Measuring Listening Indirectly Through Monitor Control

Harold S. Madsen
MEASURING LISTENING INDIRECTLY
THROUGH MONITOR CONTROL

Harold S. Madsen

Overseas requirements to evaluate the aural competence of large numbers of students, coupled with contemporary research in testing and in language acquisition, have prompted serious investigation into the feasibility of alternate modality listening examinations. For example, in Egypt alone, two million secondary-school students are required to study English and to become proficient in oral skills as well as in reading and writing. All students are evaluated annually in nationally sponsored exams, but no satisfactory method has yet been identified to assess their listening or speaking ability. And even if an appropriate instrument were available, there are insufficient facilities, equipment, and personnel available nationwide for large-scale conventional aural testing. These are the needs which have prompted experimentation with alternate-modality listening tests.

Interest in indirect measurement of language skill is of course not new. Reasonably successful language aptitude batteries have been available for decades (Carroll and Sapon 1959; Pimsleur 1966). And indirect measures of writing skill, as found in the TOEFL exam, for example, have demonstrated high concurrent validity. Indirect measures of listening and speaking, however, have not proven as successful. Despite the ingenious efforts of many, the United States Foreign Service Institute, to mention only one example, stands by its costly interview procedure as "the most valid measurement of general speaking proficiency currently available" (Jones 1975:4). Paper and pencil tests of pronunciation had become suspect even when discrete measures of listening were still in vogue (Harris 1969:90).

Nevertheless, test research points towards a possible breakthrough. Bernard Spolsky has postulated that indirect test devices could be developed as surrogates for more expensive direct evaluation procedures (1968:88-94). And ESL test expert John Oller, in reviewing integrative test research carried out in recent years, likewise advances this hypothesis. He notes, for example, that written cloze tests when correlated with a variety of subtests always tend to have the highest relationship to listening comprehension tests, despite the fact that the former include neither phonology nor the tight time constraints of the latter. And he even suggests the feasibility of substituting one modality for another:
For example, with non-native speakers, it should be useful to compare performance on a written cloze test with performance on an oral cloze test. If the correlation is sufficiently high, it would be possible to substitute one type of test for the other. (Oller 1973:108; see also Clark 1975:11; and Oller et al 1972)

Given the feasibility of an alternate modality listening test, it is now useful to consider research findings that might suggest the special properties and constraints of the surrogate exam. Recent language acquisition studies, it happens, are particularly instructive.

Experimental research by Dulay and Burt on the order of difficulty of grammatical morphemes (or functors) strongly suggests universal language acquisition strategies among child second language learners, regardless of language background (1973, 1974). And Taylor points out important cognitive similarities in child and adult acquisition of language (1974:33). Bailey and others corroborate this view with experimental work which discloses a high degree of agreement in the difficulty of various grammatical structures for children and adults, and between adults of various language backgrounds (1974:242). Diane Larsen-Freeman has replicated and extended the comparison of adult and child order of acquisition studies and has examined the effect of varying data collection measures. She has found, as had Bailey and others, that the Bilingual Syntax Measure (BSM) provides a very strong correlation between acquisition orders for people of differing language background and ages, but that this does not hold for all of the other measures used. She speculates that such lack of consistency might be due to one or more causes: "modality differences, specific task effects, skill differences, etc. (1975:418; see also 1976)" Krashen and others corroborated and extended the Bailey findings (Krashen and Seliger 1976; Krashen and others 1976).

Moreover, Krashen hypothesized that the inconsistent Larsen-Freeman results across varying data collection tasks stemmed from and illustrated the very nature of adult language learning. Positing a Monitor theory, Krashen suggested that even in formal school-type settings adults master language through two processes: 1) acquisition or unsystematic exposure much as children do their native language, and 2) learning or a systematized, step-at-a-time procedure with regular feedback. Under proper conditions, the formalized, rule governed system imposes itself and monitors output. Referring to Larsen-Freeman's results using the BSM and supplementary evaluation instruments, Krashen reasoned that:

one feature these supplementary tests had in common is that they allowed more response time (and hence more
processing time) than did the BSM, and that during this extra time subjects were able to involve more consciously learned linguistic knowledge in their responses... Thus, the change in difficulty order may have been brought about by the subjects' having altered their output, under the influence of a consciously learned and more idiosyncratic pedagogical grammar. (Krashen and others 1976:150) [In brief, then,] according to Krashen's Monitor Model, adult second language performers depart from the child's L2 order when monitoring time is allowed and when they focus on form. This accounts for Larsen-Freeman's results with written tests. (Krashen and others 1977:340)

