Cloze method: what difference does it make?

Carol A. Chapelle and Roberta G. Abraham

Language Testing 1990; 7; 121
DOI: 10.1177/026553229000700201

Language Testing 7, 2 (1990); pp. 121–146

Cloze method: what difference does it make?

Carol A. Chapelle and **Roberta G. Abraham** Iowa State University

Considerable evidence suggests that cloze techniques can create tests which measure aspects of students' second language competence. However, it remains unclear how variations in the cloze procedure affect measurement. This study compared results obtained from cloze passages constructed from the same text using four different procedures: fixed-ratio, rational, (rational) multiple choice, and C-test. The four procedures produced tests similar in reliabilities but distinct in levels of difficulty and patterns of correlations with other tests. These results are discussed in view of theoretically-based expectations for convergent and discriminate relationships of the four cloze tests with other tests.

I Introduction

Although cloze procedures do not produce perfect tests of overall language proficiency, they do hold potential for measuring aspects of students' written grammatical competence, consisting of 'knowledge of vocabulary, morphology, syntax and phonology/graphology', and textual competence, knowledge of the cohesive and rhetorical properties of text (Bachman, 1990: 87–88). The specific traits measured by a particular cloze test should depend, in part, on methods of test construction and student response. Because the effects of various cloze methods are not well understood, this study hypothesized that performance on four types of cloze tests (a fixed-ratio, a rational cloze, a (rational) multiple-choice cloze, and a C-test) would vary and then evaluated empirical data to test this hypothesis. The four tests are compared on the basis of their difficulty, reliability, and convergent and discriminant correlations with five other tests.

II Four types of cloze tests

The cloze procedure is used to construct a language test by deleting from a passage some information which the test-taker must fill in. This basic procedure for writing second language tests has been realized in several different ways, presumably resulting in tests that

differ in the specific language trait they measure, or their accuracy of measurement. The first type, the fixed-ratio cloze test, is constructed by deleting words according to a fixed pattern (e.g., every seventh word). This procedure is intended to sample regularly various types of words, some of which are governed by local grammatical constraints, others of which are governed by long-range textual constraints. A second cloze procedure, the rational cloze, allows the test developer control over the types of words deleted, and thus the language traits measured. A third cloze is constructed by altering the mode of expected response, having the student not construct an answer to fill in a blank but simply select the correct word from choices given. A fourth cloze-like procedure, the C-test, specifies that deletions are made on the second half of every other word in a short segment of text. Because of the shorter segment of text and the importance of clues in the immediate environment, this procedure most likely results in tests of more grammatical and less textual competence. Each of these types of tests has been used and investigated; however, it remains unknown exactly how these four different test-construction procedures compare.

Fixed-ratio cloze

The fixed-ratio cloze procedure for second language testing was proposed as a test of nothing less than global language proficiency (Oller, 1979). Consequently, much cloze research seeks evidence for this claim, offering little substance to the definition of specific traits that may comprise global proficiency (as indicated by cloze performance). Such studies consisted of analyses of the cloze and other language tests, which were all shown to correlate to some degree (e.g., Oller and Conrad, 1971; Irvine, Atai, and Oller, 1974; Hanania and Shikhani, 1986) or to load on one general factor (e.g., Oller, 1983). With foci on common language-test variance attributable to overall proficiency and predictive evidence for cloze validity, the theoretical relevance of different correlations between the cloze and other tests was not thoroughly explored (Oller, 1979).[1] Along with this external component of construct validation, however, a substantial amount of research on cloze items has found evidence for some cloze items as

[1] In fact, some differences in the strengths of correlation were, of course, found. The one that is given most attention is the stronger correlation between the cloze and listening tests (e.g., the Listening part of the TOEFL, or dictation tests). However, rather than interpreting these correlations as indicators of specific shared variance among these tests, they are used as an argument that both types of tests must be measures of overall language proficiency. The cloze-listening correlations observed in early studies have not been found consistently (e.g., Ilyin et al., 1987).

measures of textual competence, and others as measures of grammatical competence (Shanahan, Kamil, and Tobin, 1982; Chavez-Oller, Chihara, Weaver and Oller, 1985; Lado, 1986; Markham, 1987). Overall, these and other studies have found positive evidence for the fixed-ratio cloze as a measure of language traits, more specifically, written grammatical and textual competence.

Investigating the suggestion (Carroll, Carton and Wilds, 1959) that cloze performance may be related to cognitive abilities other than language, a few studies have uncovered evidence suggesting that cloze variance may also be related to the nonlinguistic trait, field independence (the ability to perceive analytically; McKenna, 1984), as measured by the Group Embedded Figures Test (GEFT). Hansen and Stansfield (1981) report significant disattenuated correlations between an apparently fixed-ratio cloze test (multiple-choice) and the GEFT ($r = .43$, and $r = .22$, with 'scholastic aptitude' partialed out). They concluded that cloze tests may be biased toward field independent test takers. In another study, for four out of nine groups ($n = 19$–59), Hansen (1984) found significant correlations between a fixed-ratio cloze and the GEFT ($r = .33$-$.48$). On the basis of higher correlations between the cloze and GEFT than between the GEFT and other language measures, Hansen asserts 'some support [for] the Stansfield and Hansen hypothesis of field sensitivity bias in the cloze procedure' but notes that 'the wide variation in the relationship between FD/I cognitive style and cloze test performance among the nine classes tested . . . speaks for a cautious interpretation' (pp. 320–321).[2] In a third study, Chapelle (1988) found no correlation between a fixed-ratio cloze and the GEFT for ESL students but on the same test found moderate disattenuated correlations between the two measures for remedial native speakers of English ($r = .42$) and regular freshman native speakers ($r = .63$). When verbal ability as measured by the English ACT was partialed out for the two groups of native speakers, the partial correlation was significant for the latter ($r = .53$; $n = 29$). If cloze tests do in fact measure a non-language trait as well as distinct language traits, inconsistent results from one cloze to another can be expected.

