Language Testing in Asia
a SpringerOpen Journal

METHODOLOGY Open Access

# Doubts on the validity of correlation as a validation tool in second language testing research: the case of cloze testing

Karim Sadeghi

Correspondence: ksadeghi03@gmail.com
English Language Department, Urmia University, Urmia 165, Iran

**Abstract**

The statistical analysis technique of correlation has extensively been used in second/foreign language testing research for a variety of purposes. Most commonly, the technique has been used to validate newly constructed language tests (particularly cloze tests) against previous supposedly valid measures. Tests such validated are meant to substitute the older ones, and based on these newly validated tests, important decisions are made on the candidates' suitability for certain careers, entry into universities, and so on. All such decisions can be appropriate and justified only when the validating technique is the right one. Doubts are cast in this paper on such widespread and unquestioned use of the technique of correlation for substitution purposes. After a short introduction, the meaning of correlation and cloze test is clarified. Then, a few studies in which the technique of correlation has been misapplied are reviewed. Finally, the argument against correlational validation is presented, and suggestions as to alternative validation techniques are offered.

**Keywords:** Validation, Correlation, Cloze testing

## Background

The validity of a test has traditionally been established through matching the test content against the syllabus for content validity and/or through correlating the scorers on a given measure with those of a criterion for empirical or criterion-related validation (Bachman, 1990; Messick 1990). A similar procedure (i.e., inter-item correlation which refers to cross-item correlations between two separate tests) has been in use for accounting for the construct validity of a test. (Messick 1989a, 1989b; 1990, 1994a, 1994b, 1994c, 1996), however, challenged this traditional notion of validation and introduced a new meaning for validity where validity is to do with the meaning of test scores rather than the test itself (Fulcher 1999) and where 'the consideration of values and consequences of score use has an essential role in validity considerations' (Bachman 1990, pp. 241–2). (Messick 1990) believes that the traditional notion of validity is fragmented and re-defines validity as 'an overall evaluative judgement' (Messick 1996, p. 6) 'of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness of inferences* and *actions* based on test score' (Messick 1989a, p. 13). For Messick, whose theory of validity has been so influential (McNamara 2006), establishing validity is essentially a matter of construct evaluation

in which 'the meaning and consequences of measurement' are empirically evaluated (Messick 1996, p. 6). Validity is seen as an integral or 'a unitary concept' (APA, 1985, p. 9) with six distinct but interdependent aspects of content, substantive, structure, generalizability, external and consequential validity, which jointly function as general validity criteria for all educational and psychological measurement (Messick 1994c, pp. 11–12; Messick 1996, p. 9). As such, validating a test for (Messick 1996) involves gathering enough evidence form a variety of sources to counter the two major threats to validity: construct under-representation (where the concern is 'nothing important be left out of the assessment' (p. 4) of the construct in question, which if not taken care of properly, will lead to Type I error), and construct-irrelevant variance (where the threat is making the assessment 'too broad, containing excess reliable variance' (p. 5), which, if not controlled, will result in Type II error). Indeed it may not be an easy task for the test-maker to observe the right balance between these variables, since striving to control for under-representation may push the tester to add more construct-irrelevance variance to the test, over-representing the ability being measured.

The prevalent technique applied in empirical validation the traditional type as well as in obtaining external validity evidence in Messick's framework is that of correlation. This paper aims to clarify the correct and incorrect uses of this technique for validation purposes, bearing in mind that, when applied properly for acquiring relevant validity evidence, data from other multi-dimensional aspects of validity should be added before any claims may be made on the overall validity of a test (in the traditional sense) or of the score interpretation and consequences of its use (in its modern sense). What comes below is a clarification of the concept of correlation and evaluation of whether it has been properly applied in language testing validation studies. Sine this paper is primarily on the use of correlation in cloze tests, a brief account of what cloze is will also be included before reviewing the correct/incorrect applications of correlation in cloze testing.

## The meaning of correlation and cloze

Although the main concern of the paper is challenging the use of correlation for validating language tests/test score meaning, a brief look at the meaning of correlation seems desirable at this point. The following is accordingly an attempt to show what correlation is, what it may be used for and what it should not be applied for.

Correlation is a statistical analysis technique intended to study the relationship between two or more variables. The theory and mathematical basis for correlational analysis were developed by Sir Francis Galton and Karl Pearson in the late 19th century (Hopkins & Glass 1978, p. 111). Since its introduction by Karl Pearson (1857–1936) about a century ago, the technique has been extensively used in 'virtually all empirical disciplines' (Hopkins and Glass 1978, p. 111) and 'almost all social sciences' (Glass & Hopkins 1984, p. 80) to study the relationships between variables of interest to researchers. According to (Guilford and Fruchter 1978, p. 77), 'No single statistical procedure has opened up so many new avenues of discovery in psychology, and possibly in the behavioural sciences in general, as that of correlation.' (Baggaley 1964, p. 2) confirms Guilford and Fruchter: '… the correlational methods have been the most widely used statistical techniques in non-experimental investigations.' He also states that

correlation was used to discover 'many of the relationships involved in the basic laws' of psychology, sociology, education, social work and allied fields (p. 1). Such a widespread use of correlation has not, however, prevented the technique from its improper application.

