International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:4, No:7, 2010

How Valid Are Our Language Test Interpretations? A Demonstrative Example

Masoud Saeedi, Shirin Rahimi Kazerooni, and Vahid Parvaresh

***Abstract*—**Validity is an overriding consideration in language testing. If a test score is intended for a particular purpose, this must be supported through empirical evidence. This article addresses the validity of a multiple-choice achievement test (MCT). The test is administered at the end of each semester to decide about students' mastery of a course in general English. To provide empirical evidence pertaining to the validity of this test, two criterion measures were used. In so doing, a Cloze test and a C-test which are reported to gauge general English proficiency were utilized. The results of analyses show that there is a statistically significant correlation among participants' scores on the MCT, Cloze, and C-test. Drawing on the findings of the study, it can be cautiously deduced that these tests measure the same underlying trait. However, allowing for the limitations of using criterion measures to validate tests, we cannot make any absolute claim as to the validity of this MCT test.

***Keywords*—**C-test, Cloze test, Multiple-choice test, Validity argument.

I. INTRODUCTION[1]

THE reason behind giving a language test and obtaining a test score is interpreting that score as an indicator of what a test taker knows or what he/she can do with that knowledge. Furthermore, our interpretation of that test score forms the basis for decision making. As such, when using a test score, we make an implicit link between test performance and a domain of language knowledge the test taker has or something the test taker can do with language in some language use domain beyond the test itself.

In other words, when we use test scores, we are essentially reasoning from evidence, using the test score as the evidence for inferences or interpretations and decisions we want to make [1]. Yet, we cannot simply draw on test score to make inferences and decisions without efficient justification. If we want to use a test score for a particular purpose, we must justify it through a rationale and supporting evidence. As Bachman [2] puts it, "We need to demonstrate, with logical argumentation and empirical evidence, that the intended interpretations and uses are valid." Validity in testing and assessment has traditionally been understood to mean "discovering whether a test measures accurately what it is intended to measure" [3], or uncovering the "appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure" [4].

Masoud Saeedi is with Payam-e-Noor University, Isfahan, Iran.
Shirin Rahimi Kazerooni is with Khorasgan Azad University, Iran.
Vahid Parvaresh is with the University of Isfahan, Isfahan, Iran. (corresponding author email: vparvaresh@fgn.ui.ac.ir Vahid Parvaresh)

Validation in language assessment is ominously important, judging educational and linguistic policies, institutional decisions, pedagogical practices, as well as underpinnings of language theory and research. However, establishing validity in language assessment is by all accounts problematic, conceptually challenging, and difficult to achieve [5], [6]. Test validation is the process of generating evidence to support the well-foundedness of inferences concerning trait from test scores, i.e., essentially, testing should be concerned with evidence-based validity.

Test developers need to provide a clear argument for a test's validity in measuring a particular trait with credible evidence to support the plausibility of this interpretative argument [7]. This process entails providing data pertaining to *context-based*, *theory-based* and *criterion-related* validities, together with *reliability*, or *scoring validity*. Educational measurement and language testing offer an elaborate set of procedures for conducting validation research, but rather than a prescribed invariant path – or a menu of equally-appropriate choices – the tools of validation require context-specific decisions about what and how to validate.

II. VALIDATION AS A COMPLEX PROCESS

In the early days of validity investigation, the term was conceptualized as comprising three distinct types. In this view, each type of validity (criterion-oriented validity, content validity, and construct validity) pertained to the kinds of evidence that aimed at demonstrating test validity [8]. Since that time, however, the investigation of validity has become a fundamental consideration in language testing. As a corollary, the cline of validity theory has witnessed significant changes since Lado's conceptualization of validity as a test property. These changes have resulted in different approaches to validity [8]-[10].

Among the many validity conceptualizations, perhaps Messick's framework has been the most significant. In his seminal article, Messick [11] set out to produce a unified validity framework in which different types of evidence (content-related, criterion-related, and construct-related) contribute to our understanding of construct validity as the umbrella term. His conceptualization of validity as a unitary concept has dramatically changed our understanding of validity. Messick's way of considering validity has become the accepted paradigm in psychological, educational and language testing [12]. Messick [11] described validity as "the degree to which empirical evidence and theoretical rationales

support the adequacy and appropriateness of interpretations and actions based on test scores." As such, from this perspective engaging in the process of validation necessitates identifying the relevant theory and empirical evidence and combining these in a way that builds a validation argument that pertains to the purpose of the test. Moreover, the mention of actions in this definition is intended to include consideration of the consequences of testing as one part of validation.

The validation process is portrayed as an ongoing activity with no clear end point. The research literature of language assessment abounds in examples of empirical evidence generated through research. In contrast, the other aspects of validation – relevant theory, the validation argument, and consequences of testing – have been explored less through research and practice.

The process of validation calls for the researchers to draw on theoretical rationales, but what kind of theoretical rationales? Messick [11] invokes 'theory' to refer to the 'interpretive theories' that give substance to score meaning, the 'construct theory' that forms the basis for test specifications, and the 'theory of the construct' that provides a basis for hypothesizing relations between the test construct and other constructs. For language testing, such theories might conceivably be drawn from theories of language, language acquisition, and language performance. Discussion of such theories as they pertain to language assessment reveal the complexity of validation process [13] to the point that McNamara's review of relevant language theory refers to it as 'opening Pandora's box' [14]. If construct theories of language are so complicated, what kind of theoretical rationales can really be brought to bear on a validity argument?

