the ESPecialist, vol. 25, nº 2 (221-243) 2004 ISSN 0102-7077

SCHEMA-BASED CLOZE MULTIPLE CHOICE ITEM TESTS: MEASURES OF REDUCED REDUNDANCY AND LANGUAGE PROFICIENCY*

Testes Cloze de Múltipla Escolha Baseados em Esquemas: Medidas de Redundância Reduzida e Proficiência Lingüística

Ebrahim KHODADADY (Urmia University)

Abstract:
This article reports the performance of 34 non-native speakers on four testing methods developed on the basis of reduced redundancy testing, namely a C-Test, schema-based cloze multiple choice item test (MCIT), text-driven cloze test and traditional cloze MCIT. A disclosed Test of English as a Foreign Language (TOEFL) was used as a criterion for estimating the empirical validity of the four methods and whether they tap into the same competence. The results indicated that among the four methods, schema-based and traditional cloze MCITs had the highest empirical validity. The principal component analyses of results revealed two factors: the ability to deal with cloze-based reduced redundancy and reduced redundancy in general. When the factors were rotated, only the C-Test and text-driven cloze test loaded on the first factor and thus validated themselves as measures of cloze-based reduced redundancy. The four methods, however, loaded on the second factor which was concluded to represent reduced redundancy in general. The addition of the TOEFL and rotation of the results revealed the same pattern. It also showed that the TOEFL had the highest loading on the reduced redundancy in general and thus validated reduced redundancy as a measure of language proficiency. Since the schema-based cloze MCIT had the second highest loading on the second factor, it provides the best measure of reduced redundancy and language proficiency.

Key-words: *cloze tests; C-Tests; language proficiency; multiple choice item tests (MCITs); reduced redundanc; schema theory; schema-based MCITs.*

Resumo:
Este artigo relata o desempenho de 34 falantes não-nativos em quatro métodos de teste desenvolvidos com base em redundância reduzida, a saber: Teste-C, teste cloze de múltipla escolha baseado em esquema (MCIT, conforme iniciais em inglês), teste cloze baseado em texto e MCIT cloze tradicional. Um teste de inglês como língua estrangeira (TOEFL) foi utilizado como critério para calcular a validade empírica dos quatro métodos e para avaliar se usam a mesma competência. Os resultados indicaram que, dentre os quatro métodos, os MCITs baseados em esquema e os testes cloze tradicionais tiveram a maior validade empírica. A análise de componente principal dos resultados revelou dois fatores: a habilidade de lidar com redundância reduzida baseada em cloze e redundância reduzida de maneira geral. Quando os fatores foram rotacionados, apenas o Teste-C e o teste cloze baseado em texto foram carregados no primeiro fator, validando-se, assim, como medidas de redundância reduzida com base em cloze. No entanto, os quatro métodos foram carregados no segundo fator, o que nos permite concluir que tal fator representa a redundância reduzida de maneira geral. A adição do TOEFL e a rotação dos resultados revelaram o mesmo padrão. Também mostraram que o TOEFL possuía a maior carga na redundância reduzida de forma geral, validando, assim, a redundância reduzida como medida de proficiência lingüística. Como o MCIT cloze baseado em esquema teve a segunda maior carga no segundo fator, conclui-se que fornece a melhor medida de redundância reduzida e proficiência lingüística.

Palavras-chave: *testes cloze; Testes-C; proficiência lingüística; testes de múltipla escolha (MCITs); redundância reduzida; teoria de esquemas; MCITs baseados em esquemas.*

1. Introduction

Reduced redundancy tests are developed on the assumption that knowledge of language requires understanding a distorted message (Spolsky, 1973). This assumption has been exploited in applied

linguistics to develop tests on mutilated texts, i.e., passages from which a number of words and/or phrases have been deleted. Based on test takers' processing of tests developed on the mutilated texts, two major conclusions are drawn: they know how to perform on language tests developed on reduced redundancy, and they have acquired proficiency in language.

The acceptance of these two conclusions has great consequences. If success on reduced redundancy tests is taken as a token of test takers' ability in processing specific methods of testing, then they will be used either on small scales or along with other methods. If conclusions are drawn about their level of language proficiency as Klein-Braley (1997) did, they will affect millions of students (Carver, 1992). This is because the results of proficiency tests are employed for both educational and non-educational purposes, including diagnosis, placement, selection, awards, certification, licensure, and employment (Haladyna, 1994). A comparative study is therefore necessary to determine what method should be used to draw what conclusion.

The concept of reduced redundancy has been employed to develop three major types of testing methods, namely, cloze tests, cloze multiple choice item tests and C-Tests. Hinofotis (1980), Mullen (1980), Oller (1973) and Shohamy (1978) reported high correlations between cloze tests and tests of proficiency measuring language skills (e.g., listening and speaking) and components (e.g., grammar and vocabulary). Klein-Braley (1981), however, found cloze tests to measure language components rather than language skills.

Further studies also showed that not only the knowledge of language skills and components but also the rate of deletion is an important factor in developing cloze tests. Alderson (1983) and Farhady, Jafarpur and Birjandi (1994), for example, found that changing the rate of deletion results in having different tests. These studies indicated that we must exercise caution with respect to the conclusions we draw from the performance of test takers on cloze tests. In other words, we should bear in mind that knowing how to perform on a certain method of testing should not be taken for performing in a certain language.