(Garrett and Woodworth 1958) contend that 'Correlation method is used to examine the relationship of one variable to another' (p. 122) and that 'correlation is simply a measure of mutual association between two variables' (p. 180). Similarly, (Glass and Hopkins 1984, p. 79) assert that correlation is a measure to 'describe the degree of relationship between two variables.' (Goehring 1981, p. 145) provides an example to illustrate the meaning of correlation: 'Students' history scores, for example, might be paired with their English scores to determine whether students who earn high scores in one subject also earn them in another.' Simply put, the resulting figure when two variables are correlated (variously known as correlation coefficient or coefficient of correlation or simply correlation) is an index of how far the two variables rank the observations similarly. As all these references indicate, the concept of correlation and the resulting coefficient indices are, therefore, intended to show the degree of go-togetherness between variables or the degree of similarity in the way they rank scores.

The relationship between variables may be positive, negative or none at all. This means that the two variables may rank the observations in the same or a similar order, in an opposite order or so differently that no pattern is discernible. For two variables correlated, the degree of correlation will fall somewhere between +1.00 (the highest positive correlation) and −1.00 (the highest negative correlation). If there is no relationship between the variables, the correlation coefficient will be 0.00. It is generally accepted that correlation coefficients in the range of 0.00 - ±0.20 appear by chance only, those in the rage of ±0.20 - ±0.40 show that some kind of relationship exists, coefficients ranging from ±0.40 to ±0.70 show a considerable degree of relationship, those between ±0.70 and ±0.90 show a marked and definite relationship, and coefficients ranging from ±0.90 to ±1.00 show a very consistent relationship (Goehring 1981, p. 149). However, these indices do not have the same meaning in all situations: depending on whether subjects are heterogeneous or homogeneous, on how reliable the tests are, and also on the nature of variables and the purpose of correlation, the resulting coefficient will be interpreted differently (Garrett & Woodworth 1958, pp. 171, 176; Goehring 1981, pp. 149–50; Brown, 1988). Therefore, as discussed below, a correlation coefficient which is statistically significant, may be meaningful in one context but not meaningful in a different situation. Depending on different types of variable (continuous, rank-order and nominal), different types of correlation coefficient are calculated. Here we are not concerned with different types of correlation coefficient, nor it is our end to show how to calculate them. We are rather concerned with what the resulting number, i.e., correlation coefficient, means and how it should and/or should not be interpreted.

According to (Guilford and Fruchter 1978), correlation coefficient is 'an index of the predictive value of a test' (p. 87) and that without knowing that number, 'it would be impossible to make predictions' (p. 77). In non-technical terms, given a group of subjects' scores in a test, Reading Paper of IELTS, for example, and the degree of relationship between that test and another test, Listening Paper of IELTS say, one will be able to predict that group's scores in the latter test. Namely, the resulting correlation coefficient will give us some information about the way scores on the variables in question

are ranked. It is this predictive or ranking aspect of correlation coefficient that has been misapplied in the field of second/foreign language testing. That is to say, while it will be quite logical to predict the ranking of one's score in Listening based on their Reading score provided that the correlation coefficient between these two variables is known, it makes no sense to argue that one's Reading score can be used to talk about their Listening ability even if the degree of correlation is +1.00. That is to say, replacing Reading ability for Listening ability violates a logical assumption that not all language aspects are inter-related. Namely, the resulting correlation coefficient should not be interpreted as a validity index, which is the interchangeable term used when the criterion-related validity of a newly made test is investigated (Brown, 1988: 104), and that is probably why (Garrett and Woodworth 1958, p. 180) assert '*r* alone gives us no information as to the character of association'.

A further caution voiced by statisticians while interpreting the coefficient of correlation is that the existence of correlation between two variables does not necessarily show that one variable is caused by the other (Glass & Stanley 1970, p. 12; Hopkins & Glass 1978, p. 144; Glass & Hopkins 1984, p. 104; Brown & Rodgers, 2002, p. 190). For example, there may be a high correlation between one's Speaking and Writing ability; nevertheless, it cannot be claimed that higher Speaking ability results in better Writing performance or vice versa, since many illiterate people can fluently speak their mother tongue but may not even be able to write their names. A similar problem may exist even for educated people such as the author of the paper whose first language (Azeri Turkish) does not have a written form or at least is not formally taught in Iran's schools. Furthermore, the existence or even the lack of correlation may not rule out the possibility of a causal link between two variables (Glass & Hopkins 1984, p. 106). For instance, if as (Carlisle 2007) contends, we accept that morphological awareness leads to better reading comprehension, then the expectation is to find a high degree of correlation coefficient between the tests intended to measure these constructs; however, it may be possible to arrive at an insignificant correlation *in a particular situation* for a variety of factors including the unfamiliarity of the testees with the measurement tools. In such a case, the lack of a significant correlation will not rule out the possibility of a causal link between one's morphological awareness and reading performance. All in all, there may be correlation between two variables without any causal relationship; there may be correlation with the possibility of a causal link; and there may be a cause-and-effect relationship between two variables without any correlation between them. In brief, 'There can be correlation without causation and vice versa' (Glass & Hopkins 1984, p. 107).