Whatever the theory, the validation process is intended to combine multiple sources of information into an integrated, multifaceted evaluation concerning a language test rather than allowing a test to be considered on the basis of a single research result or a set of results. In other words, validity is conceived as 'an argument' rather than a 'thumbs up/thumbs down verdict', as Cronbach [15] puts it. Like any other argument, a validity argument may consist of many turns; consequently, test-developers, it seems, have a responsibility not to display tests as 'validated' but rather to present validity evidence as a part of a process of posing the next round of validation evidence.

This view is complex relative to the categorical hierarchies that some researchers have attempted to construct by arguing, for example, that reliability is a necessary condition for validity, or that the validity of a test is expressed through its correlation with another test of the same construct. If a variety of evidence is to be included in a validity argument, what form should such an argument take? Approaches for systematizing the process of developing a validity argument have shown the value of beginning with criteria developed on the basis of values in applied linguistics and language testing [13], and conducting an analysis that develops both the positive and negative sides of the argument [9].

III. CURRENT STUDY

The purpose of this study is to validate a Multiple-Choice achievement test of English proficiency (MCT). The researchers have, in fact, been teaching in a well-known language teaching institute in Iran for several years. In this institute assessment is done through administering a test in the multiple-choice question format along with the grades teachers give the students during a three-month general English course. Drawing on the principles of the structuralist approach to language testing, this institute has long adopted and administered this objectively scored, discrete-point, end-of-the-term test of general English proficiency. Despite being practical and objectively scored, however, there have been strong criticisms against theoretical underpinnings of discrete-point tests based on this approach. Although they are practical, opponents of discrete-point tests argue that answering individual items, regardless of their actual function in communication, may not be of much value.

Furthermore, the contribution of any item to assessing total language communication may be neither significant nor identifiable [16], [17]. The structuralist approach usually fails to allow for the interplay between language and context in which it is used. Moreover, students' feedback on the test as well as the researchers' own experience have shown that this test is not a real indication of what students know about language and what they can do with it. There have been many students who get a good mark for their class activity, yet fail to realize their full potential on the test. Therefore, there must be a problem with the test. A question arises here: does this test measure general English proficiency? Put in other words, can we interpret students' score on this test as an indication of their general English proficiency?

To address this question the researchers validated this MCT against two other tests which purport to measure the same construct: a Cloze test and a C-test. Although Cloze testing procedure was originally developed to assess readability of text, it has since been considered an effective means of assessing second language proficiency [18], [19]. This test is now one of the most popular testing techniques, especially for assessing general language proficiency of English-as-a-second-language (ESL) learners [20].

The cloze procedure is considered by many as an 'integrative' method of assessment, in contrast with 'discrete-point' methods, The Cloze procedure deals with several linguistic components at once, focusing more on language use, and typically requiring the examinees to read and comprehend a substantial amount of discourse [16]-[19]. Much research has been devoted to the validity and the reliability of Cloze tests. Cloze tests are reported to have moderate to high correlations with standardized tests and their subtests such as listening comprehension and reading comprehension. Oller [19], Shohamy [21], Hinofotis [22] and Mullen [23] report high correlations between Cloze tests and

tests of listening comprehension, writing, reading comprehension, and oral interview. The findings of these researchers reveal a relation between the scores on Cloze and global language ability tests.

As for C-test, since its inception in 1981, its principles have been applied to more than 20 languages. C-tests have been used in numerous contexts and for various purposes [24]. C-tests, then, figure prominently in the fields of language testing and assessment. C-tests, like the classic Cloze test, are an operationalization of the principle of reduced redundancy testing [25], [26]. Technically speaking, this means that noise is introduced into the channel, for example by deleting parts of the words in a written text. The way in which examinees perform under these conditions is then used as an indicator of their overall language proficiency [26]. In accordance with this view, many researchers claim that C-tests provide an integrative assessment of a construct often referred to as general language proficiency [27]. As mentioned above, C-test and Cloze test can be used as integrative tests of general English proficiency. Using these tests as criteria for validating the MCT, the study is aimed at answering the following research questions:

Q_1: Is there any statistically significant correlation between students' scores on the MCT test and the Cloze test?

Q_2: Is there any statistically significant correlation between students' scores on the MCT test and the C-test?

IV. METHODOLOGY

A. Participants

Some 52 language learners studying English as a second language at a language institute in Iran participated in this study. They were male adults at different ages. All of them were native speakers of Persian. On the basis of their performance on the placement test and the interview given at the institute they were grouped as intermediate. The placement test consisted of a multiple-choice test of listening comprehension, grammar, vocabulary and reading comprehension.

B. Instruments

In this study, a multiple-choice achievement test (MCT) comprising listening comprehension, vocabulary, grammar, and reading comprehension sub-sections was used. The test included 60 questions: 10 listening comprehension questions, 20 vocabulary questions, 20 grammar questions, and 10 questions assessing test takers' reading comprehension. It should be noted that in the listening comprehension of the MCT test no tape or CD is played. Rather, the teacher reads two texts twice giving students 30 seconds in between to look at the questions. Having listened to each text for the second time, they are given 2.5 minutes to check the correct answer in their answer sheets (see Appendix A). The test is administered at the end of each three-month general English course to assess EFL learners' mastery of the course content. In addition to the MCT, a Cloze test, comprising 25 items, was used as the second data collection instrument (See Appendix B). The third means of data elicitation was a C-test, including four short thematically distinct segments of connected discourse (See Appendix C).