Such inconsistencies, even when different deletion ratios are used for the same text, have been noted by Alderson (1979, 1980, and 1983), who attributed them to uneven sampling of traits measured

[2] That variation is difficult to interpret, as it may be a function of the differential reliabilities of the cloze tests (and the GEFT) for the groups, which, on the basis of their anecdotal descriptions (Hansen, 1984: 313–14) appear to differ in language proficiency.

from one cloze to the next.[3] Other explanations of cloze inconsistencies are text and item difficulties relative to the group tested (Klein-Braley, 1983; Brown, 1983). Brown (1984) clarifies the relevance of a test's difficulty to its reliability and hence correlations with other measures, concluding that 'effectiveness in terms of reliability and validity, appears to be related to how well a given cloze passage fits a given sample' (Brown, 1984: 118). Passage fit refers to characteristics of the text as a whole such as reading level and topic (Alderson and Urquhart, 1985a, 1985b), but it also depends on individual items, each of which contributes to overall variance. Because items are at the root of cloze performance, it has been suggested that the cloze procedure can be improved by selecting explicitly the words to be deleted, thus creating a rational cloze.

Rational Cloze

The theoretical underpinning of the rational cloze procedure, in which the test writer selects particular items, diverges somewhat from that of the fixed-ratio cloze, which relies on regular sampling of words in the text. Rational cloze research and practice rests on the assumption that different cloze items can be explicitly chosen to measure different language traits. Some evidence indicates that test writers can select words reflecting distinct aspects of the learners' grammatical and textual competence (Bachman, 1982), or at least differing in difficulty in a regular fashion (Bachman, 1985). Despite these findings, the factors influencing item performance remain under investigation (e.g., Brown, 1988). While researchers continue to seek theoretical and statistical bases for cloze item performance, it is useful to note that, practically speaking, items selected by experienced test writers may produce tests that are more reliable and more highly correlated with other language tests, especially tests measuring traits similar to those that particular cloze items were chosen to measure.

The empirical findings related to these expectations have been mixed. The rational cloze procedure produced a test that was easier than the fixed-ratio cloze in Bachman's study (1985), while in Greene's (1965) research (using native speakers), overall test difficulty of the two were the same. Rational deletion procedures resulted in a test with higher reliability than its fixed-ratio counterpart in

[3] Cloze inconsistencies are even more apparent when summarized over a number of studies. In fact, J.D. Brown (1988) notes that reported internal consistency reliability estimates have ranged from .31 to .96 and correlations between cloze and other language tests have ranged from .43 to .91.

Greene's study (Split-half = .76 and .52, respectively), but not in Bachman's study (Split-half = .86 for both tests, for all subjects in the study). Bachman's (1985) comparison of correlations of fixed-ratio and rational cloze tests found the two to correlate comparably with six other language tests (rational, r = .62-.82.; fixed-ratio, r = .68-.81).

Despite these empirical results, on theoretical grounds the rational cloze procedure should have the advantage of allowing more consistent and controllable results to the extent that distinct item types can be understood and identified. As with the fixed-ratio cloze, a problem in characterizing the rational cloze as a test genre is the individual nature of each such test. In attempting to synthesize rational cloze research, one finds that the types of items used in various studies tend to be inequivalent; moreover, rational cloze tests differ from one study to another in their 'facet of expected response' (Bachman, 1990).

Multiple-choice

How does the multiple-choice response method affect students' performance on the cloze? Research has demonstrated that constructing a test response is more difficult for test takers than selecting one (Shohamy, 1984);[4] however, as Shohamy points out, finding that one method produces an easier test than another does not indicate which method is the more valid. An overall shift in difficulty alone, if not extreme, may not significantly alter the test's reliability or its convergent and discriminant correlations. Past research provides some comparative information between the two cloze methods. Cranney (1972) reports comparable reliabilities for multiple-choice and fill-in versions of a cloze test. Hale, Stansfield, Rock, Hicks, Butler, and Oller (1989) interpreted the research of Pike (1979) and Hinofotis and Snow (1978) to indicate 'a similarity of processes measured by the MC cloze and the completion cloze procedures' (p. 51), attributing the lower-than-expected correlations in the latter study to possible unreliability of measures. Bensoussan and Ramraz (1984) also report a lower-than-expected correlation (r = .43) between their multiple-choice cloze and completion cloze, but lacking reliabilities estimated for that sample, it is difficult to interpret the strength of that relationship. Ilyin, Spurling and Seymour (1987) report for their fill-in and multiple-choice cloze tests almost equal difficulty levels and a .76 correlation.

[4] As Porter (1976) points out, the difficulty of a multiple choice cloze is also a function of the alternatives provided for items.

Comparative questions aside, some data exist for adequate reliabilities of multiple-choice cloze tests and for reasonable correlations with other tests. The following reliabilities have been estimated for various multiple-choice cloze tests: KR-20 = .76 (Jonz, 1976); KR-20 = .82 & .84 (Bensoussan and Ramraz, 1984); KR-21 = .86 (Ilyin et al., 1987); adjusted reliabilities ranging from .88 to .94 (Hale et al., 1989). Strong correlations with reading tests have been hypothesized by Porter (1976) and Ozete (1977) who suggested that the multiple-choice cloze is similar to tests of reading comprehension, in other words, tests of written textual competence, requiring selected rather than constructed responses. Ilyin et al. (1987) found a slightly higher correlation of their multiple-choice cloze with their reading test (r = .77) than with their listening tests (r = .71 and .64), but the strongest correlation was found between the multiple choice cloze and their structure test (r = .81). When Jonz (1976) calculated correlations of the multiple-choice cloze with other language tests, the correlation with reading was not among the strongest: Composition, r = .80; Structure, r = .70; Reading, r = .61; Vocabulary, r = .54 and Aural, r = .29. Hale et al. (1989) found predictably stronger correlations between the multiple-choice cloze and the written-text portions (Structure, Written Expression, Vocabulary, and Reading Comprehension) of the TOEFL (r = .88; median across language groups) and weaker correlations between the cloze and the listening portion of the TOEFL (r = .77; median across language groups). However, they failed to find notable distinctions among correlations of the cloze with the four written text parts of the TOEFL.

These results appear to relate multiple-choice cloze performance to tests of written competence more clearly than has research on the fill-in cloze, indicating that this facet of test method may indeed affect not only test difficulty but also the language trait measured. However, because previous research has combined rational items and multiple-choice responses inconsistently, the hypothesis concerning the relationship between multiple-choice cloze and reading tests, or any other tests, requires additional support.