Also important to note is, as (Goehring 1981, p. 148) rightly emphasises, that a positive or negative relationship may be observed between variables that are not logically related. (Brown and Rodgers 2002, p. 184) confirm this statement by claiming that 'one problem … in the development of correlation coefficients was that they seldom turned out to be exactly zero' even with random numbers. Such a coefficient called 'spurious' or 'nonsense' may be obtained when one correlates a group's test scores in the courses they have successfully passed, for example, in English Grammar and Research Methods in ELT (both being part of the requirements for earning a BA in ELT in Iran's universities). The resulting correlation coefficient can be anything between +1 and −1 but it cannot be used as evidence to prove any meaningful relationship between the variables in question.

As for the meaning of cloze, it comes from the gestalt concept of 'closure' and was first introduced six decades ago by (Taylor 1953). Based on the concept of 'closure', it will be possible for a person to perceive an incomplete geometrical shape by 'closing' the gaps in it. Taylor, who originally recommended cloze to be used for measuring text readability in L1, argued that readers with good reading compression ability should be able to 'cloze' the gaps (blanks) in a truncated passage (with a deletion rate of every *n*th word, *n* ranging from 5 to 7). The idea was soon embraced by L2 scholars and many proposals were made to use cloze tasks for measuring a variety of skills in L2. Indeed, inferences about what cloze tests measure were based by and large on the amount of correlation cloze results showed with other tests of reading comprehension (Oller & Jonz, 1994; Jonz & Oller, 1994), language proficiency (Hale et al. 1989; Stubbs & Tucker 1974), readability (Taylor 1957), intelligence (Taylor 1957; Rankin, 1970) and the like. Since that time, different varieties of cloze have emerged, including standard or fixed-ratio cloze, multiple-choice cloze, rational cloze, discourse cloze, etc., and all sorts of experiments have been conducted on almost all its varieties to substantiate what such incomplete texts may be measuring. This paper is indeed meant to clarify that conclusions arrived at the end of most such studies on what cloze tests measure may not be justifiable due to the inappropriate technique (i.e., correlation) used for this purpose.

Having made the meaning of correlation and cloze tests clear, we will now look at a few studies in the field of second/foreign language testing with particular focus on cloze testing in which the technique of correlation has been misapplied leading to inappropriate conclusions.

## Methods

### The misuse of correlation in second language research

In the field of second/foreign language testing research, most often ESL/EFL researchers and testers have correlated test results in an attempt to empirically validate newly made local tests, and based on the correlation indices so obtained, they have taken inappropriate decisions. In such a kind of traditional validation, 'if the performance on the two tests is sufficiently similar, then the new test can be said to have been "validated" by means of the old one' (Baker, 1989, p. 22) without there being 'little external evidence that these tests are measuring' the same abilities as their well-established criterion (Mackey & Gass 2005, p. 108). Many researchers have suggested to *replace* or *substitute* one test with another only because the observed relationship between the tests in question have been moderate or strong, neglecting any checks for content and construct validity (in the traditional sense of validity), nor have the researchers asked themselves whether the tests have been constructed for the same 'purpose' and for the same 'target population' (Baker, 1989, p. 23) and little effort is seen in such validation studies for ensuring other aspects of validity mentioned above. As an example, (Stansfield and Hansen's 1983, pp. 29, 31) review of relevant literature concludes that cloze is a valid procedure because cloze tests in different studies have correlated highly with other lengthy measures of second language proficiency. Similarly, regarding cloze tests as measures of reading comprehension, content knowledge, and text readability, (Greene 2001, p. 82) argues that the validity of cloze is established for over 30 years 'on the basis of correlation between cloze scores and results of other forms of assessment'.

Interestingly, while some researchers have assigned explicit criteria on the magnitude of correlation before the substitution of the tests is possible, others have acted very subjectively and regarded coefficients considered weak or moderate in other studies as strong in their own case, allowing them to make the offer of replacement. (Oller 1973) for instance, who was one of the strong advocates of using correlation for substitution purposes, argues that if a test in visual modality is correlated highly with a listening test, one test can conveniently substitute the other (p. 115). He does not, however, indicate how high the degree of correlation should be. (Ilyin et al. 1987, p. 150), on the other hand, set a criterion for the amount of correlation before one test can replace another: 'We generally consider that tests should correlate 0.80 or above to substitute or predict another test'.