As a result of his research, Klein noted over a dozen years ago that monitoring of one's own speech is necessary in order for the individual to control his thought processes and verbal expression; moreover, he even traced the rationale for this to Freud at the turn of the century (Klein 1965:242, 269, 270). Subsequent experimental studies have borne this out (for example, Holzman and Rousey 1970:240:241; Webster and others 1970; Klein and others 1970; Yudkovitz and others 1973; Belmore and others 1973). Labov indicates that constant audio monitoring is needed for the maintenance of prestige forms learned later in life. Even school teachers, he says, use nonstandard English "in their most casual speech," when "the minimum attention is given to the speech process." (Labov 1969:17, 15). Audio monitoring can also be disrupted by great excitement, intense interest in a subject, fatigue, distractions, and by being unable to hear oneself (Labov 1969:33-34). It should be noted as well that people vary considerably in their ability or inclination to monitor themselves (Krashen 1977:156-158).

While there is wide agreement that we need to hear what we say, minimize slips of the tongue (Bolinger 1975:389), and monitor prestige forms learned as adults, there are those in language acquisition who do not accept Krashen's Monitor Model (Hatch 1977, Frazier 1977). Nevertheless, it is not only compatible with recent research findings, cited above, but it also helps account for the sometimes dramatic differences in ESL and foreign language classes between high performance on a written classroom exercise and bumbling oral production. Krashen holds that adult second-language performance is based on the acquired system, the learned system functioning simply as monitor. Tests that permit the operation of the monitor will result in "idiosyncratic" errors, which "reflect each learner's conscious mental representation of linguistic regularities in the target language (Krashen 1977:152, 154)." It follows that reasonably accu-
rate language which results from labored, analytical monitoring may represent an unrealistic index of the person's actual ability to communicate. There would seem to be an advantage, therefore, in eliciting acquired rather than learned language, by not permitting the monitor to function. This is possible, according to Krashen and others, by concentrating on natural conversation of communication rather than on language form, and by carefully limiting processing time. ESL test specialist Upshur concurs on the need to control processing time (1975:59).

METHOD

Examination format. -- In determining what test format to employ in the AUC experiment, it was of course decided to select a "communicative" type exam. A variety of reports on cloze tests initially recommended this format. It is communicative, and it does not focus on language form. Also it correlates well with tests involving listening (Irvine and others 1974; Oller and Conrad 1971, Oller 1973). For example, compare the correlations of various tests (including cloze) with the listening comprehension section of the internationally administered Test of English as a Foreign Language (TOEFL):

<table>
<thead>
<tr>
<th>Table 1</th>
<th>CORRELATIONS OF VARIOUS TESTS WITH THE TOEFL LISTENING SECTION (159 subjects)</th>
</tr>
</thead>
<tbody>
<tr>
<td>List</td>
<td>.69</td>
</tr>
<tr>
<td>Comp.</td>
<td></td>
</tr>
<tr>
<td>TOEFL</td>
<td></td>
</tr>
</tbody>
</table>

(Adapted from Irvine et al 1974:249, 251)

Except for the adjusted total on the TOEFL, the cloze test showed the highest correlation, higher even than dictation with its aural component. While Harris (1969:20) indicates that correlations in the .70's or .80's are adequate in equating two tests of the same skill, it was decided to demand performance in the upper portion of this range -- where we might safely assume the two measures were "tapping an underlying competence" (Oller and Streif 1975:33). Despite the communicative focus in cloze, we decided to select a test type that would be less conducive to monitoring and thereby more likely to generate a higher measure of concurrent validity.