C-Test

The C-test, claimed, like the fixed-ratio cloze, to be a measure of global language proficiency, was proposed to solve several cloze problems (Klein-Braley, 1985). One major problem was the unpredictable results obtained by various fixed-ratio deletion procedures (e.g. Alderson, 1979). The every-other-word C-test procedure improves on the fixed-ratio cloze by producing a large number of 'random samples of the word classes of the text involved' (Klein-Braley,

1985: 84).[5] A second problem with the cloze, the effect of text topic and difficulty on test performance, is minimized by the C-test's use of several different short texts. A third cloze problem, the lack of criterion reference point defined by performance of educated native speakers, does not exist with the C-procedures; Klein-Braley (1985: 84) reports, 'Adult educated native speakers achieve virtually perfect scores.'

Indeed, these features of the C-test appear to improve on the psychometric properties of the cloze. However, like early research on the cloze, C-test research has failed to clarify evidence for the specific language traits that this technique may measure. This evidence must, then, be procured from descriptions of test items, analysis of the test task, and details of reported validity research. The method of C-test construction, designed to improve on cloze text and item sampling through the use of more and shorter passages, also has the effect of eliminating or at least reducing the number of cloze-type items which are governed by long-range constraints. The C-test requires students to fill in missing second halves of words. In completing a given word, the most important clue for the test taker is often in the immediate environment of the blank (Klein-Braley, 1985: 98), including the first half of the word itself. Error analysis of students' responses indicated that 'recognition of syntactical relationships comes first' in making responses, although semantic processing is essential for perfect performance (Klein-Braley, 1985: 100). On the basis of item and task description, then, the C-test appears to reflect more grammatical than textual competence.

Interpretations of validity research, on the other hand, argue that the C-test is a measure of overall language proficiency. Such arguments include the observation that C-test scores increase regularly and predictably with an individual's native language ability. However, this argument for the predictable increase in C-test performance with maturational linguistic development would be equally compelling with respect to maturational development of grammatical competence alone. A second argument is Klein-Braley's report of weak to moderate correlations between C-test performance and scores on nonverbal intelligence tests. She asserts that the increase of these correlations with the subjects' age is evidence for consistency of C-test results with theoretical expectations. An alternative interpretation is that correlation of a language test with a nonverbal ability indicates undesirable convergent relationships between constructs for which divergent relationships are predicted by theory. Such an

[5] The numbers of different word classes deleted by C-tests were calculated for over 100 English and 100 German texts (Klein-Braley, 1985).

argument is precisely the one made from similar research which found the cloze test related to the nonverbal characteristic, field independence (Stansfield and Hansen, 1983).

On the basis of the C-test research reviewed by Klein-Braley, it is clear that this procedure can produce results which are psychometrically superior to the cloze. Examination of test items and task analysis suggest the the C-test may be a measure of relatively grammatical competence, while validity research does not provide evidence for the specific traits it may measure and indicates problematic convergent relationships with a non-verbal measure. Thus, questions remain concerning exactly what the C-test measures and how it can be distinguished from comparable cloze tests.

In summary, cloze-type techniques produce tests that can measure, with some degree of accuracy, aspects of the students' written grammatical and/or textual competence. The accuracy of measurement and specific traits measured may depend on how deletions are made and the manner of students' expected response. This research sheds light on the effects of these methods by hypothesizing their theoretical impact on correlations with other tests and providing empirical evidence for the following question: How do cloze tests constructed from a single passage using the four procedures outlined above compare in difficulty, reliability, and convergent and discriminant correlation coefficients?

III Research design

Subjects

The subjects were 201 nonnative speakers of English (from a wide variety of language backgrounds) who were enrolled in intermediate and advanced ESL composition courses at Iowa State University in Fall 1985. All students had met the University's admission requirement of 500 on the TOEFL and were working toward degrees in a wide range of subject area across campus.

Measures

All four cloze tests (*Appendix A*) were constructed from a single text adapted from a *Scientific American* article entitled 'Compartmentalization of Decay in Trees' (Shigo, 1985). The academic genre of this text is typical of the type these students are likely to encounter in their classes at Iowa State, yet it is not overly technical and dependent on previous knowledge. It is apparent that the author intended a general audience, as the technical terms used are defined within the text.

The fixed-ratio cloze was constructed by deleting every 11th word – an arbitrarily chosen rate – from the text following the first two sentences, which were left intact. An *a posteriori* analysis (using the item classifications introduced by Perkins and German, 1985) revealed that the procedure had resulted in four items relying on clues in the immediate context, 12 items relying on clues within the same clause, six items relying on clues beyond the clause but within the sentence, and 13 items relying on clues beyond the sentence.

The same contextual categories were used to explicitly choose items for the rational cloze. The number of each type of item for the rational cloze was approximately the same as the ones for the fixed-ratio cloze (3, 13, 5, 14, respectively) so comparisons could be made between deliberate and chance deletions without radically confounding item type as defined by context clues. The major difference, then, between the fixed ratio and rational cloze was that for the latter we chose each item as having clearly identifiable clues in the passage.

The multiple-choice cloze had exactly the same words deleted as the rational cloze. For each blank, four alternatives were given; in most cases the three distractors were the same part of speech as the correct answer.

The C-test was constructed based on the instructions given by Klein-Braley and Raatz (1984) and Klein-Braley (1985). For each of the five paragraphs of suitable length, a short intact introduction was provided, followed by the deletion of the second half of every second word. Fifteen such deletions were made for each of the five paragraphs and then the rest of the paragraph was presented intact. This procedure differed from the proto-typical C-test which would have used paragraphs from different texts for each of the five parts. For the purpose of this study, it was necessary to keep the text constant across the four test-construction methods to make valid method comparisons.

The Iowa State University English Placement Test (EPT) has three multiple-choice parts: listening (35 items), reading (35 items), and vocabulary (30 items). The listening section, with aural questions about spoken segments of text, is considered a test of both grammatical and textual competence. Some items require students to discern the grammatical details of what they heard while others require overall comprehension of discourse. The reading test consists of short passages followed by multiple-choice questions intended to test discourse comprehension (i.e., textual competence). The vocabulary items require students to select the correct university-level vocabulary word to fit into a one-sentence defining context. The

focus on specific lexical items within a limited context makes this test primarily a grammatical one, in Bachman's sense. These tests have been used successfully over the past decade for making rough distinctions among students. KR-20 reliabilities for the whole group taking the tests each semester are adequate (above .80). However, because subjects in this study consisted only of low scorers who, after taking the tests, were required to take an ESL class, the variance in this sample was reduced; consequently the sample reliabilities are relatively low. (See Appendix B for reliabilities for all tests.)