This paper questions the use of correlation for such a purpose and discusses why conclusions based on correlational studies in second/foreign language testing may sometimes be misleading. In this section, a number of instances of the inappropriate application of the statistical tool of correlation by renowned researchers in the field of second/foreign language testing are reviewed, with a special emphasis placed on cloze testing. The reason why cloze literature has been selected as the departure point for showing the ineffectiveness of correleation for validation purposes is firstly to do with the extensiveness of validation studies on cloze compared to other language-related tests. The higher number of cloze-related validation studies partly originates from the fact that researchers have in a sense been trying to establish the 'more economical' cloze as a valid measure of 'less practical' tests of language proficiency (which, according to (Richards, Platt & Platt 1992), refers to a person's skill in using a language including his ability to read, write, speak or understand it), reading comprehension and the like. The second motive for my choice of focus on cloze is related to the observation that our understanding of what cloze tests measure is to a great extent based on what types of tests they correlate with. Although some researchers (e.g., Farhady, 1983; Lee, 1985; Sadeghi 2003) have claimed that it still is not clear what a cloze test does, others have continued to compare close test results with other language-related constructs such as reading and language proficiency in a hope to understand what cloze measures, based on how it correlates with other language-related constructs. In a review of the literature on what cloze measures, (Sadeghi 2003, p. 103) noted that the final '… picture of what cloze tests measure is not available, and whatever information exists is incomplete.'

What comes below is a brief account of a selection of the studies in chronological order using correlation for validating relevant experimental tests. Empirical validation of cloze tests dates back to the early days when cloze procedure was formally introduced by (Taylor 1953). (Taylor 1957, p. 25), for example, correlated cloze results with those of reading comprehension tests and, based on relatively high coefficients, he considered cloze tests 'valid indices of comprehensibility of English prose'. In this study and most of the others reviewed below where a cloze test has been validated, decisions on its validity or the lack of it have simply been based on the resulting correlation coefficient without any concerns on whether the measures being correlated were otherwise related in terms of content, the purpose for which they were used, and the construct which they supposedly measured. This final issue has remained unnoticed mainly because no one has been able to assert for sure what it is that is measured by a cloze test.

About a decade later, (Bormuth 1967) correlated cloze tests with multiple-choice reading comprehension tests and, based on a high validity coefficient of 0.95, concluded that cloze was a valid measure of reading comprehension. (Rankin and Culhane 1969) conducted a similar experiment and correlated five cloze tests with five multiple-choice reading comprehension tests based on the same passages and obtained coefficients in the range of 0.54 to 0.77. Based on such moderate correlations, they surprisingly concluded that 'the cloze procedure is a highly valid measure of reading comprehension' (p. 196). Similarly, (Oller and Conrad 1971) cloze test showed an overall correlation of 0.88 with subparts of the UCLA ESLPE (a language proficiency test used to place non-native students attending UCLA). Based on this figure, they concluded 'the cloze method is a very promising device for measuring ESL proficiency' (p. 183).

A further example is (Jonz 1976) who used a multiple-choice (M-C) cloze as an experimental adjunct to a placement test. The cloze test took 20 minutes to administer and score and the correlation between the cloze test and subparts of the placement test was substantial. (The correlation that was regarded as substantial was only 0.54 for a group of 33 subjects.) Based on this evidence, he concluded that the three-hour long placement test and the multiple-choice cloze 'appear to be measuring very similar things' (p. 261) and that a short cloze test can be used instead of a long placement test 'without sacrificing the quality of the information derived from that testing' (p. 255). There is, however, no evidence that the quality of information derived will be the same except for a correlation figure, which shows only a degree of relationship, and nothing about the nature of this association. With a similar goal in mind, (Hinofotis 1980, p. 121) correlated cloze tests with a second language proficiency (TOEFL) and a placement (CESL) test, and based on the high correlations obtained, he argued that cloze can be 'a viable alternative procedure for placement and proficiency testing'.

The observation that cloze tests have so conveniently been recommended to substitute other tests of reading comprehension, placement and proficiency in the cited studies stems partly from the fact that cloze tests are easily constructed, administered and scored. However, one needs to ask the question of whether a measure's 'economicality' is to take priority over its validity for the purpose it is used for. Favouring cloze tests over other testing techniques because of the economy factor has helped to misinterpret correlation results in the case of many studies reviewed here.