Since the issue of deletion rates has proved to be central in developing cloze tests, some attempts have been made to remedy this shortcoming. Farhady and Keramati (1996), for example, suggested text characteristics such as cohesive ties, noun phrases, dependent and independent clauses, verb phrases, and co-ordinations to determine the rate of deletion. To avoid the shortcoming of cloze tests and explore their remedied version, the text-driven cloze test developed by Farhady and Keramati was therefore employed to explore schema-based cloze MCITs in the context of reduced redundancy and proficiency. The results obtained by Farhady and Keramati indicated that text-driven cloze tests produce superior psychometrics and have criterion validity.

Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988) conducted a research project to examine the relation of TOEFL performance to a widely used variant of the cloze procedure, i.e., cloze multiple choice item tests (MCITs). For developing the test, they chose six passages from previously used TOEFL forms representing an appropriate range of difficulty and constructed a total of 150 cloze multiple choice items. A rational deletion procedure was used for identifying the words to omit. The test was pretested in three local universities in order to identify, modify or discard the items showing a low correlation with the total test score. On the basis of pretesting, a final set of three passages was selected with 50 cloze multiple choice items. This test, labeled traditional cloze MCIT, was included in the present study as another measure of reduced redundancy.

In contrast to traditional cloze MCITs, which are developed by "measurement specialists, with advisory input from subject specialists" (Bachman, Davidson, Ryan & Choi, 1995: 17), Khodadady (1997: 220) proposed schema theory as a sound rationale to be employed by language teachers to develop their own language tests. Since poorly constructed MCITs do not serve the purposes to which they are put, they always require experts to construct traditional cloze MCITs having attractive and plausible distracters. Khodadady applied schema theory to the

construction of cloze MCITs and developed *competitives* for the alternatives referred to as distracters in traditional cloze MCITs.

According to schema theory, a schema is viewed as an abstract or idealised entity like lexeme (Taylor, Harris & Person, 1988). Each schema or lexis has a semantic network, or a set of semantic interrelationships among different schemata, i.e., the plural of schema. These semantic interrelationships are established through certain common and distinctive semantic attributes. Khodadady (1997, 1999) and Khodadady and Herriman (2000) maintained that reading comprehension ability depends on what readers know about the schemata presented in the texts.

Khodadady (1997, 1999) argued that each lexical item used in an authentic and unmodified text should be viewed as the author's schema. Whatever schema is deleted in a cloze MCIT, it should be given as a keyed response along with the schemata that are semantically interconnected with the deleted schema. These alternative schemata are traditionally referred to as distracters. Khodadady (1997), however, used the term *competitives* to emphasize the semantic relationships between the deleted schema and the alternative schemata chosen by the test writer.

The semantic relationships between the deleted schema and its competitives are determined by contextual schemata, i.e., schemata which precede and follow the deleted schema. For example, Khodadady (1997: 126) developed the schema-based cloze multiple choice item on the deleted schema *prying*, as shown below. He selected the competitives *inquiring*, *interfering* and *probing* from the *Roget's International Thesaurus* (Chapman, 1992). The semantic features of the four schemata are presented in Figure 1. The author's deleted schema *prying* has the semantic feature of *making a search*, which is shared by the schemata *inquiring* and *probing*. However, the schema *prying* also has the semantic feature of *acting uninvitedly*, which is not shared by the schemata *inquiring* and *probing*. Instead, the schema *interfering* shares the semantic feature of *acting uninvitedly* with the schema *prying* but differs from it in terms of having its own semantic feature of *hindering*.

Fears over access to medical records

Privacy campaigners in the US have launched a fierce attack on a bill that they believe will expose medical records to too many ... (1) eyes.

1.	a. inquiring	b. prying*	c. interfering	d. probing

	Semantic features			
Microschemata	**Making a search**	**Acting uninvitedly (searching)**	**Hindering**	**Secretly**
a. inquiring	+	-	-	-
b. prying	+	+	-	+
c. interfering	-	+	+	-
d. probing	+	+	-	-

Figure 1: semantic features of the deleted microschema prying and its competitives

The provision of competitives which share some semantic features with the deleted schema involves the test takers' background knowledge in that they should comprehend the *contextual schemata* in order to determine what distinctive semantic features should be used to select the deleted schema as the keyed response. In contrast to the distracters of traditional cloze MCITs, which depend on testing *specialists*, the competitives of the schema-based cloze MCITs can easily be found in thesauri such as *Roget's Thesaurus of English words and phrases* (Chapman, 1992) and *the New Collins Thesaurus* (McLeod, 1984). Based on these explanations, a schema-based cloze MCIT was developed by the present researcher and included in the study.

In addition to the schema-based, traditional cloze MCITs and text-driven cloze test, a C-Test was also included in the study. C-Tests were invented by Klein-Braley and Raatz (1985, 1990) in order to retain the positive aspects of cloze tests and remedy their technical defects.

They consist of one hundred items developed on short texts chosen carefully from different fields. Each item is formed by mutilating the second half of every second word comprising the text so that a fixed deletion rate can be followed. C-Tests are designed on the assumption that adult educated native speakers should normally make virtually perfect scores on the test.