1978 Languages and Linguistics Symposium
Even though our prospective test would be in a printed rather than an oral modality, it was deemed important to avoid tasks which would incorporate extensive traditional reading tasks. For placement test studies at AUC between 1974 and 1976 indicated a correlation of only .65 between the reading subtest of the Michigan Test of English Language Proficiency (MTELP) and the Michigan Test of Aural Comprehension (MTAC). Referring to Table 1, we see a very similar relationship between the listening and reading subsections of the TOEFL, a correlation of just .63. This meant avoiding printed versions of lecturlette, for example.

It was finally decided to utilize standard listening comprehension-type items as found on the CELT, the Michigan, and the TOEFL (excluding the lecturlette items on the TOEFL). These short, simple conversation items would provide the desired communicative focus while avoiding complex reading comprehension tasks. To further restrict monitoring, a strict time limit would be imposed.

The alternate modality listening examination (AMLEX) consists of two sections of conversational utterances. Section one consists of 45 questions requiring appropriate responses:

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

Section two consists of 45 statements requiring selection of an appropriate paraphrase:

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

**Subjects.** --Two groups of native Egyptian applicants to the American University in Cairo were administered the AMLEX, the Michigan listening test (MTAC) and the Michigan battery (MTELP). These consisted of 72 graduates from Egyptian colleges and universities as well as 73 undergraduate transfers and graduates seeking admission to the AUC Management program, or a total of 145 applicants. In a follow-up study, two groups of Egyptian students who were enrolled in the English Language Institute at AUC also sat for the three examinations, 115 first semester and 94 second semester.

**Procedure.** --To counter practice effect, half of each of the two groups of applicants to AUC took the alternate
modality listening test prior to the Michigan listening test; the other half of both groups took the Michigan listening test first. A strict 25-minute time limit was imposed on the alternate modality exam (3 minutes more than the 90-item commercial listening test requires). This alternate modality test and the Michigan were administered the same day. Follow-up groups observed the same procedure, half sitting for the AMLEX first, the second half taking the MTAC first; however, in the follow-up administrations the two tests were administered on different days. The first semester group was limited to 25 minutes on the alternate modality test, but to assess the impact of the time restriction the second semester group was given 35 minutes—ample time, it was felt, for the Monitor to intervene.

Alternate modality exam papers of the 145 AUC applicants were triple scored in order to determine the optimum scoring procedure. The first score was simply the percent of the total items that were correct. The second score incorporated a standard guessing correction. The third score represented the percent correct out of those attempted, thereby minimizing the effect of the time restriction. To assess the three scoring procedures, each set of scores on the alternate modality test was correlated with scores obtained on the Michigan listening test. The AMLEX scores generating the highest correlation would identify the best means of scoring the surrogate listening exam.

RESULTS

Results of the Pearson product-moment correlations based on the triple-scored alternate modality (AMLEX) tests and the Michigan listening test (MTAC) indicate that the highest correlations occur when the guessing correction is imposed. The lowest occur when the time factor is minimized, the Management applicant group dropping from .82 to .73 (see Table 2)

\[ R = \frac{W - \frac{1}{N}}{D - \frac{1}{2}} \]

1The formula is \( R = \frac{W - \frac{1}{N}}{D - \frac{1}{2}} \); the number right minus the number wrong divided by the number of distractors, minus one. Thus with three distractors, the correction factor consisted of the number correct minus the number wrong divided by two.
Table 2

Comparative Correlations between the Alternate Modality Listening Examination (AMLEX) and the Michigan Test of Aural Comprehension (MTAC) Utilizing Three Scoring Procedures for the AMLEX

<table>
<thead>
<tr>
<th>Score 1</th>
<th>Score 2</th>
<th>Score 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>(% correct)</td>
<td>(Score with Guessing Correction)</td>
<td>(% correct of attempted-time discount)</td>
</tr>
<tr>
<td>Grad. Applic's (N = 72)</td>
<td>.84</td>
<td>.86</td>
</tr>
<tr>
<td>Mgt. Applic's (N = 73)</td>
<td>.80</td>
<td>.82</td>
</tr>
</tbody>
</table>

Both groups are consistent across the three measures. Applying the Kuder-Richardson Formula 21 to the graduate applicant data, we obtain a reliability figure of .969.