A well-known figure carrying out many experiments with cloze for his PhD is Alderson, who published his findings in a series of papers. He correlated cloze tests with subparts of a language proficiency test (ELBA) and concluded that cloze tests were mainly tests of lower order proficiency because they correlated highly with tests of grammar and vocabulary (which were supposed to test low-level-skills), but not significantly with tests of reading comprehension (which were supposed to test high-level skills) (Alderson 1979a, p. 225, 1979b, p. 205, 1983, pp. 219, 221). Although Alderson did not explicitly mention that cloze tests could substitute grammar and vocabulary tests, he clearly stated that cloze tests did not measure high-order reading comprehension skills because they were not highly correlated with tests of reading comprehension. Such an assertion implies that, for Alderson, while cloze tests could replace those of grammar and vocabulary, they were not valid measures of reading comprehension and were thus unable to substitute reading comprehension tests. With a similar experiment but different findings, (Chapelle and Abraham 1990) correlated an M-C cloze test with

a reading test. Based on the observed correlation of 0.86, they concluded that the M-C cloze may be a measure of reading comprehension. Their conclusion, however, is not so strong as that of those who have suggested that one test replace another simply because of a high degree of correlation.

(Hanania and Shikhani 1986) found high correlations between four cloze tests, a composition, and the American University of Beirut Entrance Exam (AUB EE), and concluded that the high correlations between the tests showed their validity, alluding that they could all replace one another. The question of which test is to be regarded as the criterion and which ones as the experimental remains open in their study, however. Likewise, (Fotos 1991, p. 333) correlated the results of a cloze test with an essay, and based on a significant correlation between the two, she recommended that a cloze test be used 'as a substitute for an essay test' without worrying about whether these two instruments could tap the same abilities and whether it would be logical to use a completion task for the purpose of measuring one's writing ability. Applying a more logical procedure, however, (Sciarone and Schoorl 1989, p. 426) correlated two similarly constructed cloze tests and noted that they measured the same thing because of a high correlation. Although what that 'same thing' is was never revealed in this study, and most similar studies, such a validation here seems more justifiable because the tests being correlated were both of the same nature (by which I mean both were tests of the same entity, i.e. 'cloze' tests; both were based on similar content, i.e. on the 'same passage'; and both included tasks requiring similar performance and/or mental activities) and intended for the same purpose, whatever it may be. So they could be considered substitutable tests measuring the same construct (whatever it may be) only if the resulting correlation were very high and near +1.00. To account for the overall validity of a test such validated, apart from this correlational evidence, other forms of evidence should be added, however.

Some other studies in which one test has been suggested to substitute another only because of a moderate to high degree of correlation include (Carroll et al. 1959), (Irvine et al. 1974) and (Shohamy 1983). While the majority of the studies cited above have recommended one test to substitute another mainly because of a high degree of correlation, some researchers have noticed a low correlation coefficient between two tests and have suggested that the tests not replace each other. This means that if correlation coefficients had been high enough and statistically significant in those studies, the researchers would certainly have recommended one test to substitute the other. Readers are, however, reminded at this point that, statistical significance does not always convey meaningfulness and is not to be confused with it as Brown (1988) and Brown and Rodgers (2002) caution against. Namely, when the variables correlated are of supposedly similar nature and content, before any firm judgements are made about the meaning of the resulting correlation indices, the statistically significant correlations should be squared to get the 'coefficient of determination' (Brown & Rodgers, 2002: 190) which is indeed an index of the shared variance between the variables. Deciding on the magnitude of the shared variance which is acceptable will depend on numerous factors including the purpose of correlation, however. Furthermore, tests of different nature with higher reliability indices tend to rank the observations similarly, and the resulting correlation coefficients will most likely be statistically significant, but such high and significant correlations are not to be considered enough evidence for

concluding the tests measure the same entity. Therefore, as Brown (1988, p. 122) rightly warns us 'just because a result is statistically significant does not mean it is necessarily meaningful'. A small number of studies which have found insignificant correlations between cloze and relevant criterion measures and have therefore refrained from recommending to substitute them are reviewed below.

(Cranney 1972–73), as an example, correlated fixed-ratio and tailored cloze tests with the Cooperative Reading Test. Based on the coefficients of 0.52 and 0.51, it was concluded that cloze tests measured different aspects of reading from those measured by the Cooperative Reading Test. Although the researcher noted that correlation in neither case was strong enough and that cloze tests measured different things from the reading test, quite strangely, the cloze tests were judged to be superior to the reading test used because, the researcher argued, cloze tests were easily constructed, and the writer's meaning was not misinterpreted (p. 64). (Porter 1978, p. 333) similarly obtained no strong correlation between two cloze tests (in one of which deletions started one word earlier) and concluded that cloze tests could not be considered equivalent, and that 'they must be testing different things' (p. 337).

