Experimental Studies Utilizing Native-Language Options on Language Tests

Harold S. Madsen

Sheila Maluf

Follow this and additional works at: https://scholarsarchive.byu.edu/dlls

BYU ScholarsArchive Citation
Madsen, Harold S. and Maluf, Sheila (1979) "Experimental Studies Utilizing Native-Language Options on Language Tests," Deseret Language and Linguistic Society Symposium: Vol. 5 : Iss. 1 , Article 15.
Available at: https://scholarsarchive.byu.edu/dlls/vol5/iss1/15

This Article is brought to you for free and open access by the All Journals at BYU ScholarsArchive. It has been accepted for inclusion in Deseret Language and Linguistic Society Symposium by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
EXPERIMENTAL STUDIES UTILIZING NATIVE-LANGUAGE OPTIONS ON LANGUAGE TESTS

Harold S. Madsen
Sheila Maluf

The purpose of the studies reported in this paper is to determine whether or not the use of native-language test questions can provide a more satisfactory measurement of low-proficiency language learners than presently available examinations.

The problem. Instructional accountability in contemporary ESL and modern language programs requires periodic evaluation of students in training. Beyond the traditional testing concerns of validity, reliability, and efficiency, new concerns are now being voiced, such as test bias (Ollier and Perkins, 1978), the relationship between test form and modality (Clifford, 1979) and test affect (Stevenson, 1979; Shohamy, 1979). Yet another matter is the need for accurate and sensitive measurement at low-profi ciency levels (Maluf, 1979).

In connection with commercial tests from the "Ilyin Oral Interview" to the "Michigan Test of English Language Proficiency" or the prestigious TOEFL (Test of English as a Foreign Language), we find caveats enjoining the user to exercise caution in interpreting results when the test is administered to low-proficiency students. Sometimes a screening process is recommended in order to prevent beginning-level students from taking the exam.

An oft-expressed caution to those planning exams for less-advanced students is to avoid the situation somewhat prevalent in listening and reading tests where the student understands the passage but has difficulty with questions based on what he has heard or read (Valette, 1977, p. 104; Clark, 1972, p. 60; Harris, 1969, p. 62; Heaton, 1975, p. 76; Porter, 1976, pp. 152-153). One solution to this problem is to provide picture cues or "pure" responses (Clark, 1972, p. 28), that obviate the need for the student to cope with written alternatives. Pictures present limitations, however, due not only to their cost but also to possible ambiguities and the difficulty of representing complex or abstract ideas (Harris, 1969, pp. 38-39).

Background. A more practical solution to the problem of how to help low-proficiency students cope with questions on the foreign language is to permit the use of native language answers. For example, in connection with listening comprehension items, Valette recommends that beginning students give open-ended responses in the native language. She also suggests that for beginning students, passage comprehension questions might be presented in the native language (Valette, 1977, pp. 102, 106). And Clark cautiously recommends printed NL multiple-choice options in general achievement listening tests, for students with limited training in reading indicating positive results in at least one such effort (Clark, 1972, p. 60; see also p. 39 and Clark, 1975, p. 57).
In Holland, 240 secondary-school students participated in a testing experiment involving the use of native-language cues. Half received all-English language examinations, and half (equally proficient) took the same examination with NL multiple-choice questions. The FL mean was 77 percent; the NL mean was 82 percent. Similar results were obtained on French and German examinations (Groot, 1975a, p. 53). However, there was a preference on the part of these rather advanced students for the all-FL form.

In view of these modest studies and the overtures in favor of native-language cues, it seemed appropriate to examine experimentally their impact on students with low target-language proficiency.