The present study was undertaken to explore the performance of non-native speakers on four methods of reduced redundancy testing, namely, C-Tests, schema-based cloze MCITs, text-driven cloze tests, and traditional cloze MCITs. The following research questions were raised to guide the study:

1. How reliable is each method?
2. How valid is each method?
3. What is the factor structure for the various tests of reduced redundancy? Which of the four methods best represents the general factor?
4. What is the factor structure if TOEFL is included? Do the reduced redundancy tests load on the same factor as TOEFL?

2. Method

2.1. Participants

34 senior undergraduate Iranian students who were majoring in teaching English as a foreign language at Kurdistan University took part in the study. They spoke Kurdish (70.6%), Turkish (17.6%) and Persian (11.8%) as their first language. Out of 34 students, 21 were male, and the rest were female. The age of participants ranged between 21 and 31 and the highest percentage (23.5%) were 24 years old. The participants took all the tests as part of a course requirement offered by the researcher.

2.2. Instrumentation

The disclosed TOEFL test 1 (Educational Testing Service, 1991, pp. 75-100) was used as the criterion measure to validate the four methods of reduced redundancy testing used in this study, i.e., C-tests, schema-based cloze MCITs, text-driven cloze MCITs and traditional cloze MCITs. The TOEFL test consisted of five subtests: listening comprehension, structure, written expression, vocabulary and reading comprehension, consisting of 50, 15, 25, 30 and 30 multiple choice items, respectively.

In a study similar to the present one, Klein-Braley (1997) developed four C-tests to compare them with cloze tests, four-choice cloze tests (Manning, 1986), cloze-elide tests (Manning, 1986) and dictation from the DELTA. She employed the Duisburg placement test DELTA, a high security test, as the validation criterion and performed a number of statistical operations on her data including factorial analysis. The tests were administered to 81 students. It was found that the C-test showed superior performance over the other test procedures in terms of difficulty level, reliability, validity, and factorial validity.

The four C-Tests developed by Klein-Braley (1997) were used in this study. Although Klein-Braley and Raatz (1985, 1990) declared that the total C-test should have at least 100 items, the one given by Klein-Braley (1997: 79-80) consisted of 99 items. With the exception of C-Test 2, which had 24 items, the other three C-Tests had 25 items each.

Schema-based cloze MCIT was developed by the present researcher (it is reproduced in Appendix 1). The test was constructed on an authentic and unmodified article, *'Miracle' jab makes fat mice thin*, published in *NewScientist* magazine (5 August 1995, No. 1989). Some of the articles in *NewScientist* have been previously used in the construction of the International English Language Testing System and are "thought to be more academic than … articles in quality newspapers" (Clapham, 1996: 145). The articles of *NewScientist* provide standard

scientific texts for public readership. The readability Ease score of Flesch (57.4) indicated that the text was fairly difficult for high school students at grades 10-12 (Flesch, 1948, 1949).

The text-driven cloze test was developed by Farhady and Keramati (1996). They chose the text *telepathy* from university level textbooks on the basis of the Fog index of readability scale. Utilizing the linguistic and discourse structure of the text, Farhady and Keramati developed nine versions of cloze test. They administered the versions to 403 students (10 graduate and 393 undergraduate university students of English). They employed Comprehensive English Language Test (CELT) as a criterion measure. Forms A and B in which the deletion rates were based on the number of cohesive ties and noun phrases, respectively, showed better psychometric qualities than the other forms. Since these forms were not given in Farhady and Keramati's index, form D designed on verb phrases was used in the present study. It correlated with the structure and vocabulary subtests of CELT .66 and .59, respectively.

Hale, Stansfield, Rock, Hicks, Butler, and Oller (1988) designed the traditional cloze multiple choice item test employed in the present study. They administered the test to 11,290 test takers who took the TOEFL at the November 1984 International administration at test centers in the United States and Canada. The test takers were the native speakers of several different language families: Indo-European (Spanish, French and Farsi); Altaic (Japanese and Korean); Sino-Tibetan (Chinese); Austro-Tai (Indonesian and Thai); and Semitic (Arabic).

Since Hale et al. (1988: 31) performed several statistical analyses on different languages separately, the results obtained on the performance of Farsi speakers (476) are given here because the participants of the present study were also the speakers of Farsi. The traditional cloze MCIT developed by Hale et al. (1988) correlated .69, .71, .74, .75, .77, and .83 with the listening comprehension, structure, written expression, vocabulary, reading comprehension and total TOEFL score, respectively.

2.3. Procedure

The participants took all the tests in four running sessions as follows. In the first session the listening comprehension subtest of TOEFL was administered. Structure, written expression and vocabulary subtests of TOEFL were presented in the second session. The reading comprehension subtest of TOEFL and the text-driven cloze test were taken in the third session and finally the schema-based cloze MCIT and C-tests were held in the fourth session. Both the text-driven cloze test and C-tests were scored on the basis of the exact method as prescribed by their developers.

2.4. Data analysis

The internal consistency reliability of the tests was estimated via Cronbach's μ by using SPSS Release 7.5 for Windows, standard version. The responses to individual items were correlated with the total test scores on each test (biserial) and their *p*-values were estimated in order to explore the difficulty level of the tests. The empirical validity of the C-Tests, schema-based cloze MCITs, text-driven cloze test, and traditional cloze MCIT was determined by correlating them with the TOEFL. Principal component analyses were also run to explore the existence of a general factor.

