Improving Second Language Proficiency Assessment

A Differential Item Functioning Study

Avi Allalouf

National Institute for Testing and Evaluation

Jerusalem, Israel

Paper presented at the annual meeting of the
American Educational Research Association
San Diego, CA, April 2004

ABSTRACT

This study investigates factors affecting knowledge and acquisition of a second language (SL) by examining differential item functioning (DIF) on SL (Hebrew) test items for two language groups: Arabic and Russian speakers. The results are consistent with the literature on English as a SL with regard to performance in grammar and vocabulary. Many items (42%) functioned differentially, indicating a potential threat to validity. The most problematic item type was Sentence Completion. To reduce the number of DIF items included in operational tests, we suggest changing the balance between item types and performing DIF analysis on piloted items.

Further research, using the age of the examinees and the length of time they have lived in Israel as explanatory variables, is currently underway. The findings are pertinent to the existing debate regarding the attributes of a critical period in language acquisition. In addition, a special "non-DIF" test form was constructed on the basis of the study's results for purposes of validation. This test form will be administered to Russian and Arabic speakers during 2004.

Several factors affect the knowledge and acquisition of a second language (SL). Among them are the nature and structure of the first language, culture, environment, age, method of acquisition and the amount of effort invested. Assessment of SL proficiency should take these factors into consideration. An examination of how different first language groups perform on SL test items can elucidate how these factors come into play. The main purpose of the present study is to suggest improvements in SL proficiency tests on the basis of a better understanding of the relation between first language and SL knowledge. This should result in more equivalent scores (according to the AERA, APA & NCME standards, 1999) for different linguistic and cultural characteristics groups. SL tests can be improved at the item level, and at the test level, by achieving better balance between item types and the content of the test.

The study has two further goals:

1. To make a contribution to scientific understanding of the relationship between first and second language knowledge and acquisition.
2. To participate in the ongoing debate regarding the existence and attributes of a critical period in language (and SL) acquisition.

A major advantage this study enjoys, comparative to previous studies on similar topics, is that it is based on much larger samples (about 33,000 individuals), tested using many (nine) test forms (396 items). Moreover, all previous studies addressed English as a second language, while this study deals with Hebrew as a SL. It is hence both a replication and a basis for comparison.

How first language and other variables affect SL acquisition and knowledge has been the subject of several studies. Some of the research on this topic, the current study included, employs DIF (Differential Item Functioning) analysis. According to the common definition of DIF, an item functions differently across groups if examinees of equal ability but from different groups (here, different first language groups) do not have an equal probability of responding correctly to that item.

Some previous studies analyzed the DIF of SL items administered to various first language groups. Investigation of DIF in tests in general -- and especially in language

proficiency tests, to quote Kim (2001) -- is crucial, because DIF items pose a considerable threat to validity.

One group of studies (among them Alderman & Holland, 1981, Chen & Hening, 1985; Sasaki, 1991) examined tests of English as a second language (such as TOEFL and ESLPE). The general conclusion was that DIF items were related to cognates, meaning that the similarity between languages affected test performance. Sasaki (1991) for example, investigated the ESLPE with 262 Chinese and 81 Spanish native speakers and found that vocabulary items containing English-Spanish cognates showed DIF in favor of the Spanish-speaking group.

A second group of studies examined DIF between English and other first language groups on verbal aptitude tests, which usually require a higher level of English than tests of English as a second language. One of the studies, Schmitt & Dorans (1988), analyzed DIF in the SAT. Findings showed that the main item characteristics related to DIF between different first language groups are: content (usually in sentence completion and reading comprehension item types) and homographs (words that are spelled alike but have different meanings), which are differentially harder for different groups.

A third group of studies examined various critical period theories, all of them related to the (original) hypothesis that individuals above a certain age are less capable of learning a language than younger individuals. There is, however, no agreement regarding the characteristics of the critical period. As Dekyser (2000) puts it, "…the concept of a critical period for SL acquisition continues to be a controversial topic" (p. 500). This controversy is evident in the following example:

(a) Krashen, Long & Scarcella (1979) found that the critical factor in SL acquisition is the length of time spent in the SL country, rather than the age of arrival.