Study I: Arabic Speakers

The need in Egypt to evaluate annually the English proficiency of two million secondary-school students has stimulated interest in the development of efficient, integrative objective tests. And the concern for appropriate assessment of students with relatively little proficiency in English led to an investigation of new testing strategies, such as the use of native-language cues. In a study conducted by Madsen and Iskander in Cairo, Egypt, it was hypothesized that low-proficiency students would perform higher on an EFL proficiency test with NL options than on the same test with FL options. Even though the NL options would appear in Colloquial Arabic (a conversational form, rare in print), it was assumed that such a form would provide a more sensitive measure of English proficiency than a test with English options, which examinees only half understood. But it was also hypothesized that this advantage would not hold for intermediate and advanced students. Experienced Egyptian teachers of English agreed that the effort required to decipher the printed representation of spoken Arabic would cancel out the benefit of responding to answers in the mother tongue.

The instrument. For legal and security reasons, it was not feasible to prepare Arabic (NL) options for available commercial EFL tests. Instead, a locally normed EFL proficiency test was selected: the Alternate Modality Listening Exam (AMLEX). Administered experimentally to American University in Cairo applicants, the AMLEX correlated in the .80's with the Michigan Test of Aural Comprehension (MTAC) and generally at this same level with the Michigan Test of English Language Proficiency (MTELP) (See Madsen, 1978, p. 341).

Table 1

| CORRELATIONS BETWEEN THE AMLEX AND TWO COMMERCIAL EFL TESTS |
|----------------|----------------|----------------|
|                | MTAC | MTELP |
| Graduate Applicants (N = 72) | .86  | .88  |
| Management Applicants (N = 73) | .82  | .79  |
The AMLEX consists of two sections of conversational utterances. In Section One, there are 45 questions requiring appropriate response:

**FL EXAMPLE:** How far is it to Helwan?
A. No, not far.
B. South of Cairo.
C. About 20 kilometers.

**EXAMPLE WITH NL OPTIONS**

何時到的埃及？
A. 南部
B. 南部
C. 南部

Section Two consists of 45 statements requiring selection of appropriate paraphrase:

**FL EXAMPLE:** They work all but three months of the year.
A. They work nine months.
B. They work only three months.
C. They work every three months.

**EXAMPLES WITH NL OPTIONS**

他們想去伊斯坦布爾，但去了蘇伊士。
A. 航海
B. 航海
C. 航海

Both the cues and the options are printed. Examinees are given only 25 minutes to complete the test. Preparation, piloting, and revision of the Arabic form were carried out by Mona Iskander, an Egyptian EFL specialist. Since the AMLEX incorporates informal responses characteristic of the spoken language, it was necessary to utilize Colloquial Arabic for the form with NL options. This presented a challenge for examiner and examinee since in Arabic-speaking countries conversational language rarely appears in writing.*

*Colloquial Arabic, or a slightly classicized version thereof, is occasionally seen in cartoons or in newspaper reports of debates in the People's Assembly, but it is still difficult to read with speed and fluency because of the lack of a close sound-spelling correspondence and because of readers' attempts to impose classical word boundaries on colloquial utterances.
Subjects. Twelve groups of Egyptian students were selected, at three levels of English proficiency. Beginning and intermediate students (levels 1 and 2) were drawn from the Division of Public Services adult English program, sponsored by the American University in Cairo. Advanced-level students (level 3) were drawn from English Department courses at AUC. Level 1 examinees consisted of 43 students in two DPS classes; placed by locally-constructed tests, these students had a proficiency of 30 to 45 on the MTELP. Level 2 examinees consisted of 79 students in six DPS classes; their proficiency ranged from 50 to 70 on the MTELP. Level 3 examinees consisted of 51 students in six English Department classes, with scores above 80 on the MTELP. Completing the study were 40 persons at level 1, 76 at level 2, and 48 at level 3, for a total of 164 subjects.

Method. It was decided that subjects would constitute their own control, by taking both forms of the test. To counter practice effect, half of the students at each level took the NL (all-English) form first followed (the same day) by the form with Arabic distractors. The other half took the tests in the reverse sequence. A two-way analysis of variance was utilized to evaluate the effect of test form, sequence, level and interaction.