3. Results and discussion

Table 1 shows the mean scores, the standard deviations, the *p*-values, reliability coefficients and the correlations of the reduced redundancy tests with the TOEFL. In terms of difficulty, as judged by mean *p*-value, the text-driven cloze test is the most difficult method (.46). This finding agrees with what Brown (1993) and Klein-Braley (1997) found in their study, indicating that deleting words on the basis of text characteristics does not change the difficulty level of cloze tests. The difficulty level of C-test (.60), schema-based cloze MCIT (.59) and traditional cloze MCIT (.61) is almost the same.

Tests	Mean	SD	*P*	rtt_{μ}	rtt_{TOEFL}
C-TEST	53.38	10.40	.54	.86	.45**
C-Test 1	14.38	2.72	.58	.43	.49**
C-Test 2	12.00	3.23	.50	.61	.32
C-Test 3	14.09	3.78	.56	.74	.36*
C-Test 4	12.91	3.50	.52	.72	.26
Schema-based cloze MCIT	36.94	6.13	.59	.71	.74**
Text-driven cloze	15.76	3.33	.45	.55	.43*
Traditional cloze MCIT	30.71	6.45	.61	.78	.75**
TOEFL	89.76	14.81	.60	.88	1.00

Table 1: basic statistics for the four methods of reduced redundancy testing and TOEFL

** Correlation is significant at the 0.01 level (2-tailed)
* Correlation is significant at the 0.05 level (2-tailed)

3.1. Reliability

The second highest reliability coefficient after the TOEFL (.88) belongs to the total C-test (.86). The third and fourth reliable tests are traditional cloze MCIT (.78) and schema-based cloze MCIT (.71), respectively. The text-driven cloze test is the least reliable test (.55) among the four methods of reduced redundancy testing. The superiority of C-Test over the schema-based cloze MCIT and traditional cloze MCIT in reliability stems from its length. It is a well-established principle in psychometrics that the longer a test is, the more reliable it becomes (e.g., Gronlund & Linn, 1990; Nunnally, 1978). As shown in Table 1, the reliability coefficient of C-test 1 (. 43) is much lower than the text-driven cloze test (.53).

3.2. Validity

As Davies (1990:23) emphasized, “reliability is necessary for a test (…) but it is not sufficient. What matters in a test is that it should be *valid*”. If we accept validity as “the agreement between two attempts to

measure the same trait through maximally different methods" (Campbell & Fiske, 1959: 83), then we can say that the TOEFL, schema-based cloze MCIT and traditional cloze MCIT on the one hand and C-test and text-driven cloze test on the other measure two different traits. As shown in Table 2, however, the C-test is the least valid test (rtt_{TOEFL} = .45), followed by the text-driven cloze test (rtt_{TOEFL} = .43). The validity coefficients for the schema-based and traditional cloze MCITs are .74 and .75, respectively, indicating that these two tests are more valid measures of English language proficiency than the C-test and text-driven cloze test.

3.3. Factor analysis

Following the study of Klein-Braley (1997), two principle component analyses were performed. The first analysis was done to answer the research question: which of the four methods best represents the general factor? Table 2 presents the results. The highest loadings on the general factor belong to the C-Test (.96) and text-driven cloze test (.74), respectively. Since these two tests are both cloze-based, I call this general factor *cloze-based reduced redundancy testing*. It is best represented by the C-Test.

Test	**Factor 1**	**Factor 2**
C-Test 1	.83	*
C-Test 2	.88	*
C-Test 3	.73	*
C-Test 4	.70	-.33
C-Test (Total)	.96	*
Schema-based cloze MCIT	.55	.70
Text-driven cloze test	.74	*
Traditional cloze MCIT	.69	.47
	Eigenvalue: 4.74	Eigenvalue: 1.07
	Variance: 59.22%	Variance: 13.40%

Table 2: Unrotated Factor Matrix using principle factor with iteration for the redundancy tests

* Loadings less than .30

As shown in Table 2, C-test 4, the schema-based cloze MCIT, and traditional cloze MCIT also load on the second factor upon which the schema-based cloze MCIT and C-Test 4 have the highest and lowest loadings, respectively, i.e., 0.70 and -0.33. Since the format of the schema-based and traditional cloze MCITs differ from C-Test 4, the second factor cannot be attributed to the characteristics of multiple choice items. Furthermore, although the schema-based and traditional cloze MCITs both share multiple choice format and must therefore have a higher loading on the second factor if it represents their nature, the traditional cloze MCIT has a much higher loading on the first rather than the second factor, i.e., 0.69 and 0.47, respectively. This calls for employing rotation to achieve "a simpler factor structure, preferably with each variable loading primarily on only one factor" (Farhady, 1983: 19).

Table 3 presents the varimax rotated factor matrix using principle component analysis of the reduced redundancy tests. As can be seen, only the C-Test (.91) and text-driven cloze test (0.37) load on the general factor, and thus provide further support for my earlier conclusion that it represents *cloze-based reduced redundancy testing*. The schema-based and traditional cloze MCITS, as well as the C-Test itself, however, load on the second factor, i.e., 0.89, 0.74 and 0.41, respectively. Since all the three measures load on this factor, I call it *reduced redundancy testing*.