C. Procedure

The end-of-the-term MCT is intended to assess test takers' mastery of general English course. As such, it must be fairly capable of providing testers with a general picture of test takers' language proficiency. If this test measures what it purports to measure, it should give testers values which are representative of the ESL learners' mastery of the general English course content. In order to investigate the validity of this MCT, the participants' scores on it must be compared with their performance on the "criterion" tests which are supposed to measure global language proficiency. Since subjects were no longer available after the end of the term, the Cloze test and C-test were administered a couple of weeks before the end of the semester when they would be given the MCT. 52 students were given the C-test, the Cloze test and the end-of-the-term MCT.

After collecting the data, the calculated descriptive statistics pertaining to subjects' performance on the administered tests, i.e., Cloze test, C-test and the MCT, were tabulated. As the next step, running the SPSS, the Pearson Product-moment correlation formula was used to estimate the correlation coefficients among the tests. Then, drawing on the obtained matrix of correlation, factor analysis was run to identify the latent structures that would underlie the database. In this way, the researchers identified the common factors underlying the measures. In doing so, the correlation coefficient between the C-test, Cloze test, and individual sub-sections of the MCT were calculated, and after that, the global correlation coefficient, that is, the correlation coefficients between the C-test, Cloze test, and the total MCT were estimated.

As for the scoring procedure, although C-test cannot adopt the multiple-choice format, it is fairly objective in terms of scoring procedure. Test takers were required to supply the missing letters of the words and each correct restoration of missing letters was given one score. Furthermore, spelling mistakes were penalized. The C-test comprised four short thematically distinct texts. Around five minutes were allowed for each text so that the whole C-test took 20 minutes to complete.

Since test performance is affected by the characteristics of the methods used to elicit test performance and the characteristic of the expected response is one of the many test facets that affect performance on language tests [5], the researchers decided to adopt the constructed- as opposed to selected- type of expected response for the Cloze test, too. Therefore, both Cloze test and C-test adopted the constructed response type. Therefore, test takers were required to supply the missing items of the Cloze test. Likewise, since the Exact Word scoring procedure is fairly objective, it was used for

scoring the Cloze. Put in a nutshell, both the Cloze test and the C-test adopted the constructed response type and were objectively scored. The Cloze test included 25 missing items and each correct restoration was given one point. Twenty minutes were allocated for its completion. The MCT contained 60 multiple-choice questions. Correct answer to each question was given one point.

V. RESULTS AND DISCUSSION

The data collection tools comprising the C-test, Cloze test, and the MCT were administered to the participants and the results were inputted to some statistical procedures to arrive at answers to the research questions. The descriptive statistics of the administered tests are tabulated in table I.

TABLE I
DESCRIPTIVE STATISTICS

	N	Minimum	Maximum	Mean	Std. Deviation
C-test	52	44.00	78.00	62.4423	10.29642
MCT total	52	31.00	54.00	43.3077	6.68489
Listening	52	2.00	8.00	5.3269	1.85474
Vocabulary	52	8.00	20.00	16.1538	2.56969
Grammar	52	10.00	20.00	16.0962	2.93898
Reading	52	3.00	9.00	5.6538	1.58281
Cloze	52	6.00	21.00	15.2308	3.66550
Valid N	52				

Having collected the data pertaining to the participants' performance on the MCT, Cloze test, and the C-test, the researchers used the Pearson Product-moment formula to estimate the correlation coefficients among the obtained test scores. The results of correlational analyses are reported in table 2. As displayed in the table, all correlations except the one between C-test and the reading comprehension section of the MCT test are significant at 0.01and 0.05 levels. Among the calculated correlation coefficients between the C-test scores and the MCT, however, the strongest relationship is the one between the C-test and the grammar sub-section of the MCT (r=.45), and the weakest correlation is the one between C-test and the reading comprehension section of the MCT(r=.18), which is not statistically significant. Also, there is a correlation index of 0.43 between the C-test and the MCT as a whole, which is significant at the 0.05 level. In other words, there is a statistically significant relationship between the C-test scores and the MCT scores as a whole.

As for the correlations among sub-sections of the MCT, a correlation coefficient of 0.58 is reported between the reading comprehension and the listening comprehension sub-sections, which is the most significant one. The weakest relationship (r=.31), however, holds between the vocabulary and reading comprehension sub-sections.

TABLE II
CORRELATIONS

Test		C-test	MCT total	Listening	vocabulary	grammar	reading	Cloze
C-test	Pearson Correlation	1	.495**	.354*	.395**	.455**	.183	.435**
	Sig. (2-tailed)	-	.000	.010	.004	.001	.195	.001
	N	52	52	52	52	52	52	52
MCT total	Pearson Correlation	.495**	1	.743**	.742**	.764**	.735**	.718**
	Sig. (2-tailed)	.000	-	.000	.000	.000	.000	.000
	N	52	52	52	52	52	52	52
Listening	Pearson Correlation	.354*	.743**	1	.386**	.411**	.580**	.510**
	Sig. (2-tailed)	.010	.000	-	.005	.002	.000	.000
	N	52	52	52	52	52	52	52
Vocabulary	Pearson Correlation	.395**	.742**	.386**	1	.318*	.501**	.544**
	Sig. (2-tailed)	.004	.000	.005	-	.022	.000	.000
	N	52	52	52	52	52	52	52
Grammar	Pearson Correlation	.455**	.764**	.411**	.318*	1	.353*	.533**
	Sig. (2-tailed)	.001	.000	.002	.022	-	.010	.000
	N	52	52	52	52	52	52	52
Reading	Pearson Correlation	.183	.735**	.580**	.501**	.353*	1	.560**
	Sig. (2-tailed)	.195	.000	.000	.000	.010	-	.000
	N	52	52	52	52	52	52	52
Cloze	Pearson Correlation	.435**	.718**	.510**	.544**	.533**	.560**	1
	Sig. (2-tailed)	.001	.000	.000	.000	.000	.000	-
	N	52	52	52	52	52	52	52

*Significant at .05 level.
**Significant at .01 level.