Results. As anticipated, sequence was statistically significant. When either test form was administered second, scores were higher than on the previous exam. Predictably, too, there were significant differences by level on both forms, level 1 scoring in the 40's, level 2 in the 60's and level 3 in the 70's. Level by sequence interaction was also significant, with the most dramatic gains being made at level 1, between the first and second administration of the test. But overall difference between performance on the all-English test and that with Arabic distractors was non-significant. And the test for interaction between test form and ability level was non-significant.

Discussion. The principal hypothesis was not borne out by the analysis of variance. Instead of the Arabic-answer form being easier for beginning-level students, the all-English form actually produced a slightly higher mean, though this difference was non-significant. As expected, the form with Arabic distractors did not constitute an advantage at advanced levels. In short, at no level did native-language answers in Colloquial Arabic constitute the advantage reported in the literature for speakers of European languages. It appears that the effort required to read the unfamiliar colloquial script cancels out the native language advantage.
Table 2

ANALYSIS OF VARIANCE: Arabic Speakers

Dependent Variable: Difference Between English and Arabic Forms

\[ \text{Diff.} = \bar{\chi} + L + S + LS + E \]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>2</td>
<td>278.60</td>
<td>139.3</td>
<td>1.1154</td>
<td>.3303 NS</td>
</tr>
<tr>
<td>Sequence</td>
<td>1</td>
<td>11675.</td>
<td>11675.</td>
<td>93.482</td>
<td>&lt; .0001 Sig.</td>
</tr>
<tr>
<td>Level x Sequence</td>
<td>2</td>
<td>2952.3</td>
<td>1476.1</td>
<td>11.820</td>
<td>&lt; .0001 Sig.</td>
</tr>
<tr>
<td>Error</td>
<td>158</td>
<td>19732.</td>
<td>124.89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Diff = Difference (English - Arabic scores)
\( \bar{\chi} = \text{Mean of population} \)
L = Level (1, 2, 3)
S = Sequence (1, 2)
LS = Level x Sequence interaction
E = Error

*It should be noted that since the dependent variable is "difference" (English minus Arabic), when one tests level, this constitutes level by language interaction.

Table 3

COMPARISON OF MEANS BY LEVEL: Arabic Speakers

<table>
<thead>
<tr>
<th>Level</th>
<th>All-English form</th>
<th>Arabic Distractor form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>45.025</td>
<td>43.925</td>
</tr>
<tr>
<td>Level 2</td>
<td>64.776</td>
<td>61.789</td>
</tr>
<tr>
<td>Level 3</td>
<td>77.500</td>
<td>77.500</td>
</tr>
</tbody>
</table>
Study II: Portuguese Speakers

The problem of the unfamiliar written Colloquial Arabic could be countered by providing oral multiple-choice options, but this would be costly and inefficient; moreover, it would introduce a heavy memory-load factor. It was decided, instead, to eliminate the problem by utilizing speakers of a European language. The follow-up study replicated the original, with a few basic modifications: To enhance the need for native-language props, examiners utilized oral stimuli on the test. And because of the erratic practice effect found on the second administration of either form of the test, it was decided to analyze only a single administration of the FL test and just a single administration of the test with NL distractors. A pre-test was used for grouping students by level. Also a questionnaire was introduced to assess affect.

The instruments. Again the AMLEX was used. But this time the stem was recorded on tape. Only the multiple-choice answers were printed. Instructions were both taped and printed—in the students' native language.

For selection purposes, J. Donald Bowen's Integrative Grammar Test (IGT) was employed. An oral modality test, the IGT measures language proficiency without sensitizing students to the AMLEX. The test has been validated but does not require the security of available commercial tests; and it can be quickly administered (19 minutes) and easily scored. The IGT consists of 50 sentences containing reductions, assimilations, and contractions; the 50 items are presented a second time on the test but in a new sequence. From contextual clues, students are required to identify the second word of the sentence (often distorted or obscured). The IGT correlates at .817 with placement tests consisting of the MTELP, MTAC, and a written composition.