Tests	**Factor 1**	**Factor 2**
C-Test 1	.71	.42
C-Test 2	.69	.44
C-Test 3	.60	.41
C-Test 4	.88	*
C-Test	.91	.41
Schema-based cloze MCIT	*	.89
Text-driven cloze test	.37	.72
Traditional cloze MCIT	*	.74
	Eigenvalue: 4.56	Eigenvalue: 1.05
	Variance: 57.05	Variance: 13.12

Table 3: Varimax with Kaiser rotated factor matrix using principal component analysis for the reduced redundancy tests

* Loadings less than .30

Table 4 presents the results of the third factor analysis performed after the TOEFL test is added to the factor analysis. As shown in Table 4, even after the addition of the TOEFL, the C-Test and text-based cloze test have the highest loading on the general factor, i.e., 0.93 and 0.74, respectively. This provides further support for the conclusion that the general factor measures *cloze-based reduced redundancy*. Since all the tests, including the TOEFL, load on the general factor, one might be forced to conclude that it represents language proficiency rather than cloze-based reduced redundancy testing. Two reasons, however, defy this conclusion.

The first reason is that the structure section of the TOEFL is cloze-based. Example 1 taken from the ETS (1991, p. 82), for instance, is based on the stem *Vegetables are an excellent source … vitamins*. The item requires test takers to choose *of* from among alternatives *has*, *where* and *that* to restore the deleted word. The TOEFL loads, therefore, on the general factor because it measures cloze-based reduced redundancy.

The loading of the TOEFL (0.69), which is much lower than that of the C-Test (0.93), provides the second reason. It is further highlighted when we remember the fact that the TOEFL is the first measure analyzed in the analysis. Normally the first variable must receive the highest loading if it stands for whatever the general factor represents. This calls for rotation solution to simplify the interpretation.

Test	**Factor 1**	**Factor 2**
TOEFL	.69	.60
C-Test 1	.81	*
C-Test 2	.78	-.31
C-Test 3	.70	*
C-Test 4	.65	-.46
C-Test (Total)	.93	-.36
Schema-based cloze MCIT	.73	.50
Text-driven cloze test	.74	*
Traditional cloze MCIT	.73	.44
	Eigenvalue: 5.13	Eigenvalue: 1.30
	Variance: 56.99	Variance: 14.43

Table 4: Unrotated Factor Matrix using principle factor with iteration for the redundancy tests and TOEFL

* Loadings less than .30

Table 5 presents varimax rotation of factors for the reduced redundancy tests and the TOEFL. As can be seen, the C-Test still claims for the highest loading (.95) on the first factor. The TOEFL does not, however, load on the first factor any more. Neither do the schema-based cloze MCIT and traditional cloze MCIT. Since only the C-Test and text-driven cloze test load on the first factor, they provide further evidence to the conclusion that the general factor represents cloze-based reduced redundancy testing. Since the TOEFL has the highest loading on the second factor, I call it *the English language proficiency*.

Tests	**Factor 1**	**Factor 2**
TOEFL	*	.90
C-Test 1	.72	.39
C-Test 2	.80	*
C-Test 3	.67	*
C-Test 4	.80	*
C-Test	.95	.30
Schema-based cloze MCIT	*	.85
Text-driven cloze test	.56	.49
Traditional cloze MCIT	*	.81
	Eigenvalue: 3.62	Eigenvalue: 2.81
	Variance: 40.27	Variance: 31.21

Table 5: Varimax with Kaiser rotated factor matrix using principal component analysis for the reduced redundancy tests with the TOEFL

* Loadings less than .30

As can be seen in Table 5, the four measures of reduced redundancy testing, i.e., the C-Test, schema based and traditional cloze MCITs and text-driven cloze test load on the second factor. This is in sharp contrast to the general factor on which only two measures load. Since the four measures load on the second factor along with the TOELF, I conclude that it represents both *reduced redundancy testing* and *English language proficiency*. Considering the fact that among the four measures of reduced redundancy only schema-based cloze MCIT has the highest loading on the second factor after the TOELF, I also conclude that schema-based cloze MCITs are the best representatives of reduced redundancy testing.

4. Conclusion

Schema-based cloze multiple choice item tests (MCITs) are the only methods of reduced redundancy that have a sound theoretical rationale to guide item writers in developing attractive and plausible competitives. The application of schema theory to the construction of multiple choice items results in developing language proficiency tests which are as functional and as objective as their traditional MCITs. In contrast to traditional cloze MCITs, whose functioning depends on the expertise or intuition of item writers, however, language teachers can employ objective sources such as thesauri to write their own schema-based cloze multiple choice items through the application of semantic traits analysis.

In addition to having a theoretical rationale, schema-based cloze MCITs have superiority over other measures of reduced redundancy testing as regards empirical and factorial validity. Though they correlate significantly highly with English language proficiency tests along with traditional cloze MCITs, they have the second highest loading on the factor representing English language proficiency and reduced redundancy testing.

