cloze tests based on (a) country of origin, and (b) program of study in terms of the academic faculty the student is enrolled in. I found that country of origin

came up on three occasions to be DIF. The results were somewhat inconsistent in that the Asian group was favored for only two of the items (5, 10). This does not lend support for my hypothesis that test format familiarity would favor test performance for the Asian group. That is, if test format were the DIF factor, it would be present in the same direction (i.e., Asian group being favored of all of the DIF items; whereas Item 24 favored the non-Asian group). When DIF was found, and explanation was sought in terms of grammatical effects and Asian language classroom curricula. That is, it is speculated that for some DIF items the Asian group approached the item with a particular strategy. The explanation provided is intended to encourage future research on this topic using qualitative methods such as a talk-aloud protocol.

What was consistent, however, was that faculty of study did not show any DIF.

Limitations

The most important limitation of this study is the interpretation of the grouping variables for the DIF analysis. This limitation holds true for most DIF studies. That is, DIF studies often use intact groups that represent a variety of impacts and factors. For example, when one studies gender DIF the variable gender represents a variety of sources of DIF (e.g., up-bring, family effects, social and educational effects, biological effects, etc..). Likewise, in this study the grouping variable "country of origin" was difficult to interpret because it represents a variety of potentially influential forces such as (a) test format familiarity, (b) educational and curricular differences, and (c) cultural social differences to name but a few sources of DIF. In short, DIF studies are quasi-experimental studies that use intact groups of individuals and hence suffer of all of the

potential threats to internal validity that these quasi-experimental studies naturally have. Perhaps a stronger way to address the hypothesis of test familiarity involves doing a study in which the examinees are asked ahead of time their familiarity with the cloze test format and then a grouping variable created from the responses.

Significance of the Study

In spite of the limitations listed above, this study has made the following contributions. First, the study was a methodological advance in terms of the complexity of the DIF model. Typically, background variable effects are not investigated using DIF methodology and therefore the conclusions from those studies are suspect. The DIF methodology has one matching examinees before investigating the background variable effects. This matching is vital to the appropriate interpretation. In addition, unlike many previous DIF studies in language testing that have used only one test performance (e.g. Bachman and Palmer, 1981), this study used multiple matching variables for test performance data.

Methodologically, this study has shown that the very notion of differential item functioning by groups are caused by the qualitatively-different test takers with diverse language backgrounds, and thereby will have supported the importance of the investigation of DIF in language testing due to the qualitatively-different test takers' curriculum effects whereas the previous study only focused on the examinees' familiarity of test content. Likewise, to my knowledge, this is the first study to report a complex DIF model. What's more, this study can provide useful insight for language testing practitioners. For example, significant relationships between the strategies applied by test

takers to the test because of their familiarity with test formats and their test performance can inform test users and test developers as well as language and curriculum developers and language teaching material writers.

Second, the result of this study can shed light on some aspects of test construction such as content validity, construct validation and item writing to language test developers and researchers regarding the factors that influence test performance, and, therefore, about the validity of the theoretical underpinnings that inform these language tests.

Table 1 COUNTRY OF ORIGIN

Country of Origin	Frequency	Percent
Asian		
China	112	51.9
Taiwan	3	1.4
Japan	7	3.2
Korea	3	1.4
Non-Asian		
Afghanis	2	.9
Africa	1	.5
Bolivia	1	.5
Canada	1	.5
Columbia	2	.9
Ecuador	1	.5
Egypt	9	4.2
El Salvador	1	.5
France	1	.5
India	3	1.4
Indonesia	1	.5
Iran	14	6.5
Iraq	4	1.9
Italy	2	.9
Jordan	5	2.3
Lebanon	1	.5
Libya	3	2.8
Mexico	4	1.9
Pakistan	4	1.9
Palestine	2	.9
Romania	2	.9
Russia	1	.5
Kuwait	1	.5
Saudi Arabia	3	1.4
Sri Lanka	3	1.4
Somalia	3	.9
Turkey	2	.5
UAE	1	.9
Ukraine	2	.5
Vietnam	2	.9
Yemen	1	.5
Yugoslavia	1	.5

References

Ackerman, T. A., Simpson, M. A., & de la Torre, J. (2000). A comparison of the dimensionally of TOEFL response data from different first language groups. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, Lousiana.