Also added to this study was a questionnaire which generated demographic information. Included, also, were two questions on test affect: one to determine which test was preferred and another to determine which test was perceived as being easier.

Subjects. Forty Brazilians, 18 to 30 years of age (mean age - 23.0) were involved in this study. There were 18 males and 22 females. All were native Portuguese speakers, currently residing in the United States. Their exposure to English ranged from a few months to 13 years, the average being 2.5 years. All but two were students; some were attending high school, others college, and some graduate school. In the low group there were 9 males and 11 females, the mean age being 23.1. In the high group there were also 9 males and 11 females, the mean age being 22.8.

Method. Subjects were first administered the IGT pre-test, individually. On the basis of this test, they were divided into two equal groups. By random selection, half of each group received the all-English (FL) test; the other half received the form with NL cues. To facilitate evaluation of affect, each group was also tested on the alternate form of the test, a week later. Like the pre-test, both forms of AMLEX were individually administered. The questionnaire was administered after the second test
administration. Again, the independent variable was test form, the moderator variable was level of proficiency, and the dependent variable was performance on the AMLEX. Statistical analysis consisted of a two-way analysis of variance; and affect was evaluated by chi square.

Results. As expected, there were significant differences among the three levels. The overall differences by test type did not reach significance, however. And the interaction of achievement level and test type likewise failed to reach significance. The low group of Brazilians performed better on the form with NL distractors, though this did not reach significance. The upper groups did better on the all-FL test, with a difference significant at the .05 level.

Table 4

ANALYSIS OF VARIANCE: Portuguese Speakers

\[ Y = A + T + AT + E \]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>2</td>
<td>4310.4</td>
<td>2155.2</td>
<td>41.492</td>
<td>&lt; .0001 (Sig.)</td>
</tr>
<tr>
<td>Test</td>
<td>1</td>
<td>14.074</td>
<td>14.074</td>
<td>0.27095</td>
<td>NS</td>
</tr>
<tr>
<td>Level x Test</td>
<td>2</td>
<td>136.41</td>
<td>68.204</td>
<td>1.3130</td>
<td>NS</td>
</tr>
<tr>
<td>Error</td>
<td>34</td>
<td>1766.1</td>
<td>51.944</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A = Achievement level  
T = Test form  
AT = Interaction of level and test  
E = Error

Table 5

COMPARISON OF MEANS BY LEVEL: Portuguese Speakers

<table>
<thead>
<tr>
<th>Level</th>
<th>All-English form</th>
<th>Portuguese Distractor form</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>49.857</td>
<td>53.833</td>
<td>NS</td>
</tr>
<tr>
<td>Level 2</td>
<td>71.143</td>
<td>68.625</td>
<td>NS</td>
</tr>
<tr>
<td>Level 3</td>
<td>79.857</td>
<td>74.800</td>
<td>.05 (Sig.)</td>
</tr>
<tr>
<td>Total</td>
<td>66.952</td>
<td>65.753</td>
<td>NS</td>
</tr>
</tbody>
</table>
The majority at all levels preferred the FL form. Dividing the Portuguese speakers into a high and a low group, we see a three to one preference for the all-foreign language test by advanced students. This approaches significance. A similar relationship holds for examinees' perception of exam difficulty; but in this category, differences for advanced students are significant.

Discussion. Unlike the Arabic groups, Portuguese speakers performed nearly as expected: the least proficient scored higher on the form with native-language cues (though not significantly higher)—an average of four points, instead of the five points reported in Groot's study. And this time the most advanced group did significantly better on the all-English (FL) form. This suggests that highly proficient language learners have essentially ceased to utilize the native language when communicating in the target language and that NL cues actually constitute something of a handicap for them. This is reflected in both their preference for the all FL form as well as their perception of its being easier. The fact that an even more dramatic difference among the groups did not occur seems to reflect the fact that most of the Brazilian examinees are fairly advanced in language skills and are used to functioning daily in the target language here in an English-speaking country.
Study III: English Speakers

To evaluate the impact of NL cues on language learners at near beginning level, researchers next evaluated approximately 100 Americans who were studying Portuguese, half of whom had had only minimal exposure to the target language. As with the Brazilian group, a listening test was utilized; and again, only a single administration of the test was evaluated. The IGT pre-test was eliminated.