Although schema-based cloze MCITs measure reduced redundancy, they are not as reliable as C-Tests. In contrast to schema-based cloze MCITs, nonetheless, C-Tests share the major shortcoming of traditional cloze MCITs in that "they should be used only under the supervision of the test expert who should evaluate the suitability of the C-Test in question for the specific target group in question" (Klein-Braley, 1997: 72).

Finally, schema-based cloze MCITs function better than the text-driven cloze tests. Although developing cloze tests on the basis of text characteristics reduces the element of subjectivity in their construction, it does not bring about any changes in the low reliability and validity of cloze tests. Among the measures of reduced redundancy, text-driven cloze tests suffer from the least reliable and the most difficult items. Although their empirical validity is slightly superior to C-Tests, they suffer from low factorial validity in contrast to schema-based cloze MCITs and traditional cloze MCITs.

Acknowledgements:

I would like to thank Professor Jafarpur from Shiraz University and Dr. Nassaji from Centennial College for their fruitful comments on the procedure and factorial analysis of the study.

Recebido em: 01/2004; Aceito em: 04/2004.

References:

ALDERSON, J.C. 1983 The cloze procedure and proficiency in English as a foreign language. IN J.W. OLLER, Jr. (ed.) *Issues in language testing research*. Newbury House.

BACHMAN, L.F., DAVIDSON, F., RYAN, K. & CHOI, I. 1995 *An investigation into the comparability of two tests of English as a foreign language: the Cambridge-TOEFL comparability study*. CUP.

BROWN, J.D. 1993 What are the characteristics of natural cloze test? *Language Testing,* **10:** 93-116.

CAMPBELL, D.T. & FISKE, D.W. 1959 Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* **56:** 81-105.

CARVER, R.P. 1992 What do standardized tests of reading comprehension measure in terms of efficiency, accuracy and rate? *Reading Research Quarterly*, **27.4:** 346-359.

CHAPMAN, R.L. (ed.) 1992 *Roget's international thesaurus* (5th ed.) Harper Perennial.

CLAPHAM, C. 1996 *The development of IELTS: a study of the effect of background knowledge on reading comprehension*. CUP.

DAVIES, A. 1990 *Principles of language testing*. Basil Blackwell.

EDUCATIONAL TESTING SERVICE 1991 *Reading for TOEFL*. ETS.

FARHADY, H. 1983 On the plausibility of the unitary language proficiency factor. IN J.W. OLLER, Jr. (ed.) *Issues in language testing research*. Newbury House.

FARHADY, H.; JAFARPUR, A. & BIRJANDI, P. 1994 *Testing language skills: from theory to practice*. SAMT Publication.

FARHADY, H. & KERAMATI, M.N. 1996 A text-driven method for the deletion procedure in cloze passages. *Language testing,* **13.2:** 191-207.

FLESCH, R.F. 1948 A new readability yardstick. *Journal of Applied Psychology*, **32:** 221-33.

FLESCH, R.F. 1949 *The art of readable writing*. Harper & Row.

GRONLUND, N.E. & LINN, R.L. 1990 *Measurement and evaluation in teaching* (6th ed.) Macmillan.

HALADYNA, T. M. 1994 *Developing and validating multiple-choice test items*. Lawrence Erlbaum Associates.

HALE, G.A.; STANSFIELD, C.W.; ROCK, D.A.; HICKS, M.M.; BUTLER, F.A. & OLLER, J.W. Jr. 1988 *Multiple-choice items and the Test of English as a Foreign Language* (TOEFL Research Report No. 26). Educational Testing Service.

HINOFOTIS, F. 1980 Cloze as an alternative method of ESL placement and proficiency testing. IN J.W. OLLER, Jr. (ed.) *Issues in language testing research*. Newbury House.

KHODADADY, E. 1997 *Schemata theory and multiple choice item tests measuring reading comprehension*. Unpublished PhD dissertation, the University of Western Australia.

_____ & HERRIMAN, M. 2000 Schema theory and selected response item tests: from theory to practice. IN: A.J. KUNNAN (ed.) *Fairness and validation on language assessment.* CUP.

_____ 1999 *Multiple choice items in testing: practice and theory*. Rahnama.

KLEIN-BRALEY, C. 1981 *Empirical investigations of cloze tests*. Unpublished PhD dissertation, University of Duisburg.

_____ (1997). C-Tests in the context of reduced redundancy testing: an appraisal. *Language Testing*, **14/1**, 47-84.

_____ & RAATZ, U. (eds.) 1985 *Fremdsprachen und Hochschule 13/ 14: Thematischer Teil: C-Tests in der Praxis*. AKS.

_____ & RAATZ, U. 1990 Die objektive Erfassung des Sprachstands im mutter- und fremdsprachlichen Unterricht durch C-Tests. IN A. WOLFF & H. RÖSSLER (eds.) *Deutsch als Fremdsprache in Europa*. Arbeitskreis Deutsch als Fremdsprache.

MANNING, W.H. 1986 *Development of cloze-elide tests of English as a second language*. Educational Testing Service.