The relationship between performance on the surrogate listening test and on the commercial proficiency battery was likewise determined through Pearson product-moment correlations (see Table 3).

Table 3

Comparative Correlations between the Alternate Modality (AMLEX), the Michigan Listening (MTAC) and the Subscores and Total of the Michigan Test of English Language Proficiency (MTELP)

<table>
<thead>
<tr>
<th>GRADUATE APPLICANTS (N = 72)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>Vocabulary</td>
<td>Reading</td>
<td>Total MTELP</td>
</tr>
<tr>
<td>AMLEX</td>
<td>.85</td>
<td>.73</td>
<td>.74</td>
</tr>
<tr>
<td>MTAC</td>
<td>.78</td>
<td>.61</td>
<td>.64</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MANAGEMENT APPLICANTS (N = 73)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Grammar</td>
<td>Vocabulary</td>
<td>Reading</td>
<td>Total MTELP</td>
</tr>
<tr>
<td>AMLEX</td>
<td>.81</td>
<td>.65</td>
<td>.62</td>
</tr>
<tr>
<td>MTAC</td>
<td>.76</td>
<td>.58</td>
<td>.48</td>
</tr>
</tbody>
</table>

The surrogate test was consistently a better predictor of subscores and total score on the battery. The highest subscore correlation for both groups was with grammar, virtually as high as the correlation in Table 2 between the surrogate and commercial listening tests. The performance of
the commercial listening test parallels that reported by Irvine and others (Table 1), correlations with grammar and total score being higher than with vocabulary and reading.

In the follow-up administrations of the alternate modality test and the commercial listening test to students enrolled in the English Language Institute at AUC, the timed AMLEX again reaches the .80's while the untimed AMLEX drops into the .50's (Table 4).

Table 4

<table>
<thead>
<tr>
<th></th>
<th>ELI First Semester</th>
<th>ELI Second Semester</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 Min. Timing</td>
<td>.82</td>
<td>.82</td>
</tr>
<tr>
<td>(N = 15)</td>
<td></td>
<td>(35 Min.)</td>
</tr>
<tr>
<td>(N = 94)</td>
<td>.57</td>
<td></td>
</tr>
</tbody>
</table>

DISCUSSION

Concurrent validity measures of the alternate modality examination consistently achieved experimental objectives with correlations in the .80's when a 25-minute time limit was adhered to. These correlations of .86, .82, and .82 are sufficiently higher than the .76 correlation between cloze and listening to recommend the surrogate test over traditional cloze as a measure of listening comprehension proficiency. But the reliability estimate of .969 is undoubtedly inflated, since a number of persons failed to complete the exam.

The significance of the time restriction is borne out in two sets of data. First, when time was discounted by scoring the percent correct of those attempted by AUC applicants, correlations dropped from .86 to .81 for graduate applicants and from .82 to .73 for Management applicants. Presumably the lower correlations reflect greater intrusion by the Monitor. The second set of data (Table 4) reflects an even greater difference when time was discounted. One explanation for this greater difference is that during the test administration to AUC applicants, all were aware that a strict time limit was being imposed; thus we can assume that virtually everyone would be striving to answer the questions as quickly as possible, thereby restricting Monitor function. On the other hand, the second-semester ELI group realized they would have ample time for the test, thereby encouraging Monitor function.
In a personal conversation, John Oller has suggested a second possible explanation for the modest .57 correlation. Rather than Monitor intrusion, the lower correlation he feels could be due simply to "noise" (Oller 1978). "Noise" could include such factors as a greater opportunity for cheating, the possibility of fatigue, previous exposure to the exam, coaching, indifference as a result of knowing the exam didn't constitute part of their grade, or higher scores which could depress the correlation.