Considering the correlations between subjects' total MCT scores and their performance on its individual component parts, the most noticeable correlation is that between the MCT and the grammar sub-section (r=.76), while the weakest relationship (r=.73) pertains to the one between the global MCT and reading comprehension scores. The same statistical procedures were replicated for the participants' performance on the Cloze test and the MCT. Having collected the data, using Pearson Product-moment formula, the researchers ran a correlational analysis to calculate the correlation coefficients between the Cloze test and the MCT scores. As presented in table II, there is a statistically significant correlation between the Cloze test scores and the MCT scores. The estimated correlation coefficients are reported as 0.71, 0.51, 0.54, 0.53, and 0.56 between the Cloze test scores and MCQ test scores as a whole, listening comprehension, vocabulary, grammar, and reading comprehension sub-sections, respectively. As can be seen in the same table (Table II), all correlations are statistically significant.

Drawing on the matrix of correlations, exploratory factor analysis was also run to identify the hypothetical variables or communalities (factors) which underlie the observed variables

(test scores). The output is presented in tables III and IV and Fig.1. In should be noted that in running factor analysis, the researchers used the "Maximum Likelihood" method of extraction together with oblique or oblimin rotation method. The results encountered one or more communality estimates whose values were greater than one. As such, the resulting solution should be interpreted with caution.

TABLE III
HYPOTHETICAL VARIABLES

Test	Initial
C-test	.358
MCT total	.998
Listening	.972
Vocabulary	.985
Grammar	.989
Reading	.957
Cloze	.524

TABLE IV
TOTAL VARIANCE EXPLAINED

No	Total	Variance	Cumulative
1	4.062	58.027	58.027
2	.934	13.347	71.375
3	.712	10.170	81.545
4	.550	7.854	89.399
5	.419	5.992	95.390
6	.321	4.587	99.978
7	.002	.022	100.000

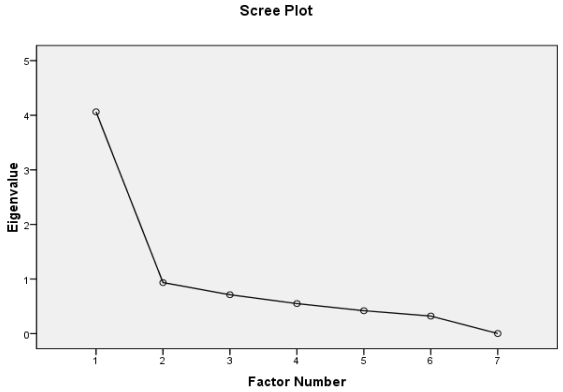

Fig. 1 Scree plot based on the exploratory factor analysis

In summary, the correlation coefficient between the C-test scores (the first criterion test) and the total MCT scores was statistically significant. Consequently, the first research question of this study is answered. In other words, there is a statistically significant relationship between the C-test sores and the total MCT scores.

Regarding the second research question raised at the outset of the study, the obtained correlation coefficient between the MCT scores and the second criterion test scores (Cloze test), was also significant. This has to do with the second research question of the study; there was a statistically significant relationship between the Cloze test scores and the total MCT scores.

Allowing for the aforementioned results, the researchers came to the conclusion that as the correlation matrix displays, the obtained correlation coefficients among the end-of-the-term MCT and the two criterion tests (C-test and Cloze test) were statistically significant. This piece of evidence can be brought to bear on our validity argument. Since the MCT significantly correlated with the criterion measures which in the related literature are reported by many researchers to be valid tests of general English proficiency, it can be cautiously claimed that the MCT is measuring what it purports to gauge. However, considering the limitations of using criterion measures to validate tests, we are not on the right track for making any absolute validity claims. The most serious limitation is that this evidence only considers the extent to which measures of the same ability tend to agree. It does not allow for the equally overriding consideration of the extent to which scores on the test are different from indicators of different abilities.

Furthermore, language tests used as the criterion, the use of which for this purpose may be supported by considerable experience and empirical evidence, cannot, on these grounds alone, be interpreted as valid measures of any particular ability. As such, information about criterion-relatedness seems to be by itself insufficient for validation [5], [28], [29].

APPENDIX A

The MCT

Vocabulary

1. The police went to Mrs. Brown's house and…..her that everything was ok.
A. reassured
B. revealed
C. represented
D. recorded

2. Our rule of law is still…..but it's better than having no laws at all.
A. imperfect
B. stylish
C. moral
D. identical

3. He was looking at me with a (n)…..smile, which I could not stand.
A. superstitious
B. envious
C. mischievous
D. obvious

4. -"I'm going to kick Mike out of the company."
-"Don't get so mad; just…..your horses."
A. make
B. hold
C. keep
D. take

5. She was wearing a very…..dress for her son's wedding.
A. giant
B. medieval
C. joyful
D. exquisite

6. Islamic countries try to have…..relations with one another.
A. inquiring
B. eager
C. cordial
D. meditative

7. Scoring of that goal was a (n)…..of good things to come.
A. omen
B. spirit
C. crucifix
D. devil

8. Did you notice that her hand was…..as she lifted her cup?
A. resisting
B. trembling
C. swinging
D. prowling

9. Residents…..the firemen for their quick action.
A. glanced
B. entitled
C. declared
D. praised

10. He was placed under…..by federal police yesterday.
A. shame
B. reality
C. reputation
D. arrest

11. "There's no point in arguing," he said and walked off….. .
A. eagerly
B. apparently
C. disgustedly
D. impressively