MCLEOD, W.T. 1984 *The new Collins thesaurus.* Collins.
MULLEN, K. 1980 Rater reliability and oral proficiency evaluation. IN: J.W. OLLER, Jr. & K. PERKINS (eds.) *Research in language testing*. Newbury House.
NUNNALLY, J.C. 1978 *Psychometric theory* (2nd ed.) McGraw-Hill.
OLLER, J.W. Jr. 1973 Cloze tests of second language proficiency and what they measure. *Language learning,* **23.1:** 105-18.
SHOHAMY, E. 1978 *Investigation of concurrent validity of oral interview with cloze procedure for measuring proficiency in Hebrew as a second language*. Unpublished PhD dissertation, University of Minnesota.
SPOLSKY, B. 1973 What does it mean to know a language; or how do you get somebody to perform his competence? IN J. W. OLLER Jr. & J.R. RICHARDS (eds.) *Focus on the learner.* Newbury House.
TAYLOR, B.; HARRIS, L.A. & PEARSON, P.D. 1988 *Reading difficulties: instruction and assessment*. McGraw-Hill.

APPENDIX 1

SCHEMA-BASED CLOZE MULTIPLE CHOICE ITEM TEST

Directions: 63 words from the following passage have been deleted and replaced with a numbered blank space. For each deleted word four choices marked **a**, **b**, **c**, and **d** have been offered. Choose the **word** which you think is the most appropriate to fill the blank. Your choice should be based on what comes before and after the blank and the text as a whole. Indicate your choice by **circling** one of the four letters. Time allotted: **30** minutes

'Miracle' jab makes fat mice thin

After a four-week course of treatment with a protein ... (1)ob, the fat simply falls off, leaving vastly overweight mice ... (2), active and sensible eaters. If the protein has the ... (3) effect on people, it could be the miracle cure millions have been ... (4) for. That, at least, is the theory. But sceptics ... (5) that too little is known about the way the human ... (6) of the ob protein works to be sure that extra doses would ... (7) people to lose weight.

1.	a.	labelled	b.	known	c.	called	d. entitled
2.	a.	Slim	b.	lean	c.	lanky	d spare
3.	a.	Alike	b.	very	c.	identical	d. same
4.	a.	lingering	b.	waiting	c.	delaying	d. remaining
5.	a.	Alert	b.	inform	c.	warn	d. frighten
6.	a.	model	b.	version	c.	replica	d. duplication
7.	a.	relieve	b.	advise	c.	heal	d. help

But when the results of the tests were ... (8) last week, Amgen, the Californian biotechnology company which ... (9) the exclusive rights to develop products based on the protein, saw an overnight ... (10) in its share prices.

8.	a.	leaked	b.	displayed	c.	advertised	d stated
9.	a.	occupies	b.	seizes	c.	retains	d owns
10.	a.	surge	b.	jump	c.	advance	d. boom

Last December, a team led by Jefferey Friedman and ... (11) colleagues at the Howard Hughes Medical Institute at the Rockefeller University, New York ... (12) a gene which they called *ob*. In mice, a ... (13) in this gene makes them grow hugely obese. Humans have an almost .... (14) gene, suggesting that the product of the gene - the ob protein - ... (15) a part in appetite control. The ob protein is a hormone, which Friedman had ... (16) Leptin.

11.	a.	her	b.	his	c.	their	d. its
12.	a.	invented	b.	innovated	c.	pioneered	d discovered
13.	a.	deficit	b.	shortage	c.	defect	d. gap
14.	a.	alike	b.	identical	c.	similar	d. identifiable
15.	a.	plays	b.	does	c.	executes	d. conducts
16.	a.	nominated	b.	specified	c.	dubbed	d. known

In April, Amgen, which is based in Thousand Oaks, California, ... (17) the institute $20 million for exclusive rights to ... (18) products based on the discovery. Amgen will carry out ... (19) tests on the protein in animals next year, and hopes to begin ... (20) trials on people within a year.

17.	a.	spent	b.	invested	c.	squandered	d. paid
18.	a.	improve	b.	cultivate	c.	develop	d evolve
19.	a.	security	b.	safety	c.	protection	d. defense
20.	a.	surgical	b.	hygienic	c.	psychic	d. clinical

The excitement began last week when the journal *Science* ... (21) the findings of three groups which have been ... (22) on the protein. The results in obese mice with a(n) ... (23) gene that prevents them making the protein were ... (24). Mary Ann Pelleymounter and her colleagues at Amgen gave obese mice ... (25) of the protein every day for a month. Those on the highest dose ... (26)an average of 22 per cent of their weight.

21.	a.	published	b.	distributed	c.	broadcasted	d. announced
22.	a.	acting	b.	working	c.	performing	d operating
23.	a.	insufficient	b.	unsound	c.	partial	d. defective
24.	a.	tense	b.	romantic	c.	dramatic	d. moving
25.	a.	ejections	b.	bullets	c.	tosses	d. shots
26.	a.	lost	b.	wasted	c.	spent	d. squandered

"Before treatment, these mice ... (27), had lower metabolic rates than normal, lower temperatures, and ... (28) levels of insulin and glucose in their blood," says Pelleymounter. "The protein brought all ... (29) back to normal levels," she says.