The writing test required students to compose an essay on the topic, 'describe something you have learned in one of your non-English courses this semester.' Each composition was rated by two ESL instructors (other than the student's own) using Jacobs, Zinkgraf, Wormuth, Hartfiel and Hughey's (1981) ESL Composition Profile, which produces a score between 34 and 100.

The Group Embedded Figures Test (GEFT) by Oltman, Raskin, and Witkin (1971) was used as a measure of field independence. The GEFT consists of a booklet containing 18 'complex figures' within which subjects are asked to find a given 'simple figure'. Their ability to identify the simple figure from the distracting context of the complex figure is used as a measure of their field independence. One point is given for each correctly identified simple figure, producing scores from 0 to 18. The GEFT, used extensively in second language research, is not without problems (see Brown, 1987; Chapelle, 1988); however, it was chosen because of the unanswered question of whether it, a nonverbal measure, would produce the desirable discriminant correlations with cloze tests.

Procedures

Most students took the listening, reading, and vocabulary tests at the beginning of Fall semester, 1985. Twenty-one (about 10%) had taken the same tests two and a half months earlier. All subjects wrote the composition during the eighth week of the semester. Within the next three weeks, each student was given a cloze test and the GEFT during one class period. To distribute the four types of cloze tests, the same type of cloze was assigned to every fourth student on alphabetized class lists, so approximately equal numbers of students took each type. Students were given 25 minutes to complete their assigned cloze tests; the GEFT was administered according to the procedures given in the test manual. Verification of the desired similarities among the four groups was obtained from the results of an ANOVA comparing the scores on all tests (except the cloze). None of the small observed differences among groups was statistically significant.

Analysis

SPSS-X (SPSS-X, Inc.) was used to estimate KR-20 reliabilities for the four cloze tests and the GEFT. Reliability for the composition was estimated by calculating the correlation between raters, and then applying the Spearman-Brown Prophecy correction for use of two raters (Thorndike and Hagen, 1977: 90–91). Reliabilities for the listening, reading, and vocabulary tests were estimated using the KR-21 formula for practical reasons: it was not possible to recover the item data for the sample. SPSS-X was used to perform an ANOVA with Sheffé follow-up tests indicating differences between cloze tests, and to perform correlations among all tests. To allow accurate comparisons of correlations, all were corrected for attenuation (Thorndike and Hagen, 1977: 101). Disattenuated correlations among all tests are given in Appendix C.

IV Results

How does the difficulty of the four types of cloze tests compare?

The fixed-ratio and rational tests were predicted to be the most difficult and the multiple-choice cloze the easiest. The C-test, hypothesized to measure relatively more grammatical competence, but requiring the student to construct a response, should be easier than the two cloze tests, but more difficult than the multiple-choice cloze. Of the two most difficult cloze tests, the fixed-ratio was predicted to be the more difficult because no deliberation over items took place during test construction in contrast to the rational cloze in which items with clearly identifiable clues were selected. An ANOVA with a Scheffé follow-up test comparing the tests' percentage scores found the differences among mean scores, which fell in the predicted order, to be statistically significant (fixed-ratio, $\bar{x} = 41.7\%$; rational, $\bar{x} = 50.9\%$; multiple-choice, $\bar{x} = 82.3\%$; C-test, $\bar{x} = 61.8\%$; $F = 79.3$; $p < .001$)

How did the reliabilities of the four tests compare?

Estimation of reliability for cloze tests is theoretically problematic because of the interdependence of cloze items; however, Brown (1983) concluded that in practice this problem is negligible. Moreover, for the purpose of comparing the cloze reliabilities in this study, if there is an overestimation problem, it should affect the three word-deletion cloze tests equally. The interdependence of C-test items, however, requires reliability to be calculated in the manner described

by Raatz (1985), which treats each segment of text with a series of blanks as one item (referred to as a 'super item'). The C-test used in this study had 5 'super items', each scored from 0 to 15. Lower reliabilities were predicted for tests with the most extreme difficulties: the fixed-ratio and multiple-choice tests. These predictions were accurate, but the differences among reliabilities were not great: fixed-ratio, KR-20 = .76; rational, KR-20 = .80; multiple-choice, KR-20 = .76; C-test, KR-20 = .81.

How do the convergent and discriminant correlations of the four tests compare?

Messick (1989) terms convergent and discriminant evidence for the validity of a test 'the external component of construct validity' which 'refers to the extent to which the test's relationships with other tests . . . reflect the expected high, low, and interactive relations implied in the theory of the construct being assessed' (p. 45). Theory indicates that these relations are affected by two sources of variance: that attributable to the trait measured by the test and that associated with the test method. Bachman's (1990) theoretical foundations – asserting the relevance of traits and methods for test performance – suggest that such a framework 'may provide language testers with an appropriate means of codifying and describing, at a very useful level of detail, the tests they are developing, using, or researching, for purposes of improved . . . communication within the field of language testing' (Bachman, 1990: 154). Accordingly, aspects of Bachman's overall framework are used here to predict convergent and discriminant relationships of each of the cloze tests with the listening, reading, vocabulary, writing, and GEFT tests.

The language tests are sufficiently narrow in the traits they measure to be characterized by using only one component of Bachman's definition of language competence and by distinguishing aural from written competence. Within Bachman's 'organizational competence' exists what we shall consider a continuum from grammatical competence (vocabulary, morphology, etc.) to textual competence (cohesion and rhetorical organization). To clarify which of these traits each test was hypothesized to measure, each language test was placed at the appropriate position along the three-point continuum under either written or aural components (Table 1). Under written, the reading and writing tests are most textually based; vocabulary is most grammatically based. The listening test, the only one under aural trait, was placed in the middle because it addresses specific grammar points with some items, while other segments require comprehension of longer discourse. The three word-deletion cloze tests were placed at

Table 1 Hypothesized sources of trait variance

Part A Similarities and differences between the cloze tests (fixed-ratio, rational, and multiple-choice) and the other tests

TRAITS		**Language** *(same)*					**Non-language** *(different)*
		written *(same)*			**aural** *(same-)*		
Organizational Competence	**Grammatical←** *(same-)*	*(same)*	**→Textual** *(same-)*	**Grammatical←** *(same--)*	*(same-)*	**→Textual** *(same--)*	
	Vocabulary	CLOZE (F-R) CLOZE (RAT) CLOZE (M-C)	Reading Writing		Listening		GEFT

Part B Similarities and differences between the C-test and the other tests

TRAITS		**Language** *(same)*					**Non-language** *(different)*
		written *(same)*			**aural** *(same-)*		
Organizational Competence	**Grammatical←** *(same)*	*(same-)*	**→Textual** *(same--)*	**Grammatical←** *(same-)*	*(same--)*	**→Textual** *(same---)*	
	C-TEST Vocabulary		Reading Writing		Listening		GEFT

Table 2 Hypothesized sources of method variance: similarities and differences of input format between the cloze/C-tests and the other tests

FORM OF PRESENTATION	**language** *(same)*		**non-language** *(different)*
Channel & Mode	**Visual** *(same)*	**Aural** *(same-)*	
	Vocabulary Reading Writing CLOZEs	Listening	GEFT

the written mid-point and the C-test at the grammatical end for reasons described above. A non-language trait was distinguished from the language traits and the GEFT, a non-language measure, was placed under that heading.