12. Notices about the seminar were…..up on walls all over the university.
A. pasted
B. unloaded
C. estimated
D. punched

13. Whatever his movies, the…..did save a hundred thousand lives.
A. relief
B. deed
C. dilemma
D. tap

14. This drug can be safely used in…..with other medicines.
A. reflection
B. combination
C. precaution
D. motivation

15. My sister plays a (n)…..role in this organization.
A. emotional
B. bright
C. intense
D. prominent

16. Like so many young people before them, they…..to change the world.
A. pulled over
B. broke in
C. set out
D. turned up

17. After the accident both drivers got out and…..their cars for damage.
A. inspected
B. swayed
C. enrolled
D. proceeded

18. The actor had forgotten some of his lines, so he had to…. .
A. quarrel
B. hail
C. spit
D. improvise

19. For the French team, winning tomorrow's game is a matter of national….. .
A. passion
B. honor
C. mastery
D. talent

20. It takes a long time to…..from a heart surgery.
A. rescue
B. overwhelm
C. defeat
D. recover

Grammar
21. You seem very run-down. You'd better…..to work today.
A. don't go
B. not go

C. not to go
D. not going

22. The letters will have all…..by the end of this week.
A. been sent
B. being sent
C. being sending
D. been sending

23. The accident was your fault-you…..have been driving so fast.
A. mustn't
B. wouldn't
C. couldn't
D. shouldn't

24. There used to…..a movie theater here, but they knocked it down last year.
A. have been
B. being
C. be
D. having been

25. Phosphates are chemicals that need…..to most farm land to make the soil rich.
A. adding
B. to add
C. to be added
D. being added

26. Oh! I can feel something…..up my leg! It must be an insect.
A. crawl
B. crawling
C. crawls
D. to crawl

27. They fined David because he had failed…..his electricity bill.
A. paying
B. to have paid
C. having paid
D. to pay

28. As time went go, Peter grew more and more….. .
A. impatient
B. impatience
C. impatiently
D. of impatience

29. I wish I…..able to see my Aunt Rachel when she visited Rome last month.
A. was
B. had been
C. have been
D. were

30. This kind of jacket…..very fashionable these days.
A. considered
B. considers
C. has considered
D. is considered

31. To be honest, I've never seen anyone…..as much as you do.
A. eating
B. eats
C. eat
D. to eat

32. I'm cold, and I feel like…..a cup of hot tea.
A. drinking
B. to drink
C. drink
D. having drunk

33. Alice…..me she…..on a diet since the week before.
A. told, has gone
B. told, had gone
C. said, has gone
D. said, had gone

34. The problem…..discussed by the teachers before it was taken to the principal.
A. is already
B. had already
C. has already
D. had already been

35. I would have bought those sneakers if they…..so expensive.
A. were not
B. had not been
C. would not be
D. have not been

36. "Those children look very….."
"Yes, but sometimes they play….."
A. quiet/noisily
B. quietly/noisily
C. quiet/noisy
D. quietly/noisy

37. I'm not sure, but I think Jim…..a lot of opportunities when he was a young man.
A. might have missed
B. might miss
C. must have missed
D. must miss

38. "I spent the whole day at home."
"Why? You….. swimming with your friends."

A. could go
B. may go
C. could have gone
D. may have gone

39. They said they…..into their new apartment….. .
A. will move/tomorrow
B. will move/the next day
C. would move/tomorrow
D. would move/the next day

40. I can't find my wallet anywhere; it….. .
A. must get lost
B. must have gotten lost
C. should get lost
D. should have gotten lost

Reading comprehension

In most societies the main social unit is a family group, but the forms of such groups differ in different societies. In modern Western societies a family unit is usually small, made up of a man, his wife, and their young children. In the nineteenth century the Western family unit was usually a little larger than this. Unmarried sons and daughters were part of the family even after they were grown up. Nowadays, unmarried sons and daughters do not always live with their parents after they begin going to work.

In other societies we find family units which are quite different. In some places, for example, a man may have more than one wife. When two or more wives live together with their husband in the same home, a child will have brothers or sisters (or both) who have the same father but a different mother. In these families the younger children sleep in their mother's room, but all the older boys may sleep together in a large room. In some societies a married son and his family always live with his father. Chinese families were of this kind. In other societies, for example, in South India, it was the custom for married sons and daughters to live with their mother, not their father. In these societies men did not have much power over their own children. The uncles, the mother's brothers, had power over her children. In other words, men were masters in their sisters' homes, not in the home where their wives and their own children lived. In their wives' homes they were only visitors.

In those societies where married children live with one of their parents (the father or the mother) the family unit is usually large. It is called a joint family. The joint family expands and gets bigger and bigger, until in the end there may be a hundred fathers and mothers living under one roof, with their children. In Calcutta there are still one or two families as big as this. But at some point in history of the family there must be a division. At some time a son will break away to form a new family unit. No home is being enough to contain more than a hundred married couples.

41. What is the passage mainly about?
A. History of western families
B. History of some family groups in the world
C. Four types of joint families
D. Importance of the family group as a social unit

42. In a joint family
A. One of the parents is dead.
B. Two families combine to form a large family.
C. All children are married.
D. The family unit is usually large.

43. China and India are the countries in which
A. The family was not a basic social unit.
B. The uncles had power over children.
C. Married sons lived with one of their parents.
D. Men were masters.