(b) Dekeyser (2000) supports the "original" critical period hypothesis, finding that SL learning ability gradually deteriorates from ages 6 to 17. According to Dekeyser, the decline is particularly evident in the ability to apprehend grammar.

Method

Examinees

The examinees were candidates for higher education studies in Israel, either Arabic speakers (born in Israel) or Russian speakers (most of whom immigrated to Israel between the ages of twelve and twenty) who took the Hebrew Proficiency Test (HPT) one to six years after immigrating.

Tests and items

The HPT is primarily administered to candidates for Israeli universities whose first language is not Hebrew. Most Israeli educational institutions use the HPT scores to place students that have been accepted in the appropriate Hebrew language courses. The HPT consists of two parts. The first part has two sections of 22 multiple choice items each. The (primary) item types are: Restatements (RS), Reading Comprehension (RC) and Sentence Completion (SC). For the purposes of this study, all the items also underwent a secondary classification, which is elaborated upon below.

Secondary Classification of SL Item Types

Sentence Completion
1. Prepositions **2**. Conjunctions **3**. Vocabulary **4**. Verbs **5**. Syntax
Restatements
1. Idioms, **2**. Vocabulary **3**. Syntax
Reading Comprehension
1. Lexical-semantic **2**. Text comprehension: local **3**. Text comprehension: global

---

In the second part of the test, examinees are asked to write a short essay on an assigned topic. Nine forms, comprising 396 (44 X 9) items, were analyzed.

DIF Method

The study employed the Mantel-Haenszel (MH) DIF detection method (see Holland & Thayer, 1988). This commonly used method (Holland & Thayer, 1988) applies a matching criterion to determine whether reference and focal group item performance is equal at various levels along the ability continuum. The MH procedure does not require

large samples, employs a χ^2 index for testing statistical significance and has simple classification rules. It provides an MH D-DIF index given in the Delta Metric. The MH D-DIF for a specific item and for the same examinee ability level represents the difference in difficulty between reference and focal groups in terms of the delta metric. The DICHODIF computer program (Rogers, Swaminathan & Hambleton, 1993) was used. DIF classification rules used in this study were based on the DIF classification rules of the Educational Testing Service (Dorans & Holland, 1993). Two categories were defined: (1) Large – C (an absolute MH D- DIF value of at least 1.5); and (2) Moderate – B (an absolute MH D- DIF value of at least 1.0). In order to refine the matching criteria (purification), the DIF detection process involved two stages. The first stage used the total raw score as the stratifying variable. The second stage used the score of the items that did not display large DIF (during the first stage) as the stratifying variable. Items were classified on the basis of second stage results. In this study, a statistically significant difference in performance (at the 0.05 level) between the reference and focal groups was found for the two categories of DIF.

It should be noted that when the criterion is raw score, the results of DIF analysis are symmetrical in nature; if one group has an advantage in one of the item types, the second group has an advantage in the other item type.

Two hypotheses were formulated with regard to DIF results:

1. Arabic speakers will perform better on vocabulary items (because of the similarity between Arabic and Hebrew, in some cases, and also because they have been exposed to Hebrew for longer), and in grammar, in accordance with Dekeyser (2000).
2. Based on the symmetrical nature of DIF results, Russian speakers will perform better on the other items, especially those in which vocabulary and grammar are not critical (such as RC items).

Results

Summary statistics for the nine test forms used in this study are presented in Table 1.

Table 1

Descriptive Statistics for Arabic and Russian Speakers on Nine Test Forms

Version	Sample Size		Score Mean[1]		Standard Deviation[1]	
	Arabic Speakers	Russian Speakers	Arabic Speakers	Russian Speakers	Arabic Speakers	Russian Speakers
9	1567	1258	18.58	18.42	5.86	7.40
10	2097	1402	21.74	21.01	6.28	7.58
12	4150	984	18.76	15.40	4.82	5.71
15	3755	1519	17.93	12.95	5.63	5.12
17	1136	1342	17.94	12.47	5.61	5.12
26	1748	1659	19.64	19.06	6.57	8.08
31	1919	2065	18.77	19.30	5.39	6.35
33	1451	1382	21.75	19.07	6.30	8.05
34	1931	2308	18.76	18.46	5.17	6.66

[1] Raw scores of non-large DIF items (which served as the stratified criteria for the DIF analysis).