(Klein-Braley 1983) conducted a similar experiment and because the correlations between cloze tests were low, she concluded that cloze tests were not parallel tests and did not measure the same thing (p. 223). (Bensoussan and Ramraz 1984) refrained from substituting one test with another simply because the correlation between them was moderate. They correlated a 'fill-in' or M-C cloze test with a fixed-ratio cloze and another multiple-choice reading test for 354 subjects and the resulting moderate correlation ruled out the possibility of substitution. And finally, while (Lutji Spelberg et al. 2000) considered the relationship of 0.4 between two tests of the Dutch Reynell Test and the BELL Test 'relatively high' (p. 311) and whereas (Jonz 1976) regarded coefficient of 0.54 as 'strong' and proposed the substitution of one test for another, (Mauranen 1989, p. 341) claimed that the correlation of 0.6 - 0.68 was only 'moderate' and concluded that her 'semantic' cloze tested different aspects of reading comprehension from what other multiple-choice reading tests measured.

Apparently the researchers reviewed above use the terms 'prediction', 'ranking' and 'substitution' interchangeably, and that is perhaps why so many mistakes have been made in interpreting correlation coefficients. As it was made clear in the previous section, it may be possible to predict or rank, but not substitute, a subject's score in one test given their scores in the other and the degree of correlation between these two tests (Mackey & Gass 2005). Substitution is, however, possible as far as it can be implied from the concept of correlation only if both variables are of the same nature (i.e., with arguably similar content and requiring similar processes for completing the tasks) and intended for the same purpose (for testing reading comprehension, for example) and if the degree of correlation is perfect or near perfect. The substitution proposed in most cloze testing research referred to above cannot be sustained mainly because the variables correlated are of different nature, intended for different purposes, and the observed correlation is rarely near perfect. Moreover, there seems to be a lack of any acknowledgement for the evidence from the other aspects of validity in its modern sense.

The review above does not, however, mean that all studies in language testing have applied correlation improperly. There are numerous other studies in which the technique of correlation has been used correctly. The most recent studies in this category

are (Choi et al. 2003), (Trites and McGroatry 2005) and (Roever 2006), to name a few. (Roever 2006), for instance, compared web-based tests of implicatures, routines and speech-acts and found that the tests were related, without claiming that they could be substituted. (Correlation in that study was used to correlate tests that were related to one another pragmatically, supposedly intended to tap a similar construct.)

Using correlation for a similar purpose, (Trites and McGroatry 2005) correlated test scores intended to show the differences between three types of reading, i.e., Basic Comprehension, Reading to Learn, and Reading to Integrate. Since all their tests were meant to measure different layers of the same construct, a highly positive index of correlation could have been interpreted as all reading types fulfilling similar functions, while the lower observed correlations between tests of Basic Comprehension and the others suggested 'a possible distinction' (p. 174) between them. This means that the type of reading called Basic Comprehension could possibly be a distinct kind of reading as compared to Reading to Learn and Reading to Integrate, both of which could be assumed to tap more similar abilities (somewhat different from what was tapped by Basic Comprehension) for having a relatively high correlation with each other.

(Choi et al. 2003) used correlational techniques to understand the relationship between two language proficiency tests at Seoul National University, one being paper-based (PBLT) and the other computer-based (CBLT) with the same subtests of listening, grammar, vocabulary and reading comprehension. As the underlying constructs were the same in their study, their suggestions of comparability of the tests involved seem tenable.

Most of what was cited above is only a small sample of the research on foreign/second language testing in which the technique of correlation has been misapplied and the results misinterpreted. The next part of the paper is intended to clarify why such a use of correlation for substitution purposes is not allowed and is unjustified.

## Results and Discussion

### The argument against correlation as a validation tool in second/foreign language testing

There are several grounds to question the inadequacy of correlation as a validation tool in which one test is suggested to replace another, as was the case in most of the studies cited above. First things first, the concept of correlation conveys merely whether two variables tend to rank the observations in the same manner. The presence of similar rankings, however, does not mean that the variables are the same or replaceable. Rather it means that, based on the data available from one variable, one may be able to make predictions about the ranking of observations in the other. That is to say, the concept of correlation does not imply that two variables that can highly predict each other can be regarded as the same and interchangeable. The variables are essentially different, and no matter how highly they are correlated, they cannot be considered the same.

In other words, even if the correlation coefficient between the scores on two tests is +1.00, the scores on one test can then predict the scores on the other with perfect confidence and they are ranked exactly the same in both variables without implying that they measure the same ability and can thus replace each other as if they were two halves of an apple. Although the idea that two instruments which measure the same thing may be highly correlated is not disputed here, the reverse side of the issue – i.e.,

that because the scores on two instruments are highly correlated, they measure the same thing – is challenged in this paper.

