TESL REPORTER

Teaching English as a Second Language

Vol 12 No. 1 Laie, Hawaii Fall 1978

# The Identification of Irrelevant Lexical Distraction: An Editing Task

## by J. Donald Bowen

As cognitive approaches to language teaching have won favor, language-testing theory and practice have been directed toward the assessment of communicative skills, rather than being limited to a determination of the mastery of specific structural or lexical points or patterns. This has prompted an effort to develop tests that are meaningfully related to the communicative function of language. More specifically, it is felt that assessment tasks should be designed to reflect some real-life activity for which language skills are authentically employed. The term that reflects communicative competence is integrative testing, in which bundles of features are assumed to be working together to carry a message, with no necessity felt to analyze the function, or indeed the identity, of individual features.

Unlike the discrete-point tests with their objective of testing a single point per item, more or less disembodied from context, integrative tests can be related to various functions of the language of real life. Thus a dictation test reflects the secretary's task of normal stenographic transcription; a test to see how aural comprehension is affected by an overlay of noise is comparable to communication in a cocktail-party atmosphere; an oral cloze test with intermittent fading or "gapped listening" is similar to defective short-wave radio reception; and a regular written cloze test is an echo of reading manuscript with unknown vocabulary items whose meanings must be inferred from context. Encouraging results from these tests have stimulated a continuing search for other formats that will measure communicative competence.

A recent paper by Alan Davies suggests a format for an integrative test of communicative competence, a speeded reading test, one which is in a sense the opposite of a written cloze test. Instead of filling in missing items the examinee is asked to identify and cross out superfluous words that have been inserted in the text. The present report is a replication of that format, conceived as one

## CONTENTS

possible subtest in a proficiency battery that might be employed to measure the preparation of non-native English-speaking applicants for enrollment in an English-medium educational institution.

The test was constructed as follows: A text was selected as a sample of written English appropriate to the interest and proficiency level of the prospective examinees. The selection chosen was a slightly modified version of the first six paragraphs of "Clocks Through Time," Reading 11 from *A Reading Spectrum* (Book 6 of the *Progressive Reading Series*, by Virginia French Allen). To the original text of 450 words were added 40 irrelevant additional words. These were selected and inserted by a randomizing process. The throw of three dice determined the interval of the text between insertions. The source of the insertions was a separate book, opened to a random page. The word to be inserted was the first word on the second line of the left-hand page, unless that line began a paragraph, in which case another page was turned. For each subsequent insertion, an additional page was turned. The resulting modified text was duplicated on a single sheet of paper with the heading "Editing Test," with instructions as follows:

> **Instructions**: In the following passage, unnecessary words have been added to the text. Find them and cross them out. For example:
>
> Have you trees eaten your dinner yet? The word 'trees' is unnecessary and is therefore crossed out. The test will be timed, so work fast. Stop when you are told to stop.

As the instruction indicates, the test is taken under time pressure. Fifteen minutes was allowed, which proved adequate for virtually all examinees to finish. The purpose was not to allow insufficient time, but to specify an attitude of urgency to complete the task.

The test was given experimentally on December 13, 1976 as a "caboose" to 145 applicants for admission to the American University in Cairo. The regular admissions battery consisted of the Michigan Test of English Language Proficiency, the Michigan Test of Aural Comprehension, and a locally set written composition test. (Also another experimental "caboose" test was administered.) This joint administration allows a comparison of the experimental test with different aspects of the Admissions Battery.

Data on the results of this administration are:

| | |
|---|---|
| N | = 145 |
| Mean | = 76.3 |
| SD | = 14.53 |
| Range | = 34 - 100 |
| Reliability | = .95 (estimated by the Diederich formula and by Kuder-Richardson Formula 21) |

Scoring the Editing Test involves some problems. Scoring can be a very tedious task, and a confusing one, since there are two kinds of errors possible, faults of omission and of commission, i.e., failing to mark words that should be omitted and marking words that should not be omitted. Following a key laid alongside a completed test and mentally matching performance with these two error types can be very disorienting. A procedure that improves accuracy and speed is to prepare a key by blacking out on a test form all words to be omitted, aligning this key under the test paper but above a light source, then placing a check mark in a distinctive color (e.g., green or purple) before each word that should be omitted. A tally can then be made quickly by counting all words marked only once, those marked twice being items successfully completed. Performance scores are determined by subtracting the total number of errors (omission and commission) from 100.