Sample sizes were large, close to 2200 Arabic speaking examinees and over 1500 Russian speaking examinees. With the exception of one test form, the Arabic speakers performed better. With the exception of two test forms, the Russian speakers were more heterogeneous. On most forms, the ability differences were small, making DIF detection more accurate (The large score gap on Form 17 was checked. It was found that most of the Russian speakers that took this form had only spent a year in Israel before taking the exam, resulting in a very low level of Hebrew at the time they took the Hebrew Proficiency Test).

Table 2 presents DIF by test form. The number of DIF items in a test form ranges from 15 items to 22 items (out of 44 items in each test form). In total, 167 items, which amount to about 42% of the all the items included in the analysis, displayed DIF. This is large amount of DIF, meaning that many items functioned differently across the two groups.

Table 2

Number of DIF items by Test Form

Test Form	Favoring		Total DIF Items
	Arabic Speakers	Russian Speakers	
9	10	9	19
10	6	9	15
12	10	9	19
15	12	10	22
17	5	10	15
26	9	11	20
31	11	9	20
33	9	8	17
34	11	9	20
Total	**83**	**84**	**167**

Table 3 presents the percentage of items showing DIF according to the primary item type: Sentence Completion, Restatements and Reading Comprehension.

Table 3

Percentage of Items Showing DIF, According to Item Type

Favoring:	**Total**	**Sentence Completion**	**Restatements**	**Reading Comprehension**
Arabic Speakers	21	39	19	6
Russian Speakers	21	19	26	20
Total	42	59	42	25

The findings show that Arabic speakers perform much better on SC items while Russian speakers do better on RC items. No group outperformed the other in the RS items. Comparison of the item types shows that SC items have more DIF than RS items while RC items have the least DIF.

Tables 4.1 – 4.3 present DIF results according the secondary classification within each item type.

Table 4.1 – **Sentence Completion**

Number of DIF Items Favoring Each Group, by Secondary Classification

	Total	Verb	Syntax	Conjunctions	Prepositions	Vocabulary
Total	93	12	19	13	16	33
Favoring Arabic	60	8	11	6	11	24
Favoring Russian	33	4	8	7	5	9

Table 4.2 – Restatements

Number of DIF Items Favoring Each Group, by Secondary Classification

	Total	Syntax	Idioms	Vocabulary
Total	45	12	9	24
Favoring Arabic	19	6	5	8
Favoring Russian	26	6	4	16

Table 4.3 – Reading Comprehension

Number of DIF Items Favoring Each Group, by Secondary Classification

		Lexical-	Text Comprehension	
	Total	Semantic	Local	Global
Total	42	8	19	15
Favoring Arabic	9	5	2	2
Favoring Russian	33	3	17	13

The most noteworthy finding is as follows: Arabic speakers perform better on verbs, prepositions and vocabulary, while Russian speakers perform better on text comprehension: local & global, and on vocabulary in restatements. An effective way to understand the sources of DIF is to analyze the very-high DIF items.

Table 5 presents the 28 items in which the MH-D-DIF value is higher than |3| -- meaning that the difference in the proportion of Arabic and Russian speakers (at the same Hebrew level) who answered the item correctly is at least thirty percent[1]. This large difference implies that item performance is highly affected by the first language.

[1] The cutoff between DIF item and no-DIF item is MH D-DIF=1.00. One delta is equal to ¼ standard deviation of the delta metric difficulty distribution, which is about 10 percent proportion correct (in the middle of the difficulty range).