With this designation of the traits hypothesized to be measured by each test, it was possible to estimate the degree of similarity or difference between the cloze and the other tests. These estimations are marked under each trait in Table 1^A using the notation 'same', 'same−' (same minus) and 'different'. Any tests which fell under the 'same' category would be hypothesized to measure the same trait as the cloze; 'same−' indicates that a test measures close to the same trait as the cloze. Additional minuses added to 'same' indicate a greater trait difference between that test and the cloze. The three word-deletion cloze tests displayed in Table 1^A are presumed to measure the same trait so their trait similarities and differences with other tests should be equivalent. Table 1^B displays the similarities and differences between the other tests and the C-test. On the basis of trait similarities alone, different patterns of correlations are predicted for the C-test than for the word-deletion cloze tests. In keeping with Bachman's theory, however, method facets should also be considered.

Two facets of test method were taken into account: input format and the format of expected response. Input format refers to how the student receives information while taking a test. Input can be received either through the use of language or non-language material, the values for the parameter 'form of presentation' in Table 2. Language input can be further subdivided on the basis of its channel and mode into primarily visual, and primarily aural, represented by the two descriptors under 'language'. In Table 2, each test is listed at its appropriate position and, using the same notation described for Table 1, similarities and differences with the cloze tests and the C-test are marked.

A second method facet, format of the expected response, can also

Table 3 Hypothesized sources of method variance: expected response

Part A Similarities and differences between the fixed-ratio/rational/C-test and the other tests

FORM OF OUTPUT		**Language** *(same)*		**Non-language** *(different)*
Type of output	**Constructed** *(same)*		**Selected** *(same--)*	
Length of output				
	word *(same)*	**text** *(same-)*		
	CLOZE (F-R) CLOZE (RAT) C-TEST	Writing	Vocabulary Reading Listening	GEFT

Part B Similarities and Differences between the multiple-choice cloze and the other tests

FORM OF OUTPUT		**Language** *(same)*		**Non-language** *(different)*
Type of output	**Constructed** *(same-)*		**Selected** *(same)*	
Length of output				
	word *(same-)*	**text** *(same-)*		
		Writing	Vocabulary Reading Listening CLOZE (M-C)	GEFT

Table 4 Estimation of relative similarities between other tests and the fixed-ratio and rational cloze tests

Test		trait		input		response		
listening	→	same-	+	same-	+	same--	=	same----
reading	→	same-	+	same	+	same--	=	same---
vocabulary	→	same-	+	same	+	same--	=	same---
writing	→	same-	+	same	+	same-	=	same--
GEFT	→	different	+	different	+	different	=	different

be distinguished as 'language' or 'non-language' as indicated in Table 3. When the form is 'language', the type can be either 'constructed' when the student is expected to produce some language, or 'selected' when the student is expected to choose an answer. Constructed language can differ in its length, with a test such as a cloze or C-test requiring students to construct a word or less, and a writing test requiring students to construct an entire essay. Table 3 indicates where each of the tests is placed with respect to its method of expected response. Table 3^{A} marks each category for its similarity to or difference from the fixed-ratio cloze, the rational cloze and the C-test; Table 3^{B} does the same for the multiple-choice cloze test.

On the basis of these three sources of variance – one trait and two method facets – estimations were made of the relative degree of similarity between each cloze test and the other measures. These estimations resulted from adding, for each of the cloze tests, the 'sames', 'same minuses' and the 'differents' for each test on each of the three dimensions. For example, relative relationships of other tests with the fixed-ratio and rational cloze tests are estimated as demonstrated in Table 4. On the basis of this analysis, the performance on the writing test (same--) should be most highly correlated with the fixed-ratio and rational cloze tests, vocabulary and reading (same---) should be tied for second, listening (same----) should be third, and the GEFT (different) should be uncorrelated. By adding the similarities and differences of the tests for each cloze test, predictions for relative correlations were obtained, thereby allowing for a theoretically motivated interpretation of the obtained correlations.

Table 5 displays the estimated ranks of correlations and the comparison between the predicted and actual disattenuated correlations of the cloze tests with the other tests. In the column labeled 'Predicted Rankings', the test named at the top of each list is the one predicted to have the the strongest correlation with the cloze test indicated.

Table 5 Predicted and actual relationships of the cloze tests to the other tests

Estimates	Predicted rankings	Actual correlations	
Fixed-ratio cloze with:			
(same--)	writing	writing	.621
(same---)	vocabulary/reading	vocabulary	.490
(same----)	listening	*reading*	.380
		listening	.296
(different)	GEFT	GEFT	.026
Rational cloze with:			
(same--)	writing	*reading*	.767
		vocabulary	.695
(same---)	vocabulary/reading	*writing*	.659
(same----)	listening	listening	.338
(different)	GEFT	*GEFT*	.293
Multiple choice cloze with:			
(same-)	vocabulary/reading	reading	.862
(same--)	listening/writing	writing	.433
		listening	.366
(different)	GEFT	*GEFT*	.226
		vocabulary	180
C-test with:			
(same--)	vocabulary	vocabulary	.836
(same---)	writing	writing	.639
(same----)	reading	*reading*	.604
(same-----)	listening	listening	.472
		GEFT	.399
(different)	GEFT		

Tests predicted to have the same relative correlation (e.g., vocabulary and reading with the fixed-ratio cloze) are on the same line. The actual correlations are also listed with the strongest one on the top, and clustered (i.e., without a blank line between them) when they were close to one another (i.e., less than .1 different). The typeface of the tests listed under 'Actual Correlations' indicates accuracy of prediction. Regular type denotes perfect prediction of rank, while italics indicates slight differences between predicted and actual correlations (differences in the way that correlations are clustered or their order within a cluster). Bold indicates major discrepancies in the

order of predicted and actual correlations. These results are discussed in terms of the agreement between predicted and actual convergent and discriminant relationships as well as the absolute strengths of correlations.