As mentioned earlier, one test may substitute another only if two variables are of the same nature, used for the same purpose and are of the same level (intended for the same target population), i.e., both are tests of reading ability intended for advanced learners and requiring similar mental processes and possibly comparable products, for example. The substitution of one test with another will be acceptable in such a case providing that the magnitude of the correlation coefficient between the two is really big (and in the positive direction). If one desires to lose no information by substituting one test with another, then the degree of correlation should be +1.00. Such a use of correlation can be found in split-half and parallel-forms techniques for obtaining test reliability, which is indeed the degree of correlation coefficient between two halves of the same test or the parallel ones. Suggestions for substituting one test with another will be justified in similar situations if the validity coefficient is significantly high. When a test is being validated using (Messick 1989a) validity framework, however, the evidence from correlation will form only a small part of the required validity evidence to be complemented by evidence from other aspects of validity mentioned earlier in the introduction. Nevertheless, in most of the studies reviewed above, the recommendation for substitution has been made without the two tests having the same nature or intended for the same purpose and target population and with only a moderate degree of correlation with almost no evidence form other aspects of validity. This kind of validation is not supported simply because correlation is a technique not intended for and unable to serve such a purpose alone, and as (Brown and Rodgers 2002, p. 191) rightly emphasise, 'the only statement you can safely make about any correlation coefficient is that it shows *X* degree of relationship between the two sets of numbers involved'. Even when correlation is the right technique used for empirical validation, other sources of evidence should be sought before any firm recommendations are made on the validity of a test or the meaningfulness of the resulting test scores.

Another counterargument against the validity of correlational validation (when used improperly) is that if we were permitted to substitute one test with another simply because two tests correlated highly, the whole notion of educational measurement would be under question. Consider the following situation: Few people will dispute the idea that top and clever students tend to get better scores than average students than slow students in many school subjects. It follows that a group of students' scores in these very subjects will tend to be ranked similarly or to correlate highly with one another, because in all of them top students always tend to be ranked first, then average students, and finally weak students. This means that a high correlation coefficient is *generally* expected to be observed between any group's scores in language-related subjects like grammar, reading comprehension, writing ability and the like and even in non-language related subjects such as maths, physics, history, geography, and so on (acknowledging the fact that there will of course be some pupils with a stronger ability in one subject area but doing terribly in another). Now if we were allowed to apply the technique of correlation as the researchers in the field of second/foreign language testing have done, we would be able to substitute the tests, in at least some of the subject areas, with one another. For instance, it would be perfectly justifiable to substitute a writing test with a reading one and by the same token a geography test with a geology

one (both being related to the science of the earth having more in common than tests like cloze and general language proficiency do simply because it is, as (Sadeghi 2003) review of literature shows, not clear yet what cloze tests measure). All this essentially means that we can give a reading test to a group of students and then talk about how good they are in writing; and along the same lines, we can give them a geography test and then talk about how good they are in geology, which will eventually mean that an algebra test may be used to measure learners' knowledge of geometry (both being branches of maths) and so on. Obviously, the distance between most of the subject areas grouped together in the preceding sentence is not so big as that between cloze tests and the range of other tests with which cloze scores have been correlated. What this mini-discussion implies is that educational authorities can give only a limited number of tests (and only one test in its extreme sense) during a programme, and they will then be able to obtain information about the candidates' abilities in some other fields. Such a presumption arrived at by applying the concept of correlation in a way that has been applied by language testers will not only be against all educational measurement principles and standards, it will also be illogical and nonsense. Still such a practice in language testing has prevailed for a long time managing to escape the inspection of researchers.

More convincingly, few people will raise an objection to the expectation that there will be a relatively high degree of correlation between test scores in English (or any other school subject) for eighth graders, ninth graders, and so forth for the same group of students. Simply put, English scores of ninth graders will substantially correlate (in the majority of cases, if not always) with those of eighth graders. The same seems to be applicable for any other subject. Now based on the correlation coefficients such obtained, it can conveniently be claimed that the English tests for eighth and ninth graders are substitutable and measure the same thing (to follow the way correlation coefficients have been interpreted in the majority of the research pointed out earlier in the paper). Deriving such a conclusion will be ridiculous even to laymen and will be an insult to professionals and educational authorities. To reiterate the point, high degrees of correlation do not necessarily mean that the variables are or do the same. As it was noted earlier, even supposing that the validity evidence so produced is dependable enough, to arrive at a clearer picture and to make a final decision on the validity of the experimental measurement tool will require 'the gathering of sufficiently compelling [other] evidence' (Messick 1996, p. 4) about content, structure, generalizablity, substantive and consequential dimensions of construct validity.

The third reason why the use of correlation for validation purposes is under question is suggested by the contradictory results reported by the application of the technique in different contexts. In other words, the observation that one test such as cloze has correlated very highly with another test (listening comprehension, for example) in one context but no so highly in a different context may by itself indicate that the technique of correlation is not able to explicitly reveal the nature of the relationship between the tests in question. This observation gains more colour when the same tests correlated in similar contexts have also led to different degrees of correlation coefficient. Literature cited in the previous section revealed that the degree of correlation between two variables might be very different in different situations. This difference may be due to differences in sample size, subject types and levels, contents and difficulty levels of the

tests correlated. And simply because of such variations, no fixed degree of correlation can be assumed to exist between two variables. Even if there were such a constant degree of correlation, it would not show anything about the degree of equivalence or substitutability of the tests, unless other criteria such as the nature of the attributes being correlated, the purposes for which they are used, the level they are intended for and comparability of contents are established first. For a test to be regarded as a valid measure and to produce meaningful scores with interpretations and actions that are adequate and appropriate, apart from such external evidence, support should also be provided from many other internal aspects of validity unable to be taken care of by correlation coefficients.