44. Family groups may be different in form according to the
A. Importance of social units
B. The time children get married
C. Family size
D. Family power

45. Which sentence is NOT true?
A. Grown-up single children didn't live with their parents in the 19th century.
B. Joint families divide when they become too large.
C. In some societies men have power only in their sister's home.
D. Eighteenth century Western families included grown-up children.

Sunlight gives us heat. Some of the heat warms the atmosphere, and some of the heat escapes back into space. During the last one hundred years, we produced a huge amount of carbon dioxide. The carbon dioxide in the atmosphere works like the glass in a greenhouse. **It** allows heat to get in, but it doesn't allow much heat to get out. The atmosphere becomes warmer, because less heat can escape.

Where does the carbon dioxide come from? People and animals breathe in oxygen and breathe out carbon dioxide. Trees take carbon dioxide from the air and produce oxygen. We produce carbon dioxide when we burn coal, oil, gasoline, natural gas, or wood. In the last few years, people have burned huge areas of rain forest. This means there are fewer trees and, of course, more carbon dioxide.

Some scientists think the greenhouse effect will make the world hotter. Area near the coast will be cloudier and wetter. There will be more storms. Inland areas will have a little more rain, but because the temperature will be higher, they will be drier. The levels of the oceans will rise. They've already risen by 15 cm (6 in.) since 1880. Maybe they'll rise another 30 cm (1 ft) before 2030. But clouds reflect sunlight back into space, and maybe more clouds will make the earth cooler again. Is the world's climate changing? We don't know.

46. The carbon dioxide in the atmosphere
A. Makes the space hot
B. Allows a lot of heat in
C. Makes the earth cool
D. Escapes back into space

47. Man has made a great deal of carbon dioxide by
A. Increasing the trees
B. Developing a lot of cities
C. Burning the forests
D. Raising a lot of animals

48. What does "It' in line 5 refer to?
A. The glass
B. The atmosphere
C. The greenhouse
D. The carbon dioxide

49. The levels of the oceans by 2030 will be about
A. 60 cm higher than now
B. 45 cm higher than 1880
C. 30 cm higher than 1880
D. 15 cm higher than now

50. Which of the following is NOT true?
A. The amount of carbon dioxide has increased within the last 100 years.
B. Burning natural gas produces less carbon dioxide than burning coal, oil, or gasoline.
C. Some of the heat we receive from sunlight goes back into space.
D. The greenhouse effect would cause more storms in areas near the coasts.

Listening comprehension

[Note: Here, no tape or CD is played. Rather, the teacher reads two texts twice giving students 30 seconds in between to look at the questions. Having listened to each text for the second time, they are given 2.5 minutes to check the correct answer in their answer sheets.]

Text 1

Do you believe that there are some people who can see what is happening somewhere else? Gerard Croiset from Holland could and he used his unusual ability to help the police find missing people anywhere in the world. Professor Sandelius lived with his 24-year-oled daughter Carol in Jopeka, in the United States. One day, something strange happened. Carol disappeared. Police looked for her everywhere. First they showed her photos to everybody in town but nobody knew anything about her. Professor Sandelius was prepared to try anything to find his daughter. He had heard about Croiset and he went to him. "Can you do something to help?" he asked. "No one else can." Croiset never refused to help people. He told the professor that Carol was alive. "I see her somewhere near water and small boats. Now I see her riding with someone in a truck and now in a big red car." After six days, Carol returned home safe and well. Gerard Croiset died in 1980. He never took money for his detective work.

51. According to what you heard
A. The police were not able to find Carol because nobody helped them.
B. People can sometimes see what's happening somewhere else.
C. Some people can see what's happening somewhere else.
D. The police didn't help the professor find his daughter.

52. Croiset saw Carol
A. After six days
B. Driving a truck
C. In a safe place
D. Near some boats

53. What is the main idea of what you heard?
A. Helping people is good.
B. It is possible to see things somewhere else.
C. Police cannot find everything.
D. Some people help others without getting money.

54. Which sentence is TRUE about Croiset?
A. He was worried about his daughter.
B. He showed Carol's photos to people.
C. He always helped people as much as he could.
D. He could find Carol with difficulty.

55. Professor Sandelius
A. Got help from the police
B. Saw his daughter in a small boat
C. Had a big red car
D. Gave some money to Croiset

Text 2

Are you always sure you know what people mean when they try to show their feelings to you? Fear is a feeling that is shown in much the same way all over the world. In Chinese and English, " He went pale' means that the man is either very afraid or has just been surprised. However, 'he opened his eyes wide" is used to show anger in Chinese and surprise in English. Even in the same community, people are different in their abilities to understand and show feelings. Studies show that women are usually better than men at realizing fear, anger, love, and happiness in people's faces. The most difficult feelings to show for people are hating and suffering. Also some people give completely wrong signs of how they feel. For example, they try to show love, but they show hate. In other words, what we think we are saying through words and face may be exactly the opposite of what other people understand.

56. What is the main idea of what you heard?
A. Fear is shown in the same way all over the world.
B. Women are better in understanding feelings.
C. Sometimes people show their anger by opening their eyes wide.
D. We cannot be sure what we understand about other people's feelings.

57. According to what you heard
A. Chinese and English people show their fear differently.
B. People have difficulty in realizing anger.
C. Sometimes what you show is understood differently.
D. People always correctly show what they feel.

58. Which sentence is TRUE?
A. Women can show their fear and anger better than men.
B. People in the same community have the same ability in realizing feelings.
C. Some people like to show the opposite of what they think.
D. Some of the feelings cannot be easily shown.