The instruments. Once more, an oral-cued form of the AMLEX was administered, with printed multiple-choice options. The instructions on both forms were in the native language, English. But on one form all questions and distractors were in Portuguese; on the second form, the questions were in English and the distractors in English. An English version of the questionnaire administered to the Brazilians was prepared. And for the beginning group, an additional question was asked: whether or not the NL distractors were helpful, confusing or neither.

Subjects. The low group consisted of 54 male, native English speakers at the Missionary Training Center (MTC) in Provo, Utah, where participants were receiving intensive Portuguese instruction. They ranged in age from 19 to 24, the mean being 19.3 years. Their only FL instruction had taken place in high school; and none had previously studied Portuguese or lived in a Portuguese-speaking country. They had been studying Portuguese from three to eight weeks. The second group consisted of 47 native English speakers currently enrolled in advanced Portuguese classes at Brigham Young University. Three were females, and 44 were males. They ranged in age from 19 to 26, the mean age being 21.7. In-country exposure to the target language averaged 20 months, in either Brazil or Portugal, with Portuguese their normal tool of communication. In addition, they had had two months of intensive instruction in Portuguese prior to taking up residence in the foreign country. The beginning group ranged from 1 to 1+ on an FSI scale; the more advanced group was in the 2 to 2+ range.

Method. By random selection, half of each group received the all-Portuguese (FL) test; the other half received the form with NL cues. As in the Brazilian study, each group was also tested on the alternate form of the test to facilitate assessment of affect. To avoid differences that might be attributable to instruction, students in the beginning (intensive) group took the second form of the test the following day. The upper group took their second administration a week later. All students completed the questionnaire after they had sat for both tests. Once more, the independent variable was test form—all NL or one with NL distractors. The moderator variable was level of proficiency, and the dependent variable was performance on the AMLEX. Statistical analysis consisted of a two-way analysis of variance; and affect was again evaluated by chi square.

Results. Analysis of variance calculations disclose a significant difference between the two achievement levels (at less than .0001). Differences according to test type approached significance (.0847). But the interaction of test type and achievement level was nonsignificant.
Table 8

ANALYSIS OF VARIANCE: American Subjects
(each term ordered last in model)

\[ Y = A + T + AT + E \]

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Achievement</td>
<td>1</td>
<td>37093.</td>
<td>37093.</td>
<td>863.92</td>
<td>.0001 (sig.)</td>
</tr>
<tr>
<td>Test</td>
<td>1</td>
<td>130.25</td>
<td>130.25</td>
<td>3.0336</td>
<td>.0847 NS (Marginal)</td>
</tr>
<tr>
<td>Achieve- x</td>
<td>1</td>
<td>87.962</td>
<td>87.962</td>
<td>2.0847</td>
<td>.1556 NS</td>
</tr>
<tr>
<td>Test</td>
<td>97</td>
<td>42.936</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9

COMPARISON OF MEANS BY LEVEL: American Subjects

<table>
<thead>
<tr>
<th>All-Portuguese form</th>
<th>English Distractor form</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>30.037</td>
<td>34.185 Sig.</td>
</tr>
<tr>
<td>High</td>
<td>70.333</td>
<td>70.739 NS</td>
</tr>
<tr>
<td>Total</td>
<td>50.185</td>
<td>52.462 NS (Marginal)</td>
</tr>
</tbody>
</table>

