

**The Identification and Interpretation of Group Differences on the Canadian Language
Benchmarks Assessment Reading Items**

Marilyn Abbott

Centre for Research in Applied Measurement and Evaluation

University of Alberta

Paper presented at the Annual Meeting of the National Council on Measurement in Education

(NCME)

San Diego, California, USA

April 15, 2004

Abstract

This study was undertaken to test the hypothesis that some of the items included in the Canadian Language Benchmarks Assessment Reading Subtest favour certain cultural groups. For example, it was posited that Mandarin speakers, who have a tendency to use bottom-up, local reading strategies, would perform better on particular questions than Arabic speakers, who tend to use top-down, global reading strategies. Two samples of examinees were drawn from previously administered CLBA Form 1, Stage II Reading Assessments. One sample consisted of 250 Arabic speaking immigrants, and the other consisted of 250 Mandarin speaking immigrants. Two ESL reading experts classified each of the 32 CLBA reading items into one of five bottom-up or five top-down reading strategy categories. Differential bundle functioning analyses were conducted to determine whether groups of CLBA items, classified according to the bottom-up, top-down organizing principle, functioned differentially for equal ability Arabic and Mandarin ESL learners. Systematic group differences were found in two of the bottom-up strategy categories and three of the top-down categories. Items involving breaking lexical items into smaller parts and matching key vocabulary in the text to key vocabulary in the item were found to favour the Mandarin speaking examinees, whereas items involving skimming for gist, connecting or relating information presented in different parts of the text, and predicting what may happen in a related scenario were found to favour the Arabic speaking examinees. In sum, five of the ten hypotheses based on the reading strategy framework were supported by the bundle analyses. Therefore, to some extent, this study provides evidence for the validity of the bottom-up, top-down reading strategy organizing principle.

The Identification and Interpretation of Group Differences on the Canadian Language Benchmarks Assessment Reading Items

The Canadian federal government provides language training to immigrants who have limited or no proficiency in an official language on arrival. Many immigrants, however, are unable to access the federally funded maximum of 1500 hours of instructional support. Consequently, it is necessary that immigrants' language levels be accurately assessed so they can be placed in the most appropriate levels of instruction. Otherwise their time and the federal support they receive will be wasted.

It is also crucial to ensure that placement tests provide equal opportunities for all immigrants to demonstrate what they know about “the construct(s) the test is intended to measure” (AERA, APA, & NCME, 1999, p. 74). For example, if a reading comprehension test is made up of question types that elicit strategies that are well developed in one specific linguistic/cultural group but not in another, then the assessment may unfairly favour the first group over the second. In other words, if the questions involve reading strategies that are more familiar to members of one language or cultural group, then the assessment may be easier for individuals of that group. Fair, equitable assessment is tailored to the individual learner's instruction context and background including his or her prior knowledge, cultural experience, language proficiency, cognitive style, and interests (Joint Advisory Committee, 1993). Therefore, substantive and statistical research devoted to examining and promoting accuracy and fairness when developing and using assessment tools such as the Canadian Language Benchmarks Assessment (CLBA) is essential.

Since its inception in 1996, the CLBA has predominantly been used to assess the English language skills of newcomers to Canada. The CLBA is a task-based tool (i.e., it includes a range

of tasks of different types) designed to assess language proficiency in the areas of listening, speaking, reading, and writing. Initially, the main purpose of the assessment was to determine newcomers' entry points in English as a second language (ESL) programs. Currently, the CLBA is also being used as a means of establishing admissible levels of English language proficiency in some post-secondary institutions and professions in Canada.

To date, the extent to which the CLBA reading items may favour examinees from particular language or cultural groups has not been the focus of any empirical research. In an attempt to fill this void and to extend our understanding of cross-cultural reading strategy use, the purpose of the current study is to test the hypothesis that some of the items included in the CLBA Reading Assessment favour Arabic speaking examinees over Mandarin speaking examinees and vice versa. For example, it may be the case that Mandarin speakers, who have a tendency to use bottom-up, local reading strategies, would perform better on particular questions than Arabic speakers, who tend to use top-down, global reading strategies. Differential bundle functioning (DBF) analyses were conducted to determine whether groups of CLBA items function differentially for Arabic and Mandarin first language immigrant groups. Arabic and Mandarin ESL learners were selected for three main reasons: first, they are currently two of the largest recent immigrant groups in Canada; second, both languages are radically different from English and from each other in terms of orthographic script; and third, the two groups are culturally distinct.

In this paper, I begin with a review of the literature on ESL reading strategies and then I consider the effects of culture, education, and language on the development and use of reading strategies. Next, I discuss the studies that have examined differential item functioning in ESL language proficiency and placement tests. Finally, I illustrate an application of the Roussos-Stout

(1996) multidimensionality-based DIF analysis framework to the study of Arabic and Mandarin speaker differences in ESL reading strategies on the Canadian Language Benchmarks Reading Assessment.

Literature Review

Reading Comprehension Strategies

Researching second language reading comprehension strategies has proved to be a complex endeavour as the concept of strategy is difficult to define, observe, measure, describe, and classify. Despite the lack of consensus regarding what constitutes a strategy, numerous researchers use the term *strategies* to refer to the mental processes or behaviours that language learners employ in second language acquisition, second language use, or second language testing situations (Alderson, 1984; Cohen, 1998; Hosenfeld, 1977; O'Malley & Chamot, 1990; Oxford, 1990; Purpura, 1997). According to Cohen (1998), language use and test-taking strategies are the “mental operations or processes that learners consciously select when accomplishing language tasks” (p. 92). By adapting this definition to the context of reading, reading comprehension strategies may be defined as the mental operations or comprehension processes that readers select and apply in order to make sense of what they read. Since strategies are generally considered to be conscious or at least potentially conscious, they are open to inspection (Weinstein & Mayer, 1986).

Examples of some commonly identified reading strategies are skimming for gist, scanning for details, guessing, recognizing cognates and word families, predicting, activating general knowledge, making inferences, following references, and separating main ideas from supporting ideas (Barnett, 1988). Although some reading experts (Davis, 1968; Drum, Calfee, & Cook, 1981; Munby 1978) classify these strategies as reading skills, microskills, or subskills,

others (Alexander & Jetton, 2000; Duffy, Roehler, Sivan, Rackcliffe, Book, Meloth, Vavrus, Wesselman, Putnam, & Bassiri, 1987; Robb, 1996) refer to these behaviours as strategies as they assume that a reading skill becomes a strategy when the reader can use it independently, reflect on it, and understand what it is, how it works, and when to apply it to new texts. This assumption will be adopted in the current study.

Reading Strategy Research

While a plethora of questionnaire research results indicate that cultural background affects second language learning strategy selection and use (e.g., Bedell & Oxford, 1996; Harshbarger, Ross, Tafuya, & Via, 1986; Levine, Reves, & Leaver, 1996; Reid, 1995; Willing, 1988), few studies have specifically focused on how second language reading strategies interact with first language and cultural background to affect test performance. It has been determined, however, that ESL reading comprehension tests often focus on low-level linguistic cues, which tend to reward bottom-up as opposed to top-down reading strategies (Hill & Parry, 1989, 1992; Purpura, 1997). Bottom-up reading comprehension strategies are data-driven, whereas top-down strategies are conceptually- or hypothesis-driven (Carrell, 1983). Parry (1996) found that when attempting English academic reading tasks, different cultural groups use strikingly different reading strategies that she claims are related to their different language backgrounds and different experiences of literacy. For example, whereas Chinese students showed a definite preference for bottom-up methods, Nigerian students reported a strong tendency to use top-down strategies. In another cross-linguistic study of ESL reading, Fender (2003) discovered that native Arabic ESL learners were more accurate in comprehending and integrating words into larger phrase and clause units than Japanese ESL learners. This suggests that Arabic ESL learners may have a proclivity for using top-down reading strategies.

Examples of Bottom-up, Local and Top-down, Global Reading Strategies

Examples of bottom-up, local, language based reading strategies that focus primarily on word meaning, sentence syntax, or text details are

1. breaking lexical items into smaller parts;
2. scanning for specific details or explicitly stated information;
3. finding a synonym or a paraphrase of the literal meaning of a word, phrase, or sentence;
4. relating verbal information to accompanying visuals; and
5. matching key vocabulary in the text to key vocabulary in the item.

Some top-down, global, knowledge-based reading strategies that focus primarily on text gist, background knowledge, or discourse organization are

1. skimming for gist/identifying the main idea, theme, or concept;
2. connecting or relating information presented in different parts of the text;
3. drawing an inference based on information presented in the text;
4. predicting what may happen in a related scenario/speculating beyond the text; and
5. recognizing discourse format (discriminating between: fact and opinion or cause and effect; focusing on the way the text is organized).

These strategies appear in standard classifications employed in one or more of the following studies: Anderson (1991), Block (1986), Carrell (1989), Phakiti (2003), Pritchard (1990), and Purpura (1997).

The Influence of Culture, Education and Language on the Acquisition of EFL/ESL

Although some cultural and educational factors have been shown to influence strategy preferences (e.g., Bedell & Oxford, 1996; Levine et al., 1996; Harshbarger et al., 1986; Pritchard, 1990; Reid, 1995; Willing, 1988), little explanation has been provided as to why this

occurs. Thus, in this section, an attempt will be made to explain why intermediate Arabic and Chinese ESL learners tend to use different reading strategies.

Instructors of reading in English influence the way their students approach text by teaching them to read in particular ways. For example, it is often cited that Chinese teachers tend to use traditional teacher-centered approaches to teaching EFL (Burnaby & Sun, 1989; Penner, 1995; Parry, 1996). As a result, Chinese EFL learners are taught to pay close attention to word level cues (i.e., morphology and syntax). According to Fischer-Kohn (1986, cited in Kohn 1992), Chinese teachers of reading in English encourage their students to

1. read slowly and take care that they know each word as they go;
2. vocalize or voice the material, either loudly or silently;
3. reread difficult sentences until they are understood;
4. look up definitions of all unknown words in a dictionary; and
5. analyze complex structures carefully. (p. 121)

Thus, it appears that Chinese EFL learners are taught to use bottom-up strategies as they are expected to carefully scrutinize each word in the text and memorize grammar rules and exceptions (Kohn, 1992).

In contrast, the general trend in Arab nations is to place more emphasis on student centered EFL activities that encourage linguistic interaction through the use of authentic, real-life tasks (Kharma, 1998). These types of communicative activities focus on developing functional language skills in a learning environment that stresses meaning over form. As Parry (1996) suggests, authentic reading activities that emphasize reading for meaning tend to encourage a global, top-down approach to text. Therefore, it is likely that the exposure Arab EFL students receive to communicative activities promotes the development of top-down reading strategies.

Numerous ESL instructors have noticed that Chinese ESL students tend to use a dictionary more than Arab ESL students. The reason for this differential use is likely reflected in both their linguistic and educational systems. Thompson-Panos and Thomas-Ružić (1983) maintain that Arab students are not highly skilled in using dictionaries when reading and writing because the words in Arabic dictionaries are arranged according to their roots. In English, this would be similar to looking up the word *misconceived* under the entry for *cept* (Thompson-Panos & Thomas-Ružić, 1983). If the educational system does not emphasize the development of such skills when learning the first language (L1), these skills will not be available to transfer to L2 learning, and consequently will not promote a bottom-up approach to reading in an L2. On the contrary, most Chinese students tend to rely heavily upon their dictionaries and as a result usually have well-developed dictionary skills, which encourage the development of a bottom-up approach to reading (Parry, 1996).

The Effects of Linguistic Differences on the Acquisition of EFL/ESL

Research suggests that language-specific differences are related to differences in processing skills and strategies in reading (Chen, 1992; Fender, 2003). For example, as stated in the introduction, in a cross-linguistic study of ESL reading skills, Fender (2003) found that Arabic ESL learners were more accurate in comprehending and integrating words into larger phrase and clause units than the Japanese ESL learners in the study. Japanese (kanji), like Chinese, uses an orthography that encodes language at the level of morphemes, which in general correspond to words and affixes (Chen, 1992). Therefore, one may hypothesize that Chinese ESL learners would also have difficulty with word integration when reading in English.

Since the vowels are not represented in Arabic orthography, Arabs may be less dependent on local cues in the printed word when reading. If reading in Arabic encourages a

reliance on higher-level contextual cues and strategies, it is possible that the Arabic ESL learners in Fender's (2003) study were more successful integrators than the Japanese ESL learners because they effectively transferred their well-developed L1 reading strategies to the L2 reading task. It is likely that the reduction of the extent of Arab readers' dependence on the visual stimulus causes them to develop more effective top-down reading comprehension processes. As a result, when processing printed material in English, Arabic ESL learners may rely more upon their background and contextual knowledge than upon their linguistic knowledge and consequently have a proclivity for using top-down reading strategies over bottom-up ones. On the contrary, it is possible that the careful approach that Chinese ESL learners take may cause them to be distracted by less relevant textual information and as a result they may not be as skilled at integrating words into larger units.

Native speakers of Chinese, however, develop a sophisticated set of orthographic processing skills through their literacy experiences. When compared with printed words in alphabetic (e.g., English) or consonantal (e.g., Arabic) orthographies, Chinese encodes morphemes with much less phonology (Akamatsu, 1999). Consequently, while Chinese word recognition requires extensive orthographic processing skills, alphabetic or consonantal orthographies require a greater connection to phonemes and phonology. Therefore, Chinese ESL learners may be able to utilize their L1-based processing skills to develop a set of graphic ESL word representations that facilitate English word processing.

Although L1 Arabic literacy skills are developed through reliance on phonological processing skills as Arabic orthography has a highly consistent set of grapheme-phoneme (letter-sound) correspondences, more mature readers must learn to use an orthography that does not include diacritic marks that signal the vowels (Abu-Rabia, 1999). In comparison, reading in

English encourages greater reliance on (a) phonological skills for decoding words with regular grapheme-phoneme correspondences, and (b) orthographic processing skills for decoding words with grapheme-phoneme irregularities (e.g., business, cough, iron) (Katz & Frost, 1992).

Therefore, it is likely that the Arabic and Chinese ESL learners' primary L1 processing skills and strategies that have been developed through exposure to distinct languages and literacy practices will differentially influence the development of their ESL processing skills and strategies.

Differential Item Functioning (DIF)

DIF is present when examinees from distinct groups have different probabilities of answering an item correctly after controlling for overall test performance (Shepard, Camilli, & Averill, 1981). DIF methods match examinees on ability (usually total test score) to determine whether comparable examinees from different populations perform the same on individual items. For example, one would expect Arabic- and Mandarin-speaking examinees, who have the same total test score, to perform in an equivalent manner on each CLBA item. When comparable examinees do not perform the same on specific test items, the items are said to display DIF. Large DIF indices signify that the items are measuring secondary dimensions that may either be relevant or irrelevant to the construct measured by the test (Shealy & Stout, 1993). If an item is measuring a secondary dimension that is an appropriate part of the intended construct, the secondary dimension is considered auxiliary. Thus the DIF between groups reflects a true difference in the construct and is considered benign. Alternatively, if an item is measuring an unintended secondary dimension, the secondary dimension is considered nuisance. DIF caused by nuisance dimensions reflects bias which may be thought of as systematic error that distorts the meaning of test inferences for members of a specific group, and therefore poses a considerable threat to validity (Camilli & Shepard, 1994).

Much of the research regarding the effects of language background on second language test performance has been concerned with whether EFL/ESL language proficiency and placement tests measure the same constructs for different language groups (e.g., Brown, 1999; Ginther & Stevens, 1998; Kunnan, 1994; Ackerman, Simpson, & de la Torre, 2000). Only a few studies have examined how different language groups perform differently on such tests at the item level (Chen & Henning, 1985; Kim, 2001; Ryan & Bachman, 1992; Sasaki, 1991). These four studies are discussed below.

Chen and Henning (1985) utilized an adapted Angoff delta-plot method (Angoff & Ford, 1973) to identify DIF items on the UCLA English as a Second Language Placement Exam (ESLPE) across Chinese ($n = 77$) and Spanish ($n = 34$) first language groups. The ESLPE consisted of five 30-item subtests: listening, reading, grammar, vocabulary, and writing error correction. Chen and Henning modified the delta-plot DIF detection procedure by plotting difficulty estimates calibrated by the one-parameter IRT model for each item across the two groups on a scatterplot rather than plotting the traditionally used standardized p -values (the proportion of examinees answering the item correctly). The assumption of this modified delta-plot method was that if an item was unexpectedly too difficult for one group and unexpectedly too easy for the other, it would be regarded as exhibiting DIF. Items beyond the 95% confidence interval of the regression line were considered DIF items.

Results indicated that four items favoured the Spanish group. Not surprisingly, the four items were English vocabulary items with close Spanish cognate forms (e.g., the Spanish cognate for 'approximate' is 'aproximado'). The authors concluded that due to the similarities between English and Spanish, the Spanish speakers had a natural advantage over the Chinese speakers with respect to vocabulary recognition. Since vocabulary was relevant to the construct being

measured by the ESLPE, the DIF exhibited by these items may be attributed to an auxiliary dimension of ESL proficiency and deemed benign. However, if the proportion of cognate vocabulary items exceeded the proportion of naturally occurring cognates in the two languages, then the vocabulary subtest would not validly represent the English lexicon. In this case, content representativeness and thus test fairness would become an issue.

Sasaki (1991) conducted a similar study to Chen and Henning's (1985) study in that she also examined DIF in the UCLA ESLPE across Chinese (n = 262) and Spanish (n = 81) language groups. However, she studied a different version of the ESLPE than Chen and Henning, and utilized Scheuneman's chi-square method (Scheuneman, 1979) for detecting DIF in addition to the same modified delta-plot method employed in Chen and Henning's study. Scheuneman's method, like other contingency table approaches, is based on the assumption that after controlling for ability, members of each group are expected to have approximately the same probability of answering each item correctly. To control for ability, Sasaki divided the Chinese and Spanish groups into three ability levels (low, mid, and high) with approximately the same number of students at each level. Then the significance of the differences between observed frequencies and expected frequencies at each of the three ability levels was calculated for each item.

While the modified delta-plot method identified nine DIF items (5 grammar, 4 vocabulary), Scheuneman's method detected 22 DIF items (4 listening, 1 reading, 4 grammar, 7 vocabulary, 6 writing error detection). Substantive analyses of the DIF results indicated DIF favouring the Spanish group on cognate vocabulary items, and DIF favouring the Chinese group on items containing idiomatic expressions. In both cases, DIF could have been attributed to auxiliary dimensions of ESL proficiency and deemed benign. However, since idiomatic

expressions might have been heavily emphasized in the Chinese speakers' instructional backgrounds and not highly emphasized in the Spanish speakers' instructional backgrounds, it was likely that instructional and curricular differences between the two groups had an impact on item performance. Thus, additional investigation into the proportion of idiomatic expression items in the ESLPE is required to address the issue of content representativeness and test fairness.

Using the Mantel-Haenszel (MH) DIF detection procedure (Mantel & Haenszel, 1959), Ryan and Bachman (1992) examined the extent to which items on the Test of English as a Foreign Language (TOEFL) and the First Certificate in English (FCE) functioned differentially for equal ability examinees from Indo-European ($n = 792$) and non-Indo-European ($n = 632$) L1 backgrounds. Indo-European (IE) examinees were native speakers of French, German, Spanish, and Portuguese, and non-Indo-European (NIE) examinees were native speakers of Japanese, Thai, Chinese, and Arabic. The MH delta difference (MH D-DIF) (Holland & Thayer, 1986) was used to estimate the average amount by which the IE group found a given item more difficult than did comparable members of the NIE group.

On the TOEFL, 32 of the 146 items were found to be easier for the IE group, and 33 items were easier for the NIE group. These differentially functioning items were spread across all three sections of the test (i.e., Listening, Structure and Written Expression, and Vocabulary and Reading). However, on the Listening component, the high (MH D-DIF > 1.5) C level DIF items were not split evenly among the groups as five of the C level DIF items favoured the NIE group, while only two C level items favoured the IE group. On the First Certificate in English (FCE), 25 of the 40 reading and vocabulary items were found to exhibit DIF (13 favoured the IE group, 12 favoured the NIE group). However, eight C level DIF items favoured the IE group

while only three C level items favoured the NIE group. The researchers suggested that these differences were not only attributable to differences in the examinees' native languages but also to differences in the examinees' culture and education. Nevertheless, it is not clear whether the DIF may be attributed to auxiliary or nuisance dimensions of ESL proficiency as no substantive analysis was conducted in this study.

In a more recent DIF study, Kim (2001) examined DIF across Asian (n = 467) and European (n = 571) language groups on the Speaking Proficiency English Assessment Kit (SPEAK) using the likelihood ratio test (Thissen, Steinberg, & Wainer, 1988) and the ordinal logistic regression approach (Zumbo, 1999). These DIF detection procedures were selected because they were considered appropriate for examining the polytomous scoring scales used in the SPEAK test where grammar, pronunciation, and fluency were rated using an ordinal scale from 0 to 3. Of the three scoring categories examined in this study (i.e., grammar, pronunciation, and fluency), Kim found that both methods yielded similar results in that the grammar and pronunciation scales' discrimination values functioned differentially across the Asian and European groups. While the grammar scale was better at discriminating between the high and low ability European speakers of English, the pronunciation scale was more discriminating for the Asian group. However, the fluency parameter estimates were very similar across the two groups, suggesting that this scale did not show DIF.

In the studies mentioned above, the researchers used a variety of DIF detection methods with diverse populations to examine the extent to which items from ESL placement and proficiency tests functioned differentially for examinees of equal ability from different first language backgrounds. Although each of these studies provided evidence that linguistic background is one determinant of DIF in ESL test performance, the studies are not without

limitations. For instance, the small sample sizes in Chen and Henning's (1985) study and Sasaki's (1991) study may have affected the accuracy of the IRT difficulty parameter estimates. In addition, because the one parameter IRT model assumed constant item discrimination, differences in difficulty among items were confounded with differences in discrimination among items (Camilli & Shepard, 1994). Furthermore, the 95% confidence interval for determining DIF items in both studies was arbitrary. If narrower confidence intervals had been used, more DIF items would likely have been detected. Additionally, the unbalanced sample sizes in Sasaki's study may have inflated Type I error in the Scheuneman chi-square procedure. The primary limitation in Ryan and Bachman's (1992) study was that they did not conduct a substantive analysis of the DIF items identified by the MH procedure. It is likely that a content review of the items may have shed some light on the sources or factors contributing to DIF in the two language groups. Finally, the small number of scoring categories examined in Kim's (2001) study made it difficult to evaluate the comparability of the two DIF detection methods.

Although the statistical methods utilized in these studies were relatively useful for flagging DIF items, to understand the nature of DIF, content analyses were also required to determine why the items functioned differentially between the groups. However, the researchers' attempts to identify the causes of DIF in many of the items using content analyses were not very successful. For example, of the 22 DIF items identified by Scheuneman's chi-square method in Sasaki's (1991) study, only four of the items had interpretable sources of DIF. Because attempts to understand the "underlying causes of DIF using substantive analyses of statistically identified items have, with few exceptions, met with overwhelming failure" (Roussos & Stout, 1996, p. 360), Douglas, Roussos, and Stout (1996) proposed a confirmatory approach to differential bundle

functioning (DBF). This approach, which is based on the Shealy-Stout multidimensional model for DIF (Shealy & Stout, 1993), was used in the current study.

A Confirmatory Approach to DIF/DBF

The Roussos-Stout (1996) approach to DIF is a two-stage approach designed to link substantive and statistical methods in a DIF analysis framework. In the first stage of this framework, substantive analyses of the test items are conducted in order to generate DIF hypotheses. A DIF hypothesis specifies whether an item or bundle of items designed to measure the primary or intended dimension also measures a secondary dimension or unexpected dimension that is suspected of producing DIF. The second stage in the Roussos-Stout DIF analysis framework involves statistically testing the hypotheses generated in stage one of the analyses. The statistical procedure selected for testing the hypotheses in this study was the *Simultaneous Item Bias TEST* (Stout & Roussos, 1999).

The Simultaneous Item Bias Test

The *Simultaneous Item Bias Test (SIBTEST)* is a commonly used statistical procedure for detecting DIF. *SIBTEST* was selected for use in this study for three main reasons. First, *SIBTEST* has been found to be more accurate in detecting DIF than the Mantel-Haenszel and logistic regression procedures (Bolt & Stout, 1996; Ercikan, Gierl, McCreith, Puhan, & Koh, 2002; Gierl, Rogers, & Klinger, 1999; Jiang & Stout, 1998). The identification of more DIF items may result in a more thorough analysis of the test items leading to a more comprehensive evaluation of the test and the reading strategy framework that was used to group the items in this study. Second, *SIBTEST* uses a regression estimate of the true score, instead of the observed score, to match students on ability, which results in an improved conditioning variable. Third, *SIBTEST*

can be used to test bundles of DIF items. DBF analyses increase statistical power and reduce the number of statistical tests, thereby controlling Type I error (Nandakumar, 1993).

Shealy and Stout (1993) provide a comprehensive and technical discussion of the *SIBTEST* procedure. *SIBTEST* can be used to test DIF hypotheses and quantify the size of DIF by estimating a measure of the effect size ($\hat{\beta}_{\text{UNI}}$) for each item and bundle (Stout & Roussos, 1995). In the *SIBTEST* procedure, items on the test are divided into two subsets, the suspect subtest and the matching or valid subtest. The suspect subtest contains the item or bundle of items believed to measure the primary and secondary dimensions identified in the substantive analysis, whereas the matching subtest contains the items believed to measure only the primary dimension. In other words, the suspect subtest contains items that are suspected of having DIF, while the matching subtest contains items that ideally have no DIF. The matching subtest places the reference and focal group examinees into subgroups at each score level so their performance on items from the suspect subtest can be compared.

The amount of DIF in the studied subtest is reflected in the effect size estimate $\hat{\beta}_{\text{UNI}}$, which is the weighted sum of the differences between the proportion-correct true scores on the studied item or bundle for examinees in the two groups across all score levels. The true scores are estimated using a regression correction described in Shealy and Stout (1993). The weighted mean difference between the reference and focal groups on the studied subtest item or bundle across the k subgroups is given by

$$\hat{\beta}_{\text{UNI}} = \sum_{k=0}^k p_k d_k ,$$

where p_k is the proportion of focal group examinees in subgroup k and d_k is the difference in the adjusted means on the studied subtest item or bundles of items for the reference and focal

groups, respectively, in each subgroup k . For large samples, $\hat{\beta}_{\text{UNI}}$ has a standard normal distribution with a mean of 0 and standard deviation of 1 under the null hypothesis of no DIF.

The statistical hypothesis tested by *SIBTEST* is

$$H_0: \beta_{\text{UNI}} = 0$$

versus

$$H_1: \beta_{\text{UNI}} \neq 0.$$

SIBTEST yields the following test statistic for evaluating the $\hat{\beta}_{\text{UNI}}$ null hypothesis of no DIF:

$$SIB = \frac{\hat{\beta}_{\text{UNI}}}{\hat{\sigma}(\hat{\beta}_{\text{UNI}})},$$

where $\hat{\sigma}(\hat{\beta}_{\text{UNI}})$ is the estimated standard error of $\hat{\beta}_{\text{UNI}}$. *SIB* is evaluated against the standard

normal distribution. A null hypothesis of no DIF is rejected whenever $|SIB| > z_{1-\frac{\alpha}{2}}$. A

statistically significant value of $\hat{\beta}_{\text{UNI}}$ that is positive indicates DIF/DBF against the focal group and a negative value indicates DIF/DBF against the reference group.

Method

Sample

Item level performance data was collected from previously administered CLBA test forms located at the Immigrant Language Vocational Assessment - Referral Centre (ILVARC) in Calgary and the Language Assessment Referral and Counselling Centre in Edmonton. Item level data was collected from 250 examinee test forms in each of the first language groups (i.e., from 250 Mandarin speakers' and 250 Arabic speakers' assessments).

CLBA

In an attempt to control for first and second language proficiency, only those ESL learners who had (a) completed at least 11 years of education in their L1, and (b) completed both stages of Form 1 were included in this sample. It was assumed that learners with this minimum education level would have well developed reading skills and strategies in their L1 and would not have difficulties reading in their L1. It was also assumed that learners who had completed Stage II would have mastered basic decoding skills and basic vocabulary in English. The need to control for L1 and L2 linguistic proficiency is reflected in Alderson's (1984) view that the skills, strategies, and knowledge from the L1 can only be transferred to L2 reading if the reader has attained a certain level of proficiency in the L2. Also, to control for the influence of ESL education, only the initial assessments of immigrants who had spent less than two years in Canada at the time of testing were included in the sample.

Instrument

Canadian Language Benchmarks Assessment - Reading Assessment. The examinees were assessed with the CLBA Reading Assessment. The CLBA Reading Assessment is a reading comprehension test that is predominantly used to place immigrants in appropriate English as a second language classes. Examinees are required to attempt a range of different task types. The Reading Assessment is divided into two stages and there are four parallel forms for each stage.

Only Form 1, Stage II was analyzed in this study for two main reasons. First, Stage I is primarily comprised of bottom-up questions that mainly test vocabulary knowledge rather than reading comprehension (31 out of 32 questions rely on lower-level cues), whereas Stage II has a more even distribution of bottom-up and top-down questions. Second, the minimum sample size requirement of 250 Arabic speakers who had completed both Stage I and Stage II was satisfied with Form 1.

CLBA

Form 1, Stage II of the CLBA consists of 8 dichotomously scored constructed-response questions and 24 multiple-choice, four-option questions. The questions follow four passages, which represent four different genres and range in length from 251 to 547 words.

Procedure

Following the recommendations by Douglas, Roussos, and Stout (1996), a confirmatory approach to grouping the items (i.e., bundling) was used to examine and statistically test the bundles of items that elicited performance differences for equal ability native Arabic and Mandarin speakers on the CLBA reading items. A theoretical framework, hereafter referred to as the reading strategy classification schema or reading strategy framework, provided the organizing principle for grouping the test items together in terms of strategies so the effects of first language and cultural background on differential bundle functioning could be examined.

The reading classification schema was used by two ESL experts to independently code the 32 reading items included in Form 1, Stage II, of the CLBA. After a training session to introduce the raters to the coding schema, each rater was asked to independently classify the questions into the most salient bottom-up or top-down reading strategy category. Thus the items were coded according to the reading strategy that the expert judges believed to be most instrumental in arriving at the answer for each of the CLBA reading items.

Once the judges finished coding all the items, a meeting was held so they could reach a consensus on the item codings which they disagreed. Then the items were grouped into bundles based on the consensus codes and the following DBF hypotheses were tested using the CLBA item level data: Arabic ESL learners will outperform the Mandarin ESL learners on the bundles that contain items which rely on top-down processing strategies, whereas Mandarin ESL learners

will outperform the Arabic ESL learners on the bundles of items that rely on bottom-up processing strategies.

Data Analysis

The computer program titled *Simultaneous Item Bias TEST* (Stout & Roussos, 1999) was used to determine which items and bundles of items displayed statistically significant differential item functioning (DIF). A four-step analysis (see Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001) was used to test the reading strategy hypotheses. First, a single item *SIBTEST* analysis was conducted to provide effect size measures ($\hat{\beta}_{\text{UNI}}$ values) for each of the CLBA Reading Assessment items. Second, the reading classification schema was used as the organizing principle to group and graph the $\hat{\beta}_{\text{UNI}}$ values for these 32 items into the bottom-up, top-down strategy categories that were previously outlined in the section of this paper titled *Examples of Bottom-up, Local and Top-down, Global Reading Strategies*. Third, the graph was visually inspected to identify patterns in the way the bundles were functioning. Fourth, the interpretable bundles were tested at an alpha level of 0.05. To ensure that the matching subtest was a homogeneous measure across the two groups, the matching subtest for the top-down bundles consisted of the 18 bottom-up test items and the matching subtest for the bottom-up bundles consisted of the 14 top-down test items. These analyses were conducted to show whether there were systematic ways in which the two language/cultural groups responded to groups of test items that were presumed to measure the auxiliary secondary dimensions of bottom-up and top-down reading strategies.

Results

Descriptive statistics for the CLBA Reading Assessment are presented in Table 1. The mean total test scores demonstrated that the Mandarin speaking examinees outperformed the

Arabic examinees on the CLBA Reading Assessment by approximately 3% on average. Results suggested that the skewness and kurtosis values were similar indicating that the shapes of the distributions were similar for both groups of examinees. Furthermore, these results indicated overall item difficulty, discrimination, and internal consistency were comparable across groups.

The single-item *SIBTEST* results found 20 of the 32 items to exhibit moderate to high DIF. In total, nine of the items exhibited moderate DIF ($.05 \leq \hat{\beta}_{\text{UNI}} < .10$), and 11 items exhibited large DIF ($\hat{\beta}_{\text{UNI}} \geq .10$). Figure 1 provides a graphical representation of the item $\hat{\beta}_{\text{UNI}}$ values grouped into the ten strategy categories. Positive $\hat{\beta}_{\text{UNI}}$ values favoured the Mandarin speaking examinees, while negative $\hat{\beta}_{\text{UNI}}$ values favoured the Arabic speaking examinees. An inspection of the graph revealed that the items in two of the bottom-up bundles (B1 and B5) and three of the top-down bundles (T1, T2, and T4) functioned in the predicted manner. In other words, all of the items in the two bottom-up bundles (B1 and B5) favoured the Mandarin speakers and all of the items in these three top-down bundles (T1, T2, and T4) favoured the Arabic speakers. Thus, these five interpretable bundles (see Figure 2) were tested against the appropriate matching subtests (i.e., the bottom-up bundles were tested against the 14 top-down items and the top-down bundles were tested against the 18 bottom-up items).

The results for the differential bundle functioning hypotheses tests are reported in Table 2. All five of these bundles yielded statistically significant $\hat{\beta}_{\text{UNI}}$ values ($p < .05$). These results indicated that two of the five bottom-up bundles significantly favoured the Mandarin speaking examinees over the Arabic speaking examinees, and three of the five top-down bundles significantly favoured the Arabic speaking examinees over the Mandarin speaking examinees.

Conclusions and Discussion

The purpose of this study was to identify bundles of items that elicited systematic performance differences for the Mandarin and Arabic speaking examinees on the CLBA Reading Assessment. Although this study found that 62.5% of the items (i.e., 20 of the 32 items) displayed moderate to large DIF, such item level analyses have generally proved uninterpretable, thereby providing insufficient evidence for making decisions regarding the retention and deletion of test items. If all 20 of these items were removed from the CLBA Reading Assessment, this would have a devastating effect on the original test specifications. Furthermore, because it is difficult to interpret such a large number of DIF items, a confirmatory approach to grouping the items and statistically testing the bundles of items was used in this study. A bottom-up, top-down reading strategy classification schema provided the conceptual framework for classifying the items into bundles that reflected ten main reading strategies (i.e., B1 - breaking lexical items into smaller parts, B2 - scanning for details, B3 - finding synonyms or paraphrases, B4 - matching words to key visuals, B5 - matching key vocabulary in the text to key vocabulary in the item, T1 - skimming for gist, T2- connecting or relating information presented in different parts of the text, T3 - drawing an inference based on information presented in the text, T4 - predicting what may happen in a related scenario/speculating beyond the text, and T5 - recognizing discourse format). These ten strategies served as the organizing principle in the differential bundle functioning analyses.

Based on the reading strategy framework, it was predicted that Mandarin speakers would outperform Arabic speakers on the bundles containing items that were assumed to elicit bottom-up strategies, whereas Arabic speakers would outperform Mandarin speakers on bundles of items that were assumed to elicit top-down strategies. These hypotheses were tested using CLBA

Reading Assessment data. Systematic group differences were found for two of the bottom-up strategy categories: B1 - breaking lexical items into smaller parts and B5 - matching key vocabulary in the text to key vocabulary in the item favoured Mandarin speakers. Therefore, on the CLBA Reading Assessment, Arabic speaking examinees were found to systematically differ from Mandarin speaking examinees with comparable CLBA reading test scores on their skill in using these two strategies (B1 and B5). Because the items in these two bundles have a strong focus on word-level strategies, they may have contributed to differential bundle functioning (DBF) favouring Mandarin speakers.

Systematic group differences were also found for three of the top-down strategy categories: T1 - skimming for gist, T2- connecting or relating information presented in different parts of the text, and T4 - predicting what may happen in a related scenario/speculating beyond the text favoured Arabic speakers. This meant that on the CLBA Reading Assessment, Mandarin speaking examinees were found to systematically differ from Arabic speaking examinees with comparable CLBA reading test scores on their skill in using these three top-down strategies (T1, T2, and T4). In sum, five of the hypotheses based on the reading strategy framework were supported by the bundle analyses. Therefore, to some extent, this study provides evidence for the validity of the theoretical organizing principle (i.e., the reading strategy framework).

In contrast, an examination of the $\hat{\beta}_{\text{UNI}}$ values in Figure 1 showed that there was a fair amount of variation in the way the items were functioning across the bottom-up and top-down categories. For example, the items in the B2, B3, and T3 bundles did not consistently favour one group over the other, and the items in the B4 and T5 bundles functioned in the opposite direction than what was predicted. In other words, although the item in B4 was predicted to favour Mandarin speakers, it was found to favour Arabic speakers, and although the item in T5 was

predicted to favour Arabic speakers, it was found to favour Mandarin speakers. Thus, the reading strategy framework tends to inconsistently identify group performance differences for items in the following five strategy categories: B2 - scanning for details, B3 - finding synonyms or paraphrases, B4 - matching words to key visuals, T3 - drawing an inference based on information presented in the text, and T5 - recognizing discourse format.

A second problem, which complicated the DBF analyses in this study, involved the item classification schema. Although coding the items using the reading strategy framework was fairly straightforward, at times the coders found it difficult to classify the items into one specific category. This was not surprising as other researchers have also found it difficult to anticipate the cognitive processes examinees use to answer the questions correctly (Gierl, Bisanz, Bisanz, Boughton, & Khaliq, 2001). This suggests that researchers need to more closely analyze the cognitive demands of the CLBA reading items in order to develop a more valid representation of the construct of ESL reading strategies, skills, and proficiency. Presumably, this would lead to the development of more psychologically meaningful organizing principles that might assist in explaining and interpreting group performance differences on the CLBA Reading Assessment.

Future Research

Follow-up substantive and statistical DIF/DBF studies of additional CLBA test forms should be conducted to determine whether similar patterns emerge for the bundles created using the reading strategy framework. If the same statistically significant bundle differences are found in future studies, this would imply that the items in these bundles are measuring one or more secondary dimensions. As Bolt (2002) suggests, several nuisance dimensions may impact performance on items within a bundle. In the case of the CLBA Reading Assessment, distinct item format effects (e.g., constructed response versus multiple-choice) might be regarded as an

CLBA

additional dimension operating within the secondary bottom-up, top-down strategy dimensions.

Furthermore, as Douglas et al. (1996) suggest, the DIF occurring among the individual item bundles should be carefully examined to gain a better understanding of the secondary dimensions and causes of DIF. Finally, following Gierl, Bisanz, and Bisanz' (2001) recommendations for developing an interpretative framework for understanding group performance differences, further research is required to validate the dimensional interpretations and clarify why the group differences occur on the CLBA. Evidently, such comprehensive studies have the potential to illuminate the effects of linguistic/cultural background on the validity of CLBA test score interpretations. It is assumed that further studies of this nature will promote more responsible, ethical assessment practices that ensure equity in the scoring and interpretation of the CLBA results.

References

- Abu-Rabia, S. (1999). The effect of Arabic word vowels on the reading comprehension of second- and sixth-grade native Arab children. *Journal of Psycholinguistic Research*, 28, 93-101.
- Ackerman, T., Simpson, M., & de la Torre, J. (2000, April). A comparison of the dimensionality 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.
- Akamatsu, N. (1999). The effects of first language orthographic features on word recognition processing in English as a second language. *Reading and Writing: An Interdisciplinary Journal*, 11, 381-403.
- Alderson, J. (1984). Reading in a foreign language: A reading problem or a language problem? In J. Alderson & A. Urquhart (Eds.), *Reading in a Foreign Language* (pp. 1-24). London: Longman.
- Alexander, P., & Jetton, T. (2000). Learning from text: A multidimensional and developmental perspective. In M. Kamil, O. Mosenthal, P. Pearson, & R. Barr (Eds.), *Handbook of reading research*, Volume III (pp. 285-310). Mahwah, NJ: Erlbaum.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Anderson, N. (1991). Individual differences in strategy use in second language reading and testing. *Modern Language Journal*, 75, 460-472.

- Angoff, W., & Ford, S. (1973). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-106.
- Barnett, M. (1988). Reading through context: How real and perceived strategy use affects L2 comprehension. *Modern Language Journal, 72*, 150-160.
- Bedell, D., & Oxford, R. (1996). In R. Oxford (Ed.). *Language learning strategies around the world: Cross-cultural perspectives* (pp. 47-60). University of Hawaii at Manoa: Second Language Teaching and Curriculum Center.
- Block, E. (1986). The comprehension strategies of second language readers. *TESOL Quarterly, 20*, 463-494.
- Bolt, D. (2002, April). *Studying the potential of nuisance dimensions using bundle DIF and multidimensional IRT analyses*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Bolt, D., & Stout, W. (1996). Differential Item Functioning: Its multidimensional model and resulting SIBTEST detection procedure. *Behaviormetrika, 23*, 67-95.
- Brown, J. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*, 217-238.
- Burnaby, B., & Sun, Y. (1989). Chinese teachers' views of western language teaching: Context informs paradigms. *TESOL Quarterly, 23*, 219-238.
- Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items*. Newbury Park: Sage.
- Carrell, P. (1983). Some issues in studying the role of schemata, or background knowledge in second language comprehension. *Reading in a Foreign Language, 1*, 81-92.

- Carrell, P. (1989). Metacognitive awareness and second language reading. *Modern Language Journal*, 73, 121-133.
- Chen, H. (1992). Reading comprehension in Chinese: Some implications from character reading times. In H. Chen & O. Tzeng (Eds.), *Language processing in Chinese* (pp. 175-205). Amsterdam: Elsevier.
- Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, 2, 155-163.
- Cohen, A. (1998). Strategies and processes in test taking and SLA. In L. Bachman & A. Cohen, *Interfaces between second language acquisition and language testing research* (pp. 90-111). Cambridge: Cambridge University Press.
- Davis, F. (1968). Research in comprehension in reading. *Reading Research Quarterly*, 3, 499-545.
- Douglas, J., Roussos, L., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement*, 33, 465-484.
- Drum, P., Calfee, R., & Cook, L. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, 16, 486-514.
- Duffy, G., Roehler, L., Sivan, E., Rackcliffe, G., Book, C., Meloth, M., Vavrus, L., Wesselman, R., Putnam, J., & Bassiri, D. (1987). Effects of explaining the reasoning associated with using reading strategies. *Reading Research Quarterly*, 22, 347-368.
- Ercikan, K., Gierl, M., McCreith, T., Puhan, G., & Koh, K. (2002). *Comparability of*

- English and French Versions of SAIP for reading, mathematics and science items.* Paper presented at the annual meeting of the Canadian Society for the Study of Education, Toronto.
- Fender, M. (2003). English word recognition and word integration skills of native Arabic- and Japanese- speaking learners of English as a second language. *Applied Psycholinguistics*, 24, 289-315.
- Gierl, M., Bisanz, G., & Bisanz, J. (2001, July). *Developing an interpretative framework for understanding group differences on national and international achievement tests: The case of excellence in Alberta.* A research proposal submitted to Alberta Learning, Edmonton, Alberta.
- Gierl, M., Bisanz, J., Bisanz, G., Boughton, K., & Khaliq, S. (2001). Illustrating the utility of differential bundle functioning analyses to identify and interpret group differences on achievement tests. *Educational Measurement: Issues and Practice*, 20, 26-36.
- Gierl, M., Rogers, T., & Klinger, D. (1999). *Consistency between statistical procedures and content reviews for identifying translation DIF.* Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Ginther, A., & Stevens, J. (1998). Language background and ethnicity, and the internal construct validity of the Advanced Placement Spanish Language Examination. In A. Kunnan (Ed.), *Validation in language assessment* (pp. 169-194). Mahwah, NJ: Erlbaum.
- Harshbarger, B., Ross, T., Tafoya, S., & Via, J. (1986, March). Dealing with multiple learning styles in the ESL classroom. Symposium presented at the annual

- international meeting of Teachers of English to Speakers of Other Languages, San Francisco.
- Hill, C., & Parry, K. (1989). Autonomous and pragmatic models of literacy: Reading assessment in adult education. *Linguistics and Education, 1*, 233-283.
- Hill, C., & Parry, K. (1992). The test at the gate: Models of literacy in reading assessment. *TESOL Quarterly, 26*, 433-461.
- Holland P., & Thayer, D. (1986). *Differential item performance and the Mantel-Haenszel procedure*. ETS Research Report No. 86-31. Princeton, NJ: Educational Testing Service.
- Hosenfeld, C. (1977). A preliminary investigation of the reading strategies of successful and non-successful language learners. *System, 5*, 110-123.
- Jiang, H., & Stout, W. (1998). Improved Type I Error Control and Reduced Estimation Bias for DIF Detection Using SIBTEST. *Journal of Educational and Behavioural Statistics, 23*, 291-322.
- Joint Advisory Committee. (1993). *Principles for fair student assessment practices for education in Canada*. Retrieved October 9, 2003, from University of Alberta, Centre for Research in Applied Measurement and Evaluation Web site:
<http://www.education.ualberta.ca/educ/psych/crame/research.htm>
- Katz, L., & Frost, R. (1992). Reading in different orthographies: The orthographic depth hypothesis. In R. Frost and L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 67-84). Amsterdam: Elsevier.
- Kharma, N. (1998). EFL and community needs. *International Review of Applied Linguistics in Language Teaching, 36*, 49-69.

- Kim, M. (2001). Detecting DIF across the different language groups in a speaking test. *Language Testing, 18*, 89-114.
- Kohn, J. (1992). Literacy strategies for Chinese university learners. In F. Dubin & N. Kuhlman (Eds.), *Cross-cultural literacy* (pp. 113-125). Englewood Cliffs, NJ: Regents.
- Kunnan, A. (1994). Modelling relationships among some test-taker characteristics and performance on EFL tests: An approach to construct validation. *Language Testing, 11*, 225-252.
- Levine, A., Reves, T., & Leaver, B. (1996). In R. Oxford (Ed.). *Language learning strategies around the world: Cross-cultural perspectives* (pp. 35-46). University of Hawaii at Manoa: Second Language Teaching and Curriculum Center.
- Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748. Holland.
- Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.
- Nandakumar, R. (1993). Simultaneous DIF amplification and cancellation: Shealy-Stout's test for DIF. *Journal of Educational Measurement, 30*, 293-311.
- O'Malley, J., & Chamot, A. (1990). *Learning strategies in second language acquisition*. Cambridge: Cambridge University Press.
- Oxford, R. (1990). *Language learning strategies*. New York: Newbury House.
- Parry, K. (1996). Culture, literacy and L2 reading. *TESOL Quarterly, 30*, 665-692.
- Penner, J. (1995). Change and conflict: Introduction of the communicative approach in China. *TESL Canada Journal, 12* (2), 1-17.

- Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*, 26-56.
- Pritchard, R. (1990). The effects of cultural schemata on reading processing strategies. *Reading Research Quarterly, 25*, 273-295.
- Purpura, J. (1997). An analysis of the relationships between test takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning, 47*, 289-325.
- Reid, J. (1995). *Learning styles in the EFL/ESL classroom*. Boston: Heinle & Heinle.
- Robb, L. (1996). *Reading strategies that work: Teaching your students to become better readers*. New York: Scholastic.
- Roussos, L., & Stout, W. (1996). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355-371.
- Ryan, K., & Bachman, L. (1992). Differential item functioning on two tests of EFL proficiency. *Language Testing, 9*, 12-29.
- Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8*, 95-111.
- Scheuneman, J. (1979). A new method for assessing bias in test items. *Journal of Educational Measurement, 16*, 143-152.
- Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58*, 159-194.
- Shepard, L., Camilli, G., & Averill, M. (1981). Comparison of six procedures for

- detecting test item bias using both internal and external ability criteria. *Journal of Educational Statistics*, 6, 317-375.
- Stout, W., & Roussos, L. (1995). *SIBTEST manual*. University of Illinois: Department of Statistics, Statistical Laboratory for Educational and Psychological Measurement.
- Stout, W., & Roussos, L. (1999). *Dimensionality-based DIF/DBF package* [Computer program]. William Stout Institute for Measurement: University of Illinois.
- Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.
- Thompson-Panos, K., & Thomas-Ružić, M. (1983). The least you should know about Arabic: Implications for the ESL writing instructor. *TESOL Quarterly*, 17, 609-623.
- Weinstein, C., & Mayer, R. (1986). The teaching of learning strategies. In M. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 315-327). New York: Macmillan.
- Willing, K. (1988). *Learning styles in adult migration education*. Adelaide, Australia: National Curriculum Resource Centre.
- Zumbo, B. (1999). *A handbook on the theory and methods of differential item functioning: Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Author Notes

Marilyn Abbott is a PhD candidate studying at the Centre for Research in Applied Measurement and Evaluation, Department of Educational Psychology, 6-110 Education North, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada, T6G 2G5. Email: mabbott@ualberta.ca

Table 1

Descriptive Statistics for Form I Stage II of the CLBA Reading Assessment

Characteristic	Arabic	Mandarin
Number of Examinees	250	250
Number of Items	32	32
Mean	17.20	18.29
Standard Deviation	5.87	6.20
Kurtosis	-.72	-.37
Skewness	.03	-.02
Mean Item Difficulty	.54	.57
SD Item Difficulty	.21	.20
Mean Item Discrimination ^a	.40	.42
SD Item Discrimination	.12	.11
Internal Consistency ^b	.83	.85

^aBiserial Correlation^bCronbach's alpha coefficient

Table 2

Differential Bundle Functioning Results

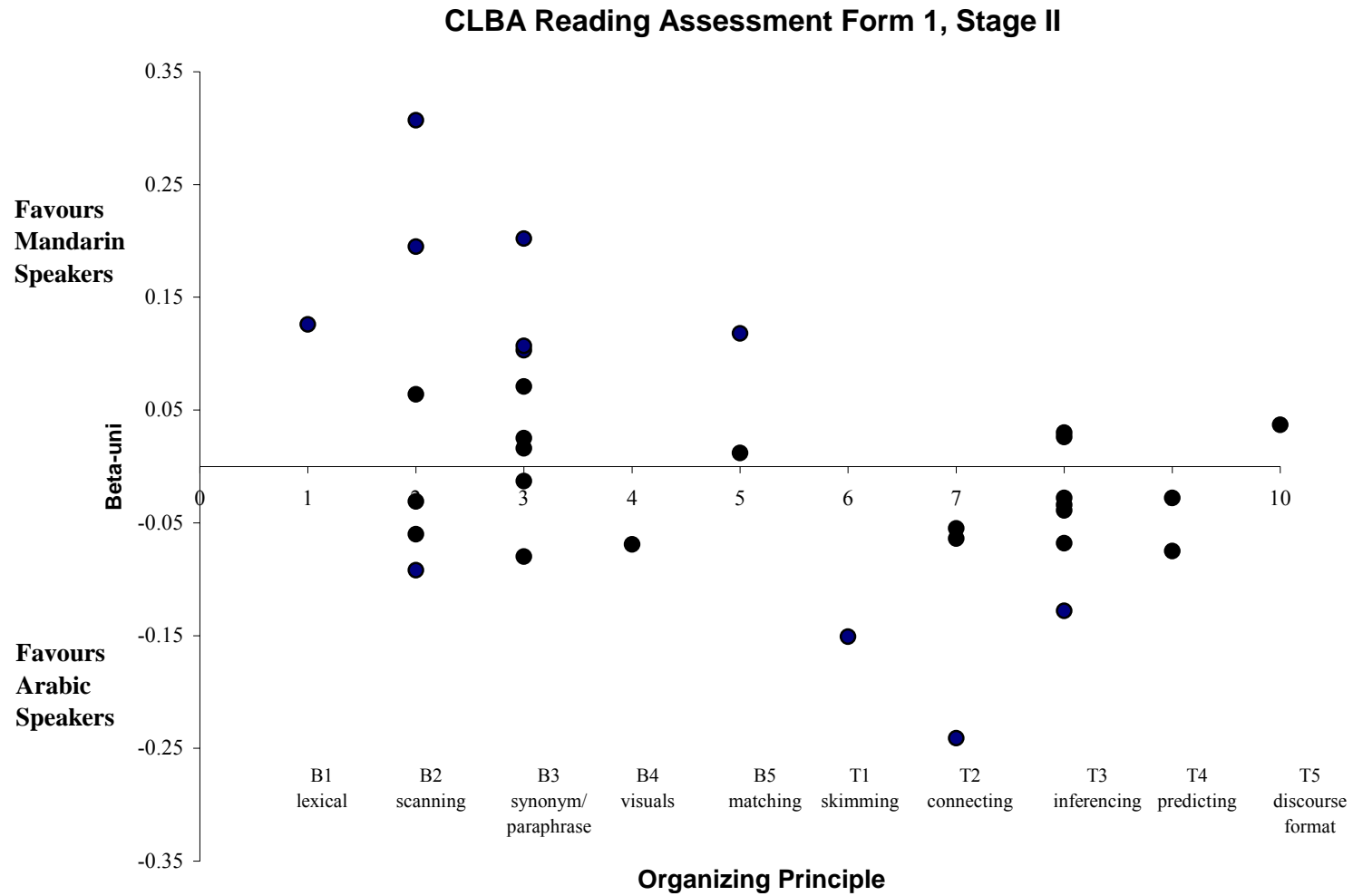
Bundle	No. of Items	No. of items in matching subtest	$\hat{\beta}_{\text{UNI}}$	Favours
Bottom-up				
B1 - Lexical	1	14	0.176*	Mandarin
B5 - Matching key words	2	14	0.192*	Mandarin
Top-down				
T1 - Skimming	1	18	0.137*	Arabic
T2 - Connecting	3	18	0.446*	Arabic
T4 - Predicting	2	18	0.132*	Arabic

* $p < .05$

Figure Captions

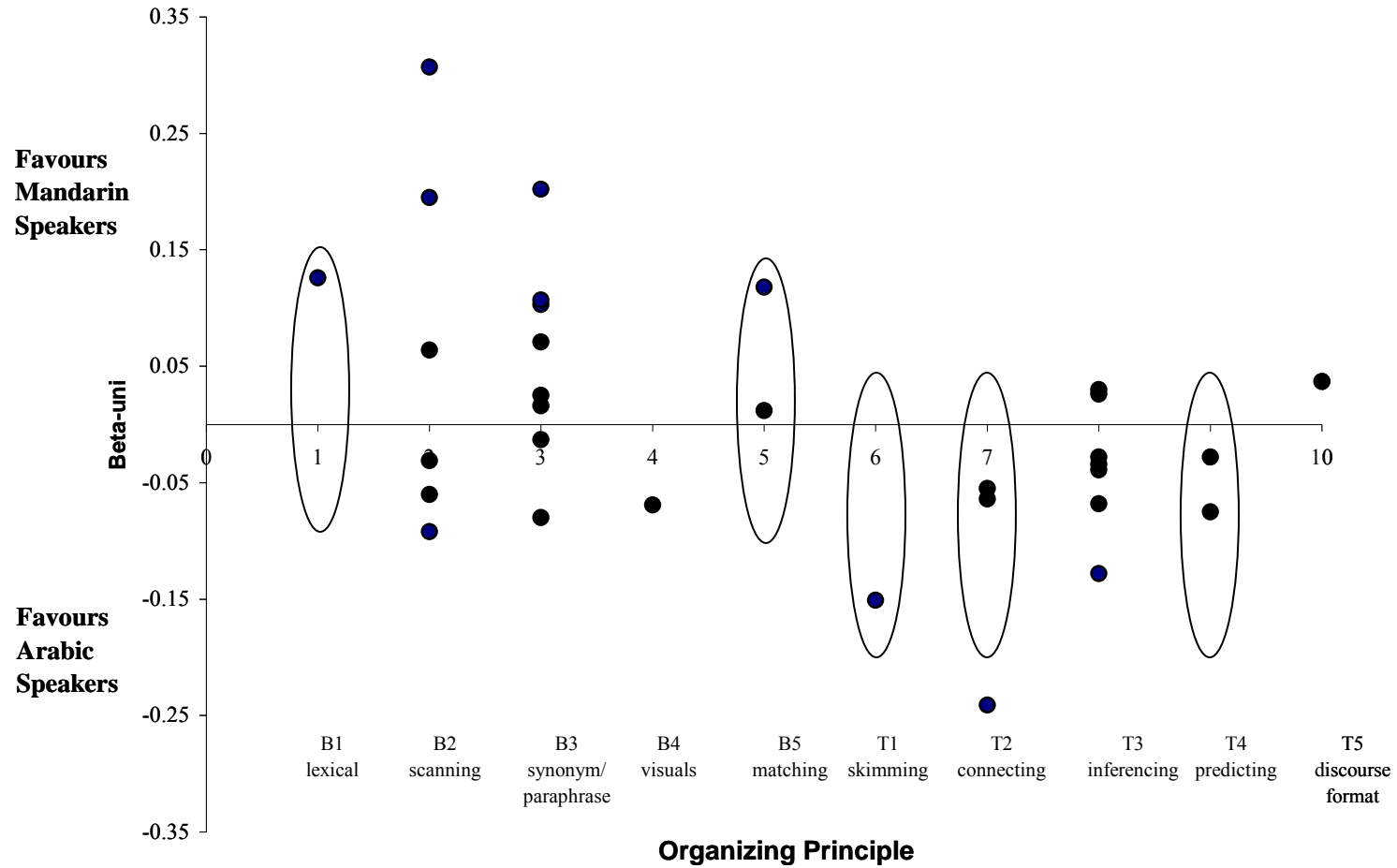
Figure 1. Arabic and Mandarin speaker differences for all items in the CLBA Reading Assessment (Form 1, Stage II) organized into bundles in the ten bottom-up and top-down ESL reading strategy categories (i.e., B1 - breaking lexical items into smaller parts, B2 - scanning for details, B3 - finding synonyms or paraphrases, B4 - matching words to key visuals, B5 - matching key vocabulary in the text to key vocabulary in the item, T1 - skimming for gist, T2- connecting or relating information presented in different parts of the text, T3 - drawing an inference based on information presented in the text, T4 - predicting what may happen in a related scenario/speculating beyond the text, and T5 - recognizing discourse format).

Figure 2. The five interpretable bundles.



Note. B1 through B5 are bottom-up strategy item bundles and T1 through T5 are top-down strategy item bundles.

CLBA Reading Assessment Form 1, Stage II



Note. B1 through B5 are bottom-up strategy item bundles and T1 through T5 are top-down strategy item bundles.