Alderson, J. C. (1978a). The effect of certain methodological variables on cloze test performance and its implications for the use of the cloze procedure in EFL testing. Paper presented at the Fifth International Congress of Applied Linguistics, Montreal.

Alderson. J. C. (1979). The cloze procedure and proficiency in English as a foreign language. TESOL Quarterly ,13 , 219-23.

Alderson. J. C. (1980). Native and non-native speaker performance on cloze tests. Language Learning, 30, 59-76.

Ackerman, T.A., Simpson, M. A. & de la Torre, J. (2000). A comparison of the dimensionally of TOEFL response data from different first language groups. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

AERA/APA/NCME (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education), (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Bachman, L. F. (1982). The trait structure of cloze scores. TESOL Quarterly, 16, 61-70.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. TESOL Quarterly, 19, 535-556.

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

Bachman, L. F., & Palmer, A. S. (1996). Language Testing in Practice. USA: Oxford University Press

Bensoussan, M., & Ramraz, R. (184). Testing EFL reading comprehension using a multiple choice rational cloze. The Modern Language Journal, 68, 230-239.

Bleistein, C. A. (1986). Application of item response theory to the study of differential item characteristics: A review of the literature. Research Memorandum. Princeton, NJ: Educational Testing Service.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting SIBTEST detection procedure. Behaviometrika, 23, 67-95.

Brazil, D. (1995). A grammar of speech. Oxford: Oxford University Press.

Brown, J. D. (1980). Relative merits of four methods of scoring cloze tests. Paper presented at the Thirteenth Annual TESOL Convocation, Boston.

Brown, J. D. (1983). A closer look at cloze: Validity and reliability. In J. W. Oller, Jr. (Ed.), Issues in language testing research (pp. 237 –250). Rowley, MA: Newbury House.

Brown, J. D. (1984). A cloze is a cloze? In J. Handscombe, R. A. Orem, & B. P. Taylor (Eds.). On TESOL '83 (pp. 109 –119). Washington, DC: TESOL.

Brown, J. D. (1988). Tailored cloze: Improved with classical item analysis technique. Language Testing, 5, 19-31

Brown, J. D. (1992). What text characteristics predict human performance on cloze test items. In the Proceedings of the Third Conferences on Language Research in Japan (pp. 1 –26). Urasa, Japan: International University Japan.

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. Language Testing,16 , 217-38.

Chapelle, C. A., & Abraham, R. (1990). Cloze method: What difference does it make? Language Testing, 7, 121-146.

Chavez-Oller, M. A., Chihara, T., Weaver, K., & Oller, J. W., Jr. (1985). When are cloze items sensitive to constraints across sentences? Language Learning, 35, 181-206.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency test. Language Testing, 2, 155-63.

Chu, C. C. (1990). Semantics and discourse in Chinese language instruction. Journal of the Chinese Language Teachers Association, 25 , 15-29.

Clark, H., & Clark, E. (1977). Psychology and language. New York: Harcout Brace Jovanovich.

Cortazzi, M., & Jin, L. (1996). Cultures of learning: Language classrooms in China. Society and the language classroom. England: Cambridge University Press.

Curley, W.E., & Schmitt, A. P. (1993). Revising SAT-Verbal items to eliminate differential item functioning. College Board Report 93-2; ETS Research Report, 93-61. New York: College Entrance Examination Board.

Darnel, D. K. (1970). Clozentropy: A procedure for testing English language proficiency of foreign students. Speech Monographs, 37, 36-46.

Diaz, R. (1983). Through and two languages: The impact of biligualism on cognitive development. In Gordon E. W., (Ed.), Review of research in education. Vol. 10. Washington DC: American Educational Research Association, 23-54.

Douglas, D. (2000). Assessing languages for specific purposes. Cambridge: Cambridge University Press.

Erickson, M., & Molloy, J. (1983). ESP test development for engineering students. In Oller, J. W. Jr., (Ed.), Issues in language testing research, Rowley, MA: Newbury House.

Flesch, R. (1951). How to test reliability. New York: Harcourt

Fodor, J., Bever, T., & Garrett, M. (1974). The psychology of language. New York: McGraw-Hill.

Fox, J. D. (2000). It's all about meaning: L2 test validation in and through the landscape of an evolving construct. Unpublished doctoral dissertation, McGill University, Montreal.