With respect to relative convergent and discriminant relationships, the fixed-ratio cloze conforms most closely to its expectations. The correlation with the writing test is the strongest, followed by vocabulary; the correlation with the reading test is slightly lower than predicted, but the listening ranks fourth and, as predicted, there is no correlation between the cloze and the GEFT. In fact, the fixed-ratio cloze is the only one for which the theoretically-predicted lack of correlation with the GEFT appears. However, despite the consistency of the predictions with the results, the fixed-ratio cloze, overall, correlates most poorly with the language measures; its highest correlation with a language measure is $r = .621$, the next is $r = .490$, and the correlation with the reading test is only $r = .380$. Despite the desirable lack of correlation with the GEFT, it would be difficult to argue for the fixed-ratio cloze as a clear measure of any of these traits given the moderate to low correlations obtained with other language measures.

Even with approximately the same item types as the fixed-ratio cloze, the rational cloze contrasts in its strength of correlations with the written language tests: $r = .767$ to $r = .659$. With these language test correlations, however, also appears a higher-than-zero correlation with the GEFT. Although this correlation is greater than one would predict in absolute terms, the observed correlations fall close to their predicted order, with less discrimination among the written-text correlations than predicted. All three were moderately high rather than the writing test correlation being higher than the reading and vocabulary tests. On the basis of these results, then, the rational cloze demonstrates moderately high, fairly predictable correlations with other tests, providing evidence for the superiority of carefully selected items over items selected by their positions in the text alone.

These benefits of the rational cloze did not appear when it was modified into a multiple choice format. While the correlations with reading correspond to expectations, the multiple choice cloze has the pattern of relationships most deviant from predictions. The strangest of these correlations is the one with the vocabulary test, $r = .180$, a correlation which is lower than that between the multiple choice and the GEFT. One might suspect something strange about the vocabulary test for this sample; however, that suspicion is allayed somewhat by the correlation of $r = .784$ between the vocabulary and the reading test, one similar to that of the other groups. (See Appendix C.) This multiple choice cloze, then, is apparently a poor measure of

vocabulary relative to its fill-in counterpart, which correlates with the vocabulary test at a predictable .695, despite the method differences between these two tests. The absolute strengths of correlations provide evidence for the hypothesis (Porter, 1976) that the multiple choice cloze may be a measure of reading comprehension. Empirical evidence supporting this assertion has not been obtained from previous research, in which the correlation between multiple choice cloze and reading has been about the same as that between the cloze and other measures (Hale et al., 1989) or less (Jonz, 1976; Ilyin et al., 1987). However, here the correlation obtained between the multiple-choice cloze and the reading test was the highest one obtained in the study ($r = .862$). The fact that the multiple choice and rational cloze tests were exactly the same except for their facet of expected response offers clear evidence for the effect of this method facet on convergent correlations.

The C-test's observed correlations were quite close to those expected, its correlation with reading being slightly higher, and with the GEFT considerably higher than predicted. While the correlation with the GEFT is relatively the lowest, in absolute terms it is not much lower than that with the listening test. The observed strong correlation between the C-test and the vocabulary test (the most grammatically-based test) provides empirical support for placing the C-test at the 'grammatical' end of the 'grammatical-textual' continuum, as suggested earlier. However, the C-test shares considerable variance with the other less grammatically-based tests, with none below $r = .472$.

V Conclusion

Examining the cloze as a measure of second language traits, this study described results obtained from altering trait and method facets of the cloze procedure while holding text and student ability constant. The data substantiated predictions of the fixed-ratio as the most difficult and the multiple choice as the easiest of these methods. The fact that the rational and multiple choice were exactly the same tests except for their format of expected response pinpoints this facet as a determinant of difficulty level. Although statistically significant, these differences in difficulty were not sufficiently large to affect reliability substantially. The reliabilities were adequate, although not as high as desired. The homogeneity of this advanced-level group had the predictable effect of yielding only moderate cloze reliabilities.[6]

[6] It should be noted that these moderate cloze reliabilities are consistently higher than those obtained with this group for the other language tests used in this study.

Differences in cloze methods had striking effects on their external relationships, suggesting directions for further investigation. The multiple-choice cloze was strongly related to the reading test, but not to the other language tests, including the vocabulary test. Why did the multiple choice cloze measure traits so similar to those measured by the reading test but so different from those measured by the vocabulary test? Facet of expected response apparently accounted for differences between the multiple choice and rational cloze tests, but why did the fixed-ratio and rational cloze produce different strengths of correlations with the other tests? Is it possible to explain these differences by comparing characteristics of explicitly selected items with those of items chosen by their position in the text or are these simply to be added to the many documented cases of cloze inconsistencies? Analysis of items by the amount of context needed to complete blanks does not account for test differences, since the number of items at each context level in the two tests was almost the same; therefore, other explanations are needed for the differences between these 'parallel' tests. The C-test, correlating most strongly with the vocabulary test, produced, on average, the highest correlations with the language tests. Why did this apparently more grammatically-based test correlate so well with written text-based tests – even better than the fixed-ratio cloze? Why did the GEFT correlate more strongly than theory predicts with the three cloze tests which correlated well with language measures? Can we learn more about cloze/GEFT correlations by identifying particular types of items with which performance on the GEFT is associated? The consistent relationship of cloze performance with the non-verbal characteristic, field independence, is worthy of further investigation and interpretation.

Along with the empirical research directions suggested by these results are observations concerning the viability of systematizing theoretical assumptions about language tests to make predictions of relationships among tests. The framework we devised for making hypotheses about expected correlations was our first attempt at formalizing and manipulating elements from Bachman's overall framework. As a first attempt, and as one which defines categories for some noncategorical constructs, it is crude and incomplete in some respects. Despite its shortcomings, we found it very useful to systematize our intuitions of expected correlational outcomes. This calculated development of hypotheses pinpointed effects of components of test variance, thereby helping us to isolate components we wished to consider. Generation of clear-cut hypotheses, moreover, facilitated discussion of correlational results because comparisons with predictions were possible as opposed to the familiar *ad hoc* explanations. There is much that could be added, rethought and reshuffled in this

initial scheme; on the basis of our work with it, we foresee that time spent on further developments will be time well spent.

VI References

Alderson, J.C. 1979: The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13, 219–27.