Table 9 indicates that there was a significant difference in the low group in the direction of the form with English distractors. The high group performed almost identically on both forms.
Turning to the questionnaire results, we note that among those who expressed a specific preference for one form, almost two-thirds of the advanced group preferred the all-foreign language test, while a slim majority of the low group expressed preference for the form with NL distractors. On perception of ease, the high group was almost evenly divided, while the same slim majority of beginning students viewed the NL distractor form as being easier. Results did not reach significance.
Table 13

PREFERENCE MEASURE (Americans): Form with NL Cues First

<table>
<thead>
<tr>
<th>FL Preference</th>
<th>NL Distractor Preference</th>
<th>Undecided</th>
<th>Significance Level (X²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>19</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>High</td>
<td>17</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

But analyzing preference according to which test form was taken first, we find a trend in favor of whichever form came last. An exception was the nearly even split among high students when the FL form was administered second. On perception of ease, there were significant differences at both levels in favor of whichever test was administered second. Finally, in regard to the question to the low group on whether or not the NL distractors were helpful, confusing, or neither, there was an indication (at the .02 level of significance) that they were helpful.

Discussion. In the third study, the major hypothesis was confirmed. Beginning-level students were able to perform significantly better on tests with native-language cues than on all-FL tests. There was also a tendency for them to prefer the NL distractors. Combining the results from studies

Table 14

COMPARISON OF MEANS BY LEVEL:

BRAZILIAN AND AMERICAN SUBJECTS

<table>
<thead>
<tr>
<th>Group</th>
<th>FL form</th>
<th>NL Cue form</th>
<th>N</th>
<th>Significance Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Americans</td>
<td>30.037</td>
<td>34.185</td>
<td>54</td>
<td>.01 (Sig.)</td>
</tr>
<tr>
<td>Low Brazilians</td>
<td>54.900</td>
<td>59.200</td>
<td>20</td>
<td>NS</td>
</tr>
<tr>
<td>High Americans</td>
<td>70.333</td>
<td>70.739</td>
<td>47</td>
<td>NS</td>
</tr>
<tr>
<td>High Brazilians</td>
<td>77.909</td>
<td>72.667</td>
<td>20</td>
<td>.05 (Sig.)</td>
</tr>
</tbody>
</table>
II and III, we find that those with the lowest mean appeared to benefit significantly from tests with NL cues. Those with the highest mean performed significantly better on the all-FL form, while the two middle groups registered nonsignificant differences. Affect results for Americans were somewhat confounded by sequence interaction,* but a low-group preference for NL cues did emerge. The high-scoring Brazilians perceived (significantly) that the all-FL form was easier than the form with NL cues; and among the strong majority expressing preference for one of the two forms, three-fourths favored the all-FL form.

Conclusions and recommendations. While rather low-proficiency language learners appear to perform better on a multiple-choice test utilizing NL distractors, the difference on the AMLEX is not of sufficient magnitude to recommend an immediate wholesale shift toward tests with native-language cues—even for the beginning student. But there is cause for cautious optimism that such cues may be desirable. For one thing, they may well reduce test anxiety, as was apparent during Maluf's test administration of low-level students. Improved test affect in turn promises not only a more positive attitude towards instruction but the likelihood of more accurate evaluation.

It is also quite possible that the effect of NL cues would be even more dramatic on a reading test, for example, where the target-language material is considerably more complex than that on the AMLEX and on which the multiple-choice options are far more difficult. Experimental replication with other test formats is highly recommended; first, in order to determine the degree to which they benefit students of lower language proficiency, and also to indicate whether affect is similarly positive. Replication with other types of language tests could determine whether or not persons approaching the intermediate level might likewise benefit from this strategy.

*It appears that the reason for the shifts in affect might well be due to the following: Sheila Maluf, in administering the tests to low-level American students, found them demoralized when they took the all-Portuguese test form first. They were greatly relieved when they took the form with NL distractors. These students rated the form with NL cues very high. Those who took the form with NL cues first, had little trouble with the all-FL test the following day because of the dramatic practice effect from the initial administration. As a result, their attitude toward the FL test was rather benign.
BIBLIOGRAPHY