Fox, J., Pychyl, T., & Zumbo, B. D. (1997). An investigation of background knowledge in the assessment of language proficiency. In A. Huhta, V. Kohonen, L. Kurki- Suonio, & S. Luoma, (Eds.), Current developments and alternatives in language assessment (pp.367 - 383). Jyväskylä: University of Jyväskylä and University of Tampere.

Gardber, R. C. (1985). Social psychology and second language learning: The role of attitude and motivation. London: Eduward Aenold

Gardber, R. C. (1988). The socio-educational model of second language learning Assumptions, findings and issues. Language Learning, 38, 101-26.

Ginther, A., & Stevens, J. (1998). Language background and ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In Kunnan, A. J., (Ed.), Validation in language assessment. Mahwah, NJ: Lawrence Erlbaum, 169-94.

Hale, G. A., Rock, D. A., & Jirele, T. (1989). Confirmatory factor analysis of the Test of English as a Foreign Language. TOEFL Research Report No. 32. Princeton, NJ: Educational Testing Service.

Hinofotis, F. B. (1987). Cloze testing: An overview. In M. Long & J. Richards, (Eds.) Methodology in TESOL: A look of readings (pp. 412 –417). Rowley, Mass: Newbury House.

Hu, C., & Hsian, L. A. (2000). Toward an understanding of writing in a second language: evidence and its implications from L2 writers of Chinese. Dissertation Abstracts International, 61, 1264A.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? TESOL Quarterly, 24 , 61-83.

Klein-Braley, C.(1985). A cloze-up on the C-Test: A study in the construct validation of authentic tests. Language Testing, 2, 76-104.

Kunnan, A. J. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: An approach to construct validation. Language Testing 11, 225-52.

Markham, P. L. (1985). The rational deletion cloze and globle comprehension in German. Language Learning, 35, 423 –430.

McPeek, W. M., & Wild, C. L. (1992). Identifying differential functioning items in the core battery. ETS Research Report No. 92-62. Princeton, NJ: Educational Testing Service

Morrow, K. (1979). Communicative language testing: Revolution of evaluation? In C. K. Brumfit, & K. Johnson (Eds.), The communicative approach to language teaching (pp. 143-159). Oxford: Oxford University Press

Morrow, K. (1982). Testing spoken language. In J.B. Heaton (Ed.), Language Testing (pp.56-58). London: Modern English Publication.

Oller, J. W., Jr. (1979). Language tests at school: A pragmatic approach. London: Longman

Oltman, P. K., Stricker, L. J., & Barrows, T. (1988). Native language, English proficiency, and the structure of the Test of English as a Foreign Language for several language groups. TOEFL Research Report No. 27. Princeton, N.J: Educational Testing Service.

Osterlind, S. J. (1985). Constructing Test Items: Multiple-choice, Constructed-Response, Performance, and Other Formats. Kluwer Academic Publisher.

Peretz, A. S. (1986). Do content area passages affect student performance on reading comprehension tests? Paper presented at the twentieth meeting of the International Association of Teachers of English as a Foreign Language, Brighton, UK.

Sasaki, K. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. Language Testing 8, 93-111.

Shin, K. (1983). Cloze-hou ni yoru nihongo nooryoku no sokutee[Assessing Japanese proficiency by cloze]. Kyoikyu Kenkyu: ICU Gakuho,25, 141-177.

Shin, K. (1990). Nihongo cloze test kara Nihongo henkee C-Test e; Some problems on rating]. Nagoya University Soogoo Gengo Center Gengo Bunka Ronshuu, 11, 213-225.

Spolsky, B. (1995). Measured words. Oxford: Oxford University Press

Swinton, S. S., & Powers, D. E. (1980). Factor analysis of the Test of English as a Foreign Language for several language groups. TOEFL Research Report No. 6. Princeton, NJ: Educational Testing Service.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. Journalism Quarterly 30, 415-33.

Taylor, W. L. (1954). Application of "cloze" and entropy measures to the study of contextual constraint in samples of continuous prose. Doctoral thesis, Urbana: University of Illinois.

Zumbo, B., & Hubley, A. M. (in press). Differential Item Functioning and Item Bias. Encyclopedia of Psychological Assessment. Thousand Oaks: Sage Press.