Alderson, J.C. 1980: Native and nonnative speaker performance on cloze tests. *Language Learning*, 30, 59–76.

Alderson, J.C. 1983: The cloze procedure and proficiency in English as a foreign language. In Oller, J., editor. *Issues in Language Testing Research*. Rowley, Mass.: Newbury House, 205–17.

Alderson, J.C. and **Urquhart, A.H.** 1985a: The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing*, 2(2), 192–204.

Alderson, J.C. and **Urquhart, A.H.** 1985b: This test is unfair: I'm not an Economist. In Hauptman, P., LeBlanc, R. and Wesche, M., editors. *Second Language Performance Testing*. Ottawa: University of Ottawa Press, 25–43.

Bachman, L. 1982: The trait structure of cloze test scores. *TESOL Quarterly* 16, 61–70.

Bachman, L. 1985: Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly*, 19(3), 535–56.

Bachman, L. 1990: *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bensoussan, M. and **Ramraz, R.** 1984: Testing EFL reading comprehension using a multiple-choice rational cloze. *Modern Language Journal* 68, 230–39.

Brown, H.D. 1987: *Principles of Language Learning and Teaching, second edition*. Englewood Cliffs, NJ: Prentice-Hall.

Brown, J.D. 1983: A closer look at cloze: validity and reliability. In Oller, J.W., editor. *Issues in Language Testing Research*. Rowley, Mass.: Newbury House: 237–50.

Brown, J.D. 1984: A cloze is a cloze is a cloze? In Handscombe, J., Orem, R. and Taylor, B., editors. *On TESOL 83*. Washington DC.: TESOL Publications, 109–19.

Brown, J.D. 1988: What makes a cloze item difficult? *University of Hawai'i Working Papers in ESL*, 7(2), 17–39.

Chapelle, C. 1988: Field independence: A source of language test variance? *Language Testing* 5, 62–82.

Chavez-Oller, M.A., Chihara, T., Weaver, K.A., and **Oller, J.** 1985: When are cloze items sensitive to constraints across sentences? *Language Learning* 35, 63–73.

Cranney, A.G. 1972: The construction of two types of cloze reading tests for college students. *Journal of Reading Behavior* 5, 60–64.

Greene, F.P. 1965: Modifications of the cloze procedure and changes in reading test performances. *Journal of Educational Measurement*, 2, 213–217

Hale, G., Stansfield, C., Rock, D., Hicks, M., Butler, F., and **Oller, J.** 1989: The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing* 6, 49–78.

Hanania, E. and **Shikhani, M.** 1986: Interrelationships among three tests of language proficiency: Standardized ESL, cloze, and writing. *TESOL Quarterly*, 20, 97–109.

Hansen, J. and **Stansfield, C.** 1981: The relationship of field dependent-independent cognitive styles to foreign language achievement. *Language Learning*, 31(2), 349–67.

Hansen, L. 1984: Field dependence-independence and language testing: Evidence from six Pacific island cultures. *TESOL Quarterly* 18, 311–24.

Hinofotis, F. and **Snow, B.G.** 1980: An alternative cloze testing procedure: multiple-choice format. In Oller, J.W. and Perkins, K., editors, *Research in Language Testing*. Rowley, MA: Newbury House.

Ilyin, D., Spurling, S., and **Seymour, S.** 1987: Do learner variables affect cloze correlations? *System*, 15, 149–160.

Irvine, P., Atai, P, and **Oller, J.** 1974: Cloze, dictation and the test of English as a foreign language. *Language Learning*, 24, 245–252.

Jacobs, H.L., Zinkgraf, S.A., Wormuth, D.R., Hartfiel, V.F., and **Hughey, J.B.** 1981: *Testing ESL Composition: A Practical Approach*. Rowley, MA: Newbury House.

Jonz. J. 1976: Improving on the basic egg: the M-C cloze. *Language Learning*, 26, 255–65.

Klein-Braley, C. 1983. A cloze is a question. In Oller, J.W., editor. *Issues in language testing research*. Rowley, MA: Newbury House.

Klein-Braley, C. and **Raatz, U.** 1984: A survey of research on the C-test. *Language Testing* 1, 134–46.

Klein-Braley, C. 1985: A cloze-up on the C-Test. *Language Testing* 2, 76–104.

Lado, R. 1986: Analysis of native speaker performance on a cloze test. *Language Testing*, 3, 2 130–46.

Markham, P. 1987: Rational deletion cloze processing strategies: ESL and native English. *System*, 15, 303–11.

McKenna, F.P. 1984: Measures of field dependence: cognitive style of cognitive ability. *Journal of Personality and Social Psychology* 47, 593–603.

Messick, S. 1989: Validity. In Linn, R., editor, *Educational measurement*. NY: Macmillan.

Oller, J.W. and **Conrad, C.A.** 1971: The cloze technique and ESL proficiency. *Language Learning*, 21, 183–95.

Oller, J.W. 1979: *Language Tests at School*. NY.: Longman.

Oller, J.W. 1983: Evidence for a general language proficiency: an expectancy grammar. In Oller, J.W., editor. *Issues in Language Testing Research*. Rowley, Mass.: Newbury House.

Oltman, P., Raskin, E. and **Witkin, H.** 1971: *Group Embedded Figures Test*. Palo Alto, CA.: Consulting Psychologists Press.

Ozete, 1977: The cloze procedure: A modification. *Foreign Language Annals*, 10, 565–68.

Pike, L.W. 1979: *An evaluation of alternative item formats for testing English as a foreign language*. TOEFL Research Report No. 2, Princeton, New Jersey: Educational Testing Service.

Perkins, K. and **German, P.** 1985: The effect of information gain on different structural category deletions in a cloze test. Paper presented at Midwest TESOL, Milwaukee, WI, October 17–19.

Porter, D. 1976: Modified cloze procedure: a more valid reading comprehension test. *English Language Teaching*, 30, 151–55.

Raatz, U. 1985: Better theory for better tests? *Language Testing* 2, 60–75.

Shigo, A.L. 1985: Compartmentalization of decay in trees. *Scientific American* 252, 96–103.

Shohamy, E. 1984: Does the testing method make a difference? the case of reading comprehension. *Language Testing* 1, 147–70.

Shanahan, T., Kamil, M., and **Tobin, A.** 1982: Cloze as a measure of intersentential comprehension. *Reading Research Quarterly* 17, 229–55.