The final reason why I believe correlation alone should not be used as the yardstick for judging the substitutability of one test with another is to do with the vicious circle observed in the validation of this kind. Namely, while sometimes the established valid tests are used as criteria against which other newly made tests are validated, these latter tests become criterion measures themselves later on against which other *a priori* valid tests are validated. Such a 'back-validation' process produces twice as many problems because, first of all, correlation (the validity of which for validation was shown to be questionable here) is used for validation in the first instance; and secondly, the tests validated in this way become supposedly valid criteria themselves, against which some other tests (new or established) are to be validated still using the improper technique of correlation. The question of whether the criterion test is itself valid or not is usually neglected. Without the criterion test being valid *a priori*, the validated test will itself be invalid even if the validating procedure used is the right one, however. In any case, the newly validated test cannot be more valid than its criterion and cannot, therefore, be used to 'back-validate' the criterion measures.

As an example, the forerunner of IELTS, i.e., ELTS (English Language Testing Service), was introduced by the British Council in 1980 as a substitute for EPTB (English Proficiency Test Battery). Part of validating ELTS involved concurrently validating the new test with test results from other proficiency tests including EPTB and ELBA (English Language Battery) (Davies, 1990). The outcome was that ELTS was regarded as a more valid measure of language proficiency and superior to its criterion measures. The proficiency tests developed after ELTS were accordingly validated against ELTS rather than its predecessors, and even the statistical validity of the predecessors themselves now owe it to the strength of their correlation with their substitute. The tests used by (Hanania and Shikhani 1986) and (Hale et al. 1989) also have similar scenarios.

It should be acknowledged, however, that there has been some concern about the improper use of statistics in general and correlation in particular in the field of language testing which has gone unnoticed by fellow researchers. For example, (Brown 1994, p. 194) asks '… how can we justify relying almost exclusively on the statistical techniques developed for other fields when we are endeavouring to understand how to test *language*?' Talking exclusively about the kind of validation which is the focus of the present paper, (Brown 1983, p. 238) criticises the suitability of such a method and expresses that such correlation coefficients 'do little to express how and exactly what cloze is testing.' While themselves applying the technique of correlation for a similar purpose, (Sciarone and Schoorl 1989, p. 434) attack (Lado 1986) use of correlation to

conclude that exact-word cloze (in which credit is given to only the exact word in the original text) measures the same thing as acceptable-word cloze (in which any word which makes sense in the blank is given credit): '… the existence of high correlations between exact scores and acceptable ones does not imply that high acceptable scores go hand in hand with high exact scores. It only means that subjects will be ranked in much the same order'. (Johnson 1981, p. 77) also looks at the conclusions drawn based on statistics with much doubt, and argues that the results based on statistics are not justifiable because statistics are data and that 'valid conclusions can only be reached by process of argument.'

The present researcher, however, does not agree that statistics and correlation should be avoided in language testing research. What he insists on instead is that right statistical procedures should be applied for the right purposes, and caution should be taken in interpreting the results so as not to interpret them in a way not allowed by the relevant statistical concepts. Doubts on the use of correlation for validation purposes in second language testing have also been raised by (Sadeghi 2002a, 2002b, 2002c, 2002d, 2003, 2004, 2006, 2010) whereby qualitative investigation has been proposed to accompany quantitative research. Based on the arguments made here, the researcher urges that the fellow researchers reconsider the appropriate use of statistical tests and particularly correlation in their validation studies. It should also be borne in mind that whenever applied correctly, correlation can provide us with some important information about the validity of a test (or test scores); but more important to remember is that this piece of validity evidence should be integrated with many other pieces from different sources and the final validity value of a test (or its scores) will be determined by an interaction of many different types of evidence, only part of which can be contributed by correlational validation.

## Conclusions

The paper began with a concern on the application of correlation for validation purposes in language testing whereby one test (usually a cloze) has been suggested to replace another. After clarifying the traditional and modern notions of validity as well as the concept of correlation, a small number of studies in which researchers have misused the technique of correlation for validation purposes were reviewed. Drawing on the lessons to be learnt from cloze validation studies, the paper concluded with the researcher's argument against the use of correlation alone for validation in language testing.

It is finally suggested here that because mere correlation does not seem to be enough for validation in language testing, researchers ought to carry out qualitative content, construct, level and objective analysis of new tests based on information obtained from students, teachers, testers and researchers. In short, evidence from many sources