So there are two kinds of error possible: insufficient and superflous editing. The insufficient errors are planned by the test—they need over correction. The superfluous editing involves an overreaction to the data, an inability to recognize the actual needs of the editing task.

An item analysis of the 40 superfluous lexical insertions, based of the 20 high and 20 low papers (scoring respectively above 89 and below 74) in one section on the 72 subjects, reveals that all fall within the

difficulty range of 40 to 85, with an average item difficulty of 66.1. The discrimination range for the same items ranges from 12.5 to 42.5, averaging 26.9. These are very encouraging results; no item needs to be revised because of item weakness.

Of the 40 planned items, 18 were correctly identified as superfluous by all 20 of the high papers, and the average error rate for the other 22 items was 1.4 per paper. All 40 items were incorrectly answered among the 20 low papers, and the average error rate was 12.15 per paper.

The error rate on unplanned items was slightly higher for the high 20 papers, considerably lower for the low 20. Among the high papers 30 errors provided distraction at an average rate of 1.6 errors per paper. For the low 20 papers 154 items provided distraction at an average rate of 7.0 points per paper. Thus both the planned and the incidental items are working efficiently. It is interesting to note that of the 450 potential unplanned items, the words in the original selection, only 164, or about 36.4 per cent, were ever selected, and two-thirds of the 164 (109) were selected by only one of the 40 papers. So the incidental items do not play a very important role in the test. Still they must be considered, since if there is no penalty for wild guessing an examinee could increase his score by indiscriminately marking everything even remotely suspected as being superfluous.

But what does the test measure? A good reliability figure and an encouraging item analysis are all very well, but if the test does not validly measure some relevant aspect of competence, it is of little use.

Coefficients of correlation for the seven scores of the AUC Admissions Battery and the Editing Test are shown in Table 1. Abbreviations and explanations are: GC, VOC, and RD are the grammar, vocabulary, and reading comprehension subtests of the Michigan Test of English Language Proficiency (MTELP), MICH is the equated score of the Michigan Test of Aural Comprehension (MTAC), WC is the percentage score of the written composition test, and AB is the Admissions Battery score, which is an average of MICH, AC, and WC.

What conclusions can reasonably be drawn from an analysis of the Editing Test? It's not an overwhelming success, nor is it a hopeless failure. As with almost any test some items are stronger than others, and the constraints of the format make it somewhat difficult to modify or reorder items.

First it seems clear that chance cannot be depended on to arrange for insertions. In the present test some insertions are very conspicuous, while others manage to partially conceal themselves. If a random procedure is followed, there will occasionally be reasonable insertions with no rational basis for their deletion. Suppose the following sequence is produced: "Since there were no

**TABLE 1**

Coefficients of Correlation on the Admissions Battery and the Editing Test for Two Groups of Applicants to AUC—December 1976

| | | GR | VOC | RD | MICH | AC | WC | AB | EDIT |
|---|---|---|---|---|---|---|---|---|---|
| | N | 70 | 71 | 71 | 70 | 70 | 54 | 53 | |
| Grad. | SD | 7.30 | 9.49 | 4.44 | 16.46 | 22.13 | 17.66 | 10.44 | 13.48 |
| | r | .736 | .564 | .647 | .684 | .645 | .456 | .690 | |
| | N | 73 | 73 | 73 | 73 | 73 | 48 | 48 | |
| Manag. | SD | 6.98 | 7.32 | 4.18 | 14.80 | 17.20 | 10.77 | 10.57 | 14.27* |
| | r | .693 | .613 | .600 | .685 | .648 | .533 | .714 | |

*Average

*(continued on page 14)*

# The Identification of Irrelevant Lexical Distraction

*(continued from page 3)*

interesting planes or trains to catch, however, people were not concerned about knowing the exact time." The word "interesting" is redundant but not grammatically incorrect. But note that if words like "interesting" are to be deleted, why not also omit "however" and "knowing" and "exact."