27.	a.	overfed	b.	overdined	c.	overate	d. overtook
28.	a.	lifted	b.	raised	c.	risen	d erected
29.	a.	them	b.	they	c.	it	d. these

More significantly, in terms of the ... (30) for a human slimming drug, the treatment also worked on normal mice, ... (31) lost what little spare fat they had. They lost ... (32) 3 and 5 per cent of their body weight, almost all of ... (33) in the form of fat, according to Pelleymounter. This is important because no one has ... (34) a mutation in the human *ob* gene that might ... (35) to obesity, suggesting that whatever the ... (36) of obesity, the ob protein might still help people lose weight.

30.	a.	talent	b.	faculty	c.	potential	d. gift
31.	a.	that	b.	who	c.	what	d which
32.	a.	among	b.	from	c.	between	d. until
33.	a.	them	b.	it	c.	its	d. their
34.	a.	identified	b.	distinguished	c.	realized	d. recognized
35.	a.	cause	b.	motivate	c.	induce	d. lead
36.	a.	reason	b.	cause	c.	principle	d. factor

Friedman and his team ... (37) similar experiments. In just one month, their obese mice ... (38) around half their body fat. In the average obese mouse, ... (39) makes up about 60 per cent of body weight. Treated mice lost their ... (40). Within a few days they were eating about 40 per cent as much as untreated animals. Their fat practically ... (41) away, falling to 28 per cent of their body weight after a month. In normal mice, treatment ... (42) the amount of fat from an average of 12.22 per cent of body weight to a spare 0.67 per cent.

37.	a.	operated	b.	dealt	c.	conducted	d. carried
38.	a.	stripped	b.	shed	c.	removed	d plucked
39.	a.	fat	b.	meat	c.	bones	d. water
40.	a.	will	b.	appetite	c.	lust	d. taste
41.	a.	vanished	b.	perished	c.	diminished	d. melted
42.	a.	subtracted	b.	concentrated	c.	reduced	d. withered

Friedman and Pelleymounter believe that the protein, which is produced by fat cells, ... (43) appetite. "We think it's something like a ... (44) hormone to tell the brain there are normal amounts of fat, or too much, in which case the brain ... (45) down your appetite," says Pelleymounter.

43.	a.	harmonizes	b.	regulates	c.	coordinates	d. adapts
44.	a.	rotating	b.	spinning	c.	circulating	d gyrating
45.	a.	declines	b.	decreases	c.	abates	d. turns

The experiments also show that ... (46) mice have an increased metabolic rate, suggesting that they burn fat more ... (47). Their appetites decrease and they are less ... (48), becoming as active as normal mice.

46.	a.	cured	b.	prescribed	c.	treated	d. remedied
47.	a.	competently	b.	efficiently	c.	adroitly	d aptly
48.	a.	sluggish	b.	apathetic	c.	dormant	d. slack

The third group of researchers from the Swiss Pharmaceutics Company Hoffmann-La Roche, are more ... (49) about how significant the ob protein might be in treating obesity. From their studies, they ... (50) that the protein is just one of many ... (51) that control appetite and weight. "This is a very important ... (52), but it's one of several," says Arthur Campfield, who led the team.

49.	a.	agnostic	b.	faithless	c.	incredulous	d. sceptical
50.	a.	gather	b.	conclude	c.	infer	d determine
51.	a.	ingredients	b.	features	c.	factors	d. items
52.	a.	token	b.	index	c.	measure	d. signal

Campfield ... (53) whether the ob protein alone will have much effect in overweight humans. His team hopes to ... (54) the signalling system that regulates weight, and is particularly ... (55) to find the receptor in the brain that responds to the ob hormone. Hoffmann-La Roche, ... (56) by the Amgen license deal from developing products based on the protein itself hopes to develop pills that ... (57) with the message pathways in appetite control. Stephen Bloom, professor of endocrinology at London's Hammersmith Hospital, ... (58). "I think the work with ob is a major advance, but

we've not got the tablet yet. That will come when people have made a pill that ... (59) the ob receptors in the brain so it switches off appetite."

53.	a.	disbelieves	b.	doubts	c.	denies	d. hesitates
54.	a.	develop	b.	extract	c.	unravel	d demonstrate
55.	a.	fervent	b.	keen	c.	intense	d. delirious
56.	a.	banished	b.	prohibited	c.	fired	d. excluded
57.	a.	interrupt	b.	insert	c.	interfere	d. infiltrate
58.	a.	complies	b.	agrees	c.	coincides	d. identifies
59.	a.	stimulates	b.	motivates	c.	animates	d. impels

Even Pelleymounter at Amgen cautions against overoptimism at this stage. "We don't know whether it would be ... (60) that people would lose weight, but you can ... (61) from mice that it would have some ... (62) effect," she says. "However, I don't think obese people should hold out for this. They should carry on with their ... (63) and dieting" (NewScientist, 5 August 1995, No. 1989).

60	a.	constant	b.	sure	c.	secure	d true
61.	a.	forecast	b.	predict	c.	forebode	d. speculate
62.	a.	absolute	b.	certified	c.	positive	d. settled
63.	a.	exercises	b.	exertion	c.	practice	d. drills

Ebrahim Khodadady is an assistant professor of Applied Linguistics at Urmia University, Urmia, Iran. He is the director of Foreign Language Services at Part Time Education Center. The center offers a joint English language teaching program with Brock University, St. Catherines, Ontario, Canada. Khodadady's main interests are language testing, research, teaching methodology and translation. ekhodadady@yahoo.com