Table 5

Items with MH-D-DIF Greater than |3|

No	MH-D-DIF	Favoring	Form	Item	Item Type
1	5.19	AR	12	43	SC
2	5.13	RU	9	41	SC
3	5.01	AR	33	44	SC
4	4.78	RU	34	26	SC
5	4.52	AR	15	23	SC
6	4.30	AR	17	23	SC
7	4.26	AR	34	42	SC
8	4.23	AR	12	19	SC
9	4.20	RU	9	20	SC
10	4.18	RU	12	1	SC
11	4.11	AR	34	23	SC
12	4.11	AR	17	25	SC
13	3.69	AR	12	33	RS
14	3.67	RU	26	10	RS
15	3.67	AR	26	20	SC
16	3.56	AR	17	29	RC
17	3.37	RU	34	3	SC
18	3.35	AR	33	19	SC
19	3.33	AR	33	33	RS
20	3.31	AR	10	12	SC
21	3.20	AR	12	27	SC
22	3.16	RU	17	42	SC
23	3.13	AR	26	44	RS
24	3.11	AR	26	32	SC
25	3.08	RU	15	41	SC
26	3.08	AR	17	4	SC
27	3.07	AR	31	35	SC
28	3.02	AR	9	13	SC

An interesting finding is that 20 of the 28 items favor Arabic speakers. Here, there is no symmetry in the extreme DIF items, and there are more items that the Arabic speakers have much higher probability of answering correctly. 23 of the 28 items with "very large DIF" are sentence completions, yet another indication of the problematic nature of this item-type.

Table 6 presents the secondary classification for the 23 "very-large DIF" SC items. It should be noted that some of the items have two secondary classifications. The Arabic speakers performed better on syntax and prepositions items.

Table 6 – Sentence Completion – 23 "Very Large DIF" Items
Number of DIF Items Favoring Each Group, by Secondary Classification

	Verb	Syntax	Conjunctions	Prepositions	Vocabulary
Total	5	8	1	6	5
Favoring Arabic	2	6	1	5	2
Favoring Russian	3	2	0	1	3

Additional analysis regarding the Russian speakers addresses the relationship between time spent learning the language (here, based on the time passed since the immigrant came to Israel) and knowledge of the second language. This relationship was estimated through the correlations between the Hebrew Proficiency Test (HPT) score and two other scores: score on the verbal section of the Psychometric Entrance Score (PET)[2] which was administered in Russian, and score on the English (as a foreign language) section pf the PET. Table 7 presents the correlations, by the estimated time spent learning Hebrew.

Table 7
Correlations Between Hebrew HPT and Verbal-Russian & English Sections of PET for the Russian Speakers, by Estimated Time Spent Learning Hebrew

Time	Verbal in Russian	English
Less than 1 year	.38	.48
Between 1 and 5 years	.44	.57
Between 5 and 15 years	.47	.64

The results show that the correlations increase, in both columns, as the time spent learning Hebrew increases. This demonstrates that time is crucial for achieving proficiency in language. It is interesting that the correlations between Hebrew and English are higher than the correlations between Hebrew and Russian, implying that the relationship between second and third language is stronger than the relationship between first and second language.

[2] The *Psychometric Entrance Test* (PET) is a high-stakes test used for admission to universities in Israel – it is translated from Hebrew into Arabic, Russian, English, Spanish and French.

Discussion

The results of the study in relation to the objectives outlined at the start are as follows:

1*. To suggest improvements in SL proficiency tests, making them more equivalent for different linguistic and cultural groups.* The study results showed that many of the items in the HPT test have DIF between different mother tongue groups. Since equivalency means (among other things) reducing the incidence of DIF, we suggest two ways of accomplishing this in operational tests: (a) piloting new items and performing DIF analysis on the piloted items in order to avoid including DIF items in operational tests, and (b) changing the balance between item types, in accordance with the other test specifications (lowering the number of SC items and increasing the number of RC items). Furthermore, the relationship between the secondary classification and DIF should be taken into consideration.

2*. To make a contribution to scientific understanding of the relationship between first and SL knowledge and acquisition* -- The study results are consistent with previous studies and with the hypothesis: the Arabic speakers have an advantage on grammar and vocabulary items, and the Russian speakers have an advantage on items where vocabulary is not crucial for understanding, such as Reading Comprehension items. It is interesting to cite Angoff & Cook (1988) who, based on a study on DIF in test translation, state that: "items with more context probably tend to retain their meaning, even in the face of translation into another language" (page 8). Here, although no translation is taking place, the situation is similar. The Russian speakers have less vocabulary and this is detrimental to their performance on short items. However, when the item is based on a wider context and lack of familiarity with some words is not critical, this disadvantage is overcome.