Since "noise" might indeed have influenced the second semester ELI results, it was decided to control for each of the variables mentioned and to evaluate Monitor function more accurately by administering two forms of the surrogate test. With alternating administrations, each group of ESL students would be restricted to 25 minutes on one form but not restricted on the other form. While this study will not be completed for several months, the results of a small pilot administration are in. The present AMLEX was recently administered to 16 non-native speakers at Brigham Young University: 7 Spanish speakers, 1 Portuguese, 2 Chinese, 1 Korean, 2 Japanese, and 3 Germans. These 16 ESL students also sat for a 50-minute experimental TOEFL listening examination. Since time did not permit their also sitting for form B of the AMLEX, students were told they would be evaluated on the number of items completed correctly in 25 minutes as well as on their total score; and that they would be permitted to have as long as needed to finish the test after the 25-minute check. Papers were triple scored and the results correlated with performance on the experimental TOEFL. The Pearson product-moment correlation between the AMLEX score at the end of 25 minutes and the experimental TOEFL was .81; but the correlation between these two examinations dropped to .75 when students were permitted to continue working on the AMLEX without a time limit, and the correlation was .76 when papers were corrected for the percentage correct out of those attempted. While the results must be interpreted cautiously because of the small N, they tend to support the notion of Monitor interference as exemplified in Table 2. The differences in the recent pilot study, however, are not nearly as dramatic as that between the two ELI correlations, but then conditions were not identical either. In brief, the data reveal that time restrictions are necessary in conducting an alternate modality listening test, although the exact magnitude of correlation differences between timed and untimed tests has not yet been determined.

2 This consists of experimental listening items prepared by TOEFL for evaluation and possible later incorporation in the official TOEFL instrument.
High correlations between the surrogate test and a variety of measures such as listening comprehension, grammar, and total scores on a proficiency battery recall Oller's observation that "when such vastly different tests consistently intercorrelate at the 0.85 level or better . . . we may reasonably conclude that they are tapping an underlying competence" (Oller and Streiff 1975:33). Thus while the alternate modality listening test is obviously not measuring such skills as proficiency in processing phonological input, it is apparently tapping language skills which undergird listening.

CONCLUSIONS AND RECOMMENDATIONS

Results of the alternate-modality experiment at the American University in Cairo support the hypothesis that printed listening-test cues utilizing an integrative conversational format can substitute for direct tests of listening. But for them to be fully effective, strict time limits must be imposed. Since the surrogate listening test is similar in form to various commercial aural comprehension tests, it appears likely that listening tests such as the Michigan and CELT could produce similar results in a printed modality with timed administration. Findings seem to corroborate Monitor theory; a timed test with a communicative focus diminishes or eliminates Monitor activity.

Certain cautions, however, should be considered. All findings have been interpreted in terms of experimental groups. Individual variation has not been examined. Moreover, the students utilized in the study tend to have reasonably equivalent proficiency in listening and reading. There is a good possibility, then, that an individual whose skill in one mode far surpasses his skill in another mode may not be properly evaluated by a surrogate test. Equally important is the possibility of a negative backwash effect on instruction. If used to replace a direct measure, a surrogate might well result in a decreased emphasis on oral activities. But if used where no formal oral evaluation had previously been made, it could provide useful information about listening skills, and it might serve as a mild catalyst for additional attention to listening comprehension. Ideally it would complement a productive measure such as a brief oral interview.

Experimental research is now needed to assess individual variation on the surrogate as well as the full effect of timed versus untimed administration. Further evaluation would be useful, too, on implications for Krashen's monitor theory that have surfaced in this research. Finally, other formats, such as timed cloze, need to be evaluated in order to determine the most powerful model for alternate modality testing.
REFERENCES


Frazier, Bruce. Discussion at the American University in Cairo; Cairo, Egypt (December 1977).


Hatch, Evelyn. Discussion at the American University in Cairo, Cairo, Egypt (December 1977).


Larsen-Freeman, Diane E. "The Acquisition of Grammatical Morphemes by Adult ESL Students," TESOL Quarterly, 9 (December 1975), 409-419.

Larsen-Freeman, Diane E. "An Explanation for the Morpheme Acquisition Order of Second Language Learners," Language Learning, 26 (June 1976), 125-134.

Oller, John W., Jr. Discussion in Mexico City, Mexico (April 1978)