59. Love is an example of
A. Feelings which many people don't like to show
B. Using wrong signs in showing feelings
C. Using words in showing feelings
D. Feelings which many people can understand easily

60. The hardest feelings to show are
A. Those people say through face
B. Anger and hate
C. Those people say through words
D. Hating and suffering

MCT Answer key
1. A
2. A
3. C
4. B
5. D
6. C
7. A
8. B
9. D
10. D
11. C
12. A
13. B
14. B
15. D
16. C
17. A
18. D
19. B
20. D
21. B
22. A
23. D
24. C
25. C
26. B
27. D
28. A
29. D
30. D
31. C
32. A
33. B
34. D
35. B
36. A
37. A
38. C
39. D
40. B
41. B
42. D
43. C
44. C
45. A
46. A
47. C
48. D
49. B
50. B
51. C
52. D
53. B
54. C
55. A
56. D
57. C
58. D
59. B
60. D

APPENDIX B

The Cloze Test

The cat has a _____1_____ as fascinating and mysterious as the creature itself. The true beginnings of the domestic cat are unknown, but the cat may have first appeared around3000 B.C. in a _____2_____ called Nubia, which bordered Egypt. By 2500 B.C., the cat was domesticated in Egypt. The cat's first _____3_____ in Egypt was Mau. The Mau's _____4_____ in Egypt grew rapidly; she was eventually considered guardian of the temple and was worshipped as a goddess. Besides being worshipped as goddesses, cats also had a practical _____5_____: they kept _____6_____ from overrunning the Egyptian grain store-houses.

The Greeks were probably the first _____7_____ to recognize cats for their mouse- catching talents. When

Egyptians refused to sell or trade any of their cats, the Greeks _____8_____ several of the Egyptian cats and sold the _____9_____ of these stolen cats to Romans. The cat became the _____10_____ of liberty in ancient Rome. By the end of the eleventh _____11_____ cats were popular among sailors because of their rat-catching skills. Sailors admired cats because they _____12_____ disease-infested rats which lived on ships. Many sailors believed that cats possessed special powers that could _____13_____ them at sea.

Although the cat was held in high regard and fancied during _____14_____ times, the cat didn't fare will in Europe in the Middle Ages. Cats were associated with evil, witchcraft, and black _____15_____. Many people believed that _____16_____ regularly transformed themselves into cats. Men and women were killed for helping a _____17_____ or injured cat. During the witch-hunts in Europe many innocent people were accused of witchcraft simply because they owned cats. Black cats were especially feared. Some legends and _____18_____ about cats exist today, like that about the nine lives of cats. Another legend that survived from Europe's Middle Ages into the present states that a black cat crossing one's path brings bad _____19_____.

Today the elegant, graceful cat has become a popular house _____20_____ throughout the _____21_____. The cat is one of the smartest of tame animals, but they are independent and harder to train. Cats are valued for their gentle, affectionate natures. They have _____22_____ memories; they _____23_____ who treats them well and who treats them badly. A cat's loyalty is earned; a cat won't stay where it is _____24_____. They respond to loving owners with loyalty, affection, and respect. Cats are noted for their keen senses: their sharp hearing, sense of smell, and ability to _____25_____ in near darkness. Perhaps Leonardo Da Vinci summed it up best when he referred to the cat as "Nature's Masterpiece."

The Cloze test answer key

1. history
2. country
3. name
4. status
5. function
6. mice
7. Europeans
8. stole
9. kittens
10. symbol
11. century
12. destroyed
13. protect
14. ancient
15. magic
16. witches
17. sick
18. superstitions
19. luck
20. pet
21. world
22. good
23. remember
24. mistreated
25. see

APPENDIX C

The C-test with answers

Read the passages bellow and fill in the missing letters. You should spend no more than 5 minutes on each passage. [Note that the bold letters are the answers. In the original test which the students took each bold letter was substituted with one dash.]

Nothing beats the heat like a refreshing dip in a swimming pool. But wh**en** it comes to wa**ter**, both kids and ad**ults** need t**o** be car**eful**. Susan King's dau**ghters** Alison, 12, a**nd** Christy, 9, a**re** in th**eir** grandparents' po**ol** every d**ay**. King's gi**rls** have ma**de** pool ru**les**, including n**ot** being all**owed** in t**he** pool ar**ea** without a**n** adult, n**o** jumping i**n** the sha**llow** end, n**o** running around the pool and no holding each other under water. "Kids drown quic**kly** and quietly," cautions Jen Costello of the National safe kid's campaign. Even less than an inch of water can be enough. "Parents need to actively supervise children at all times," she says. "Don't take your eyes off them to ans**wer** the phone, to serve food or even to wa**tch** another child."

The global dominance in word processing software held by Microsoft is under threat from a new coalition. The Cillicion-Valley bas**ed** Google and Micro systems ha**ve** announced a formi**dable** alliance. The**ir** plan was t**o** make wo**rd** processing a**nd** spreadsheet pro**grams** available o**n** the Inte**rnet**, in a direct challenge t**o** Microsoft. Indu**stry** observers s**ee** increased compe**tition** in t**he** global soft**ware** market wi**ll** be g**ood** for cons**umers**. The comp**etition** could n**ot** say wh**en** Google wo**uld** Begin carr**ying** Sun's technology, including open office which was launched in 2000.

There are many possible causes of insomnia. Sometimes th**ere** is o**ne** main ca**use**, but of**ten** several fac**tors** interacting toge**ther** will ca**use** a sl**eep** disturbance. THE causes o**f** insomnia inc**lude**: Psychological, **physical** or temp**orary** factors. A la**ck** of go**od** night's sl**eep** can le**ad** to var**ious** problems a**nd** a vic**ious** circle co**uld** develop. Professio**nal** counseling fr**om** a doc**tor**, therapist o**r** sleep specialist can help individuals cope with these conditions.