To provide a rationale, insertions should damage the grammatical or lexical integrity of the sentence. (Presumably sentence structure is what is being measured; the strongest correlation is with the grammar subtest of the MTELP.) The instructions accompanying the present test failed to do this, since the word "unnecessary" was used as the judgment criterion. This instruction leads directly to some performance errors that might have been avoided with better instructions. A few examples of misleading items are:

| Sentences not needing deletion | Marked for deletion | No. of papers (high-low) |
|---|---|---|
| It was probably around 3,000 years ago... | probably<br>around | 6 (1-5)<br>7 (2-5) |
| Candles and water clocks helped people know how much time had gone by. | by | 13 (6-7) |
| So after glass blowing was invented, the hourglass came into use. | blowing | 9 (4-5) |
| These did not always tell the correct time, either. | either | 8 (3-5) |

Note that the discrimination power of these items is relatively weak.

A typographical error in the test form was responsible for the deletion of two words, often by the same subject. The word "divisions" appeared as "decisions," with the error acting as a lightning rod for corrections:

| | | |
|---|---|---|
| As the sun passed overhead, he marked even divisions on the circle... | even<br>decisions | 15 (8-7)*<br>13 (5-8) |

Some of the better items required interpretation by means of non-adjacent data to identify their inappropriateness. One nonplanned, incidental item illustrates this:

| | | |
|---|---|---|
| One of the first such clocks was built for a king... | such | 10 (0-10) |

It takes reference to the preceding sentence, where the first clock to be built with a face and and hour hand is mentioned, to justify keeping the word "such." To sense this requires sophistication and a discerning feel for the structure of the language. To avoid (or minimize) problems of this kind, examinees should be informed that the insertions are inappropriate, not just unnecessary, that errors of grammar, usage, style, or logic result from the insertions.

Non-adjacent clues to inappropriate inclusion provide some of the strongest items. Ideally an insertion reads reasonably until an expected disharmony arises that forces a re-evaluation. The ability to reanalyze under time pressure distinguishes the strong from the weak examinees. A few examples of good items follow, with the insertions underlined for

*The only item in the test (other than random commission items marked by a single subject) to attract more high than low papers.

easy identification:

| Sentences needing deletion | Not marked or wrongly marked for deletion | No. of papers (high-low) |
|---|---|---|
| The people could tell which part of the day represented it was by noticing. . . | represented<br>was | 20 (6-14)<br>11 (5-6) |
| Of course a sundial did not work at night or on cloudy days, so men expected kept inventing other ways. . . | expected<br>kept | 22 (3-19)<br>12 (3-9) |
| Usually it considered did not even show the correct hour. | considered | 14 (1-13) |

Or a grammatical mistake may be present in the inserted word, but is noticed only by good students:

| | | |
|---|---|---|
| .the first with a face and an hour hand was imitate made | imitate | 14 (0-14) |
| The mention clock did not show minutes or seconds | mention | 20 (4-16) |
| Since there were no trains to interesting catch | interesting | 17 (1-16) |

Lexical association appears to deflect judgment; when an insertion is in the right register, it may be retained in spite of grammatical inappropriateness:

| | | |
|---|---|---|
| As the sun passed overhead, he flew marked even divisions. . . | flew | 13 (0-13) |
| . . .and water clocks had number to be refilled. | number | 16 (0-16) |

First and last words in a sentence seem to be difficult to omit, particularly for low students:

| | | |
|---|---|---|
| Above it was about 600 years ago that. . . | above | 17 (1-16) |
| Find they still did not keep correct time. | find | 16 (0-16) |
| . . .as the shadow of the stick crossed it presence. | presence | 12 (0-12) |

It would probably help examinees to know that adjacent words are never scheduled for deletion. Only one word is inserted in a location, so only one word is to be deleted. Also, some idea of the interval between insertions could be given for the guidance of the examinees: in the present test from a minimum of three to a maximum of eighteen words (the extreme possibilities of three dice). This would forestall the occasional student who seems to think an entire phrase, or even sentence, is superfluous.

In summary, I would say that the Editing Test has possibilities, at least sufficient to justify further experimental use. It is reliable, valid, and practical. An item analysis shows that even random construction of the test produces effective items. Perhaps a thorough pre-

analysis, or an experimental administration, as a test is being developed, would permit the elaboration and refinement that would strengthen the test as an instrument to measure underlying language competence and proficiency.

## REFERENCES

Davies, Alan. 1975. "Two Tests of Speeded Reading." In *Testing Language Proficiency*, Randall L. Jones and Bernard Spolsky, eds. Washington, D.C.: Center for Applied Linguistics, 119-130.

Allen, Virginia French. 1975. *A Reading Spectrum,* Book 6 of the Progressive Reading Series. Washington, D.C.: U.S. Information Agency.