SPSS Inc. SPSS-X *User's Guide 3rd Edition*. Chicago: SPSS Inc.

Stansfield, C. and **Hansen, J.** 1983: Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly*, 17, 29–38.

Thorndike, R.L. and **Hagen, E.P.** 1977: Measurement and Evaluation in Psychology and Education, Fourth Edition. NY: John Wiley & Sons.

Appendix A: Cloze text – compartmentalization of decay in trees

Trees have a spectacular survival record. Over a period of more than 400 million *years* (2)* they have evolved as the tallest, most massive and longest-lived organisms ever to inhabit the earth. Yet [trees lack a means of defense *that* (2) almost every *animal* (1) has: trees cannot move away *from* (2) destructive forces. Because they *cannot* (1) move, all types of living and non-living]** enemies – fire, storms, *microorganisms* (1), insects, animals and man – have wounded *them* (2) throughout their history. *Trees* (1) (2) have survived *because* (2) their evolution has made them into highly *compartmentalized* (1) organisms; that is, they wall off injured and infected wood.

In (1) (2) that respect trees are radically different from animals. Fundamentally, [animals *heal* (1): they preserve their lives by making billions of repairs, installing *new* (1) cells or rejuvenated cells in the positions of old *ones* (2). *Trees* (1) cannot heal; they make] no *repairs* (2). Instead, they defend themselves *against* (1) consequences of injury and infection by walling *off* (2) the damage. *In* (1) a word, they compartmentalize. At *the* (2) same time they put *new* (1) cells in new positions; in effect, they grow a new *tree* (1) over the *old* (2) one every year. The most obvious results *of* (1) the process are growth rings, which are visible on the *cross* (1) section of a trunk, *a* (2) root or a branch.

---

* Italicized words represent blanks in one or more of the clozes. Blanks designated as (1) appear in the fixed-ratio cloze; those designated as (2) appear in the rational and rational multiple-choice cloze.

** Deletions for the C-test begin with the first word in each bracketed section and continue with every second word to the end of the section.

Trees *have* (1) been guided through evolution by their need to defend against *attack* (1) while standing their ground. They always defend by *compartmentalizing* (2): they *attempt* (1) to wall off the injured or infected region.

After (2) a *tree* (1) has been injured, microorganisms can *infect* (2) the wound in several *ways* (1). Some [bacteria infect inner bark and stay *there* (2), creating diseases *known* (1) as annual cankers. *Other* (2) bacteria move into the tree's wood, *causing* (1) so-called wound rots. Still other microorganisms begin by] infecting the *inner* (1) bark *and* (2) then move in to *infect* (2) the wood as *well* (1). Finally, some microorganisms attack the wood first and *then* (2) move *to* (1) infect the inner bark *too* (2). Trees neither kill *nor* (2) arrest *the* (1) activity of these microorganisms. Nor do they respond in specific *ways* (1) (2) to specific microorganisms; the compartmentalization comes in response to the *fact* (1) of the injury.

Broadly *speaking* (2), the tree makes three responses *to* (1) injury and infection. In [the first of them, the boundaries of (1) compartments already in place are strengthened to resist the spread *of* (1) *infection* (2). *In* (2) the second, the *tree* (2) creates a new wall *by* (1)] anatomical and chemical means. The *third* (2) response the tree makes *is* (1) to continue growing. Trees survive injury and infection *if* (2) they *have* (1) enough time, energy and genetic capacity *to* (2) recognize and compartmentalize *injured* (1) and infected tissue while generating the new tissue that will maintain the *life* (2) of the tree.

This new understanding of trees as compartmentalizing organisms did not arise long ago. Indeed, [it came as a contradiction of earlier notions, some of *which* (2) were developed soon after the foundations of modern biology were established a century ago. It seems a] trite thing to say, but trees are fundamentally different from *animals* (2), and much of the failure to understand trees derives from unconsciously confusing the *two* (2).

Appendix B Descriptive statistics and reliability* estimates for all tests for the four groups

GROUP	Cloze	Listen	Vocabulary	Reading	Writing	GEFT
FIXED-RATIO						
n	53	53	53	53	52	53
$\bar{x}$	14.6	21.2	23.0	23.4	78.2	13.0
s.d.	4.9	4.3	3.7	4.9	9.6	4.9
reliability	.76	.56	.63	.70	.85	.91
RATIONAL						
n	50	50	50	50	49	50
$\bar{x}$	17.8	21.8	23.4	23.9	78.9	13.4
s.d.	5.6	4.4	3.8	4.9	7.8	3.9
Reliability	.80	.59	.67	.70	.76	.84
MULTIPLE CHOICE						
n	49	49	49	49	49	49
$\bar{x}$	28.8	22.4	23.4	23.5	77.7	13.9
s.d.	3.9	4.4	3.0	4.5	7.3	3.5
Reliability	.76	.60	.44	.64	.68	.81
C-TEST						
n	49	49	49	49	49	49
$\bar{x}$	46.6	20.7	23.4	23.6	78.4	13.8
s.d.	10.5	4.7	3.3	5.1	7.6	4.1
Reliability	.81	.64	.55	.73	.72	.87

* Reliability for writing calculated by correlations between raters corrected by the Spearman-Brown Prophecy Formula. Cloze and GEFT reliabilities calculated by KR-20. Listening, Vocabulary, and Reading reliabilities calculated by KR-21.

Appendix C Disattenuated correlations between all tests for the all groups

GROUP	Cloze	Listen	Vocabulary	Reading	Writing	GEFT
FIXED-RATIO	---					
Listen	.296	---				
Vocabulary	.490	.030	---			
Reading	.380	-.013	1.000	---		
Writing	.621	.128	.765	.583	---	
GEFT	.026	-.052	.034	.293	.073	---
RATIONAL	---					
Listen	.338	---				
Vocabulary	.695	-.000	---			
Reading	.767	.490	.654	---		
Writing	.659	.218	.776	.727	---	
GEFT	.293	.038	.012	-.069	-.238	---
MULTIPLE CHOICE	---					
Listen	.366	---				
Vocabulary	.180	.241	---			
Reading	.862	.386	.784	---		
Writing	.433	.759	.031	.487	---	
GEFT	.226	.000	-.089	.158	.016	---
C-TEST	---					
Listen	.472	---				
Vocabulary	.836	.320	---			
Reading	.604	.293	.685	---		
Writing	.639	.579	.263	.484	---	
GEFT	.399	-.280	.400	.122	0.87	---