A popular form of recreation in Britain is attendance at dog racing. The fi**rst** impression o**f** the ar**ena** is attr**active**. However, t**he** races thems**elves** are uninte**resting**; a **few** dogs cha**sing** a tin ha**re**, but thirty two mill**ion** people att**end** them

annu**ally**. Out o**f** two ho**urs**, barely fi**ve** to t**en** minutes a**re** usually dev**oted** to t**he** actual rac**ing**. There wo**uld** be n**o** interest i**n** it if it were not for the betting. Many of the audience pay little attention to the racing, but have their eyes fixed on a board which gives the number of winners.

REFERENCES

[1] R. J. Mislevy, "Test theory reconceived," *Journal of Educational Measurement*, vol. 33, no. 4, pp. 379-416, 1996.
[2] L. F. Bachman, *Statistical Analysis for Language Assessment*. Cambridge: Cambridge University Press, 2003.
[3] A. Hughes, *Testing for Language Teachers*. Cambridge: Cambridge University Press, 1989.
[4] G. Henning, *A Guide to Language Testing*. Cambridge, MA: Newbury House, 1987.
[5] L. F. Bachman, *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press, 1990.
[6] P. Groot, "Language testing in research and education: The need for standards," in J. De Jong, Ed. *Standardization in Language Testing*, London: AILA Review, 1990, pp. 7-23.
[7] M. T. Kane, "An argument-based approach to validity," *Psychological Bulletin*, vol. 122, no. 3, pp. 527–35, 1992.
[8] L. Cronbach, and P. E. Meehl, "Construct validity in psychological tests," *Psychological Bulletin*, vol. *52*, 281-302, 1955.
[9] A. Chapelle, "Are C-tests valid measures for L2 vocabulary research?" *Second Language Research*, vol. 10, no. 2, pp. 157–187, 1994.
[10] E. Tarone, "Research on interlanguage variation: implication for language testing," in L. F. Bachman, and A. D. Cohen, Eds. *Interfaces between Second Language Acquisition and Language Testing Research*, Cambridge: Cambridge University press, 1998, pp. 71-89.
[11] S. Messick, "Validity," In R. Linn, Ed. *Educational Measurement*, New York: Macmillan, 1989, pp. 13–103.
[12] G. Fulcher, and F. Davidson, *Language Testing and Assessment: An Advanced Resource Book*. New York: Routledge, 2007.
[13] L. Bachman, and A. Palmer, *Language Testing in Practice*. Oxford: Oxford University Press, 1996.
[14] T. F. McNamara, *Measuring Second Language Performance*. Harlow: Longman, 1996.
[15] L. Cronbach, "Five perspectives on validity argument," in H. Wainer, and H. Braun, Eds., *Test Validity*. Hillsdale, NJ: Erlbaum, 1988, pp. 3-17.
[16] J. B. Carrol, "The nature of data, or how to choose a correlation coefficient," *Psychometrika*, vol 26, pp. 4342-4372, 1961.
[17] Spolsky, "What does it mean to know a language; or how do you get somebody to perform his competence?" In J. Oller, and R. Richards, Eds. *Focus on the Learner*, Rowley, Massachusetts: Newbury House, 1973, pp. 164-176.
[18] J. Anderson, *Psycholinguistic Experiments in Foreign Language Testing.* St Lucia, Queensland: University of Queensland Press, 1976.
[19] J. W. Oller, J.W. *Language Tests at School: A Pragmatic Approach*. London: Longman, 1979.
[20] H. Farhady, and M. N. Keramati, "A text-driven method for the deletion procedure in cloze passages," *Language Testing*, vol. 13, 191-207, 1996.
[21] Shohamy, E, "*Investigation of concurrent validity of oral interview with cloze procedure for measuring proficiency in Hebrew as a second language*," Ph.D. Dissertation, University of Minnesota, 1978.
[22] Hinofotis, "Cloze as an alternative method of ESL placement and proficiency testing," in J. Oller & K. Perkins, Eds. *Research in Language Testing*, Rowley, MA: Newbury House, 1980, pp. 45-67.
[23] K. Mullen, K, "Rater reliability and oral proficiency evaluation," in J. Oller, and K. Perkins, Eds. *Research in Language Testing*, Rowley, MA: Newbury House, 1980.
[24] R. Grotjahn, C. Klein-Braley, and U. Raatz, "C-Tests: An overview," in R. Grotjahn, C. Klein-Braley, and U. Raatz, Eds. *University Language Testing and the C-Test*, Bochum: AKS-Verlag, 2002, pp. 93–114.
[25] C. Klein-Braley, "Language testing with the C-Test: *A linguistic and statistical investigation into the strategies used by C-Test takers and the prediction of C-Test difficulty*," Ph.D. dissertation, University of Duisburg, 1994.
[26] C. Klein-Braley, "C-Tests in the context of reduced redundancy testing: An appraisal," *Language Testing*, vol. 14, pp. 47-84, 1997.
[27] T. Eckes, and R. Grotjahn, "C-tests: Rasch analyses via the continuous rating scale model," In R. Grotjahn, Ed. *The C-test: Theory, Empirical Research, and Applications.* Frankfurt am Main: Peter Lang, 2006, pp. 167-193.
[28] R. Lado, *Language Testing: The Construction and Use of Foreign Language Te*sts. London: Longman, 1961.
[29] V. Parvaresh, and M. Tavakoli, "Discourse completion tasks as elicitation tools: How convergent are they?" *The Social Sciences*, vol. 4, no. 4, pp. 366-373, 2009.