3*. To contribute to the debate regarding the critical period hypothesis* -- Indeed, as hypothesized, the Arabic speakers were better at grammar than the Russian speakers, who started learning their SL (Hebrew) between the ages of twelve and twenty and took the HPT after spending 1 to 6 years in Israel.

An additional source of information that is pertinent to this discussion is Abramzon (2002), which summarizes HPT statistics for 2000-2001. Abramzon reports that:

1. The correlation between HPT score (first part) and verbal ability score in first languages is .54 for Arabic speakers and .40 for Russian speakers. As Khuwaileh & Shoumali (2000) noted, verbal ability in first and second language are related. The correlation is expected to be smaller for the Russian speakers who started learning the SL late and spent less time doing so.
2. The Russian speakers' scores on the closed part of the HPT go up in proportion with the amount of time they have spent in Israel. During the years 2000-2001, on a scale of [100,20] for a base population, every year in Israel is "worth" about 5 points on the closed section, and about 4 points on the open section. This shows that time is important for second language acquisition, and that time is probably more critical in learning to read than it is in learning to express oneself in writing.

Another conclusion is that since many items showed DIF, SL proficiency for different first language groups may be assessed using different tests: a special test form for each group, rather than a single form for all. It seems that a single form cannot assess proficiency when there is large variation in the nature of language ability between the groups.

A special "non-DIF" test section was constructed on the basis of the study results and will be administered to Russian and Arabic speakers for purposes of validating the study results during 2004. This section includes 22 items (6 Restatements, 8 Reading Comprehension and 6 Sentence Completion) like every common HPT test section. None of thel 22 items displayed DIF and almost all of them belonged to the "upper quarter" of each test form, based on descending MH D-DIF value. This "ideal" test section will be analyzed for DIF and other important characteristics will be estimated (validity, reliability).

Further research is in progress to obtain more information regarding the study goals that were formulated in the introduction. The 28 items that exhibited very high DIF will be analyzed by Arabic-Hebrew and Russian-Hebrew bilinguals, using the same approach as Allalouf, Hambleton & Sireci (1999). The purpose of this analysis is to explain and understand the sources of DIF. Further analysis will be performed using the age of examinees and the time they have spent in Israel as explanatory variables.

Notes

The author would like to thank Oshrit Binhas and Andrea Abramzon for their assistance throughout the study.

References

Abramzon, A. (2002). *Hebrew proficiency tests, 2000-2001.* Research Report No. 293. Jerusalem: NITE.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Alderman, D. L. & Holland, P. W. (1981) *Item performance across native language groups on the test of English as a foreign language*. Report 448, ETS.

Allalouf, A., Hambleton, R. K., Sireci, S.G. (1999) Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36,* 3, 185-198.

Angoff, W. H., & Cook, L. L. (1988). *Equating the scores of the Prueba de Aptitud Academica and the Scholastic Aptitude Test* (College Board Report No. 88-2). New York: College Entrance Examination Board.

Chen, Z & Henning, G, (1985). Linguistic and cultural bias in language proficiency test. *Language Testing, 2,* 155-163.

Dekeyser, R. M. (2000). The robustness of critical period effects in second Language acquisition. *Studies in Second Language Acquisition 22*, 499-533. (2000).

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning*. Hillsdale, NJ: Erlbaum.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. Braun (Eds.) *Test validity*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Khuwaileh, A. A., & Shoumali, A. A. (2000). Writing errors: A study of the writing ability of Arab learners of academic English and Arabic at university. *Language, Culture and Curriculum, 12,* 174-183

Kim, M. (2001). Detecting DIF across the different language groups in a speaking test, *Language Testing, 18,* 89-114

Krashen, S. D., Long, M. A., & Scarcella, R. C. (1979). Age, rate and eventual attainment in second language acquisition, *TESOL Quarterly, 13,* 573-582

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing, 8,* 95-111

Schmitt & Dorans (1990). Differential item functioning for minority examinees on the SAT. *Journal of Educational Measurement, 27,* 67-81.

Rogers, H. J., Swaminathan, H., & Hambleton, R. K. (1993). *DICHODIF : A FORTRAN program for DIF analysis of dichotomously scored item response data* [A computer program]. Amherst, MA: University of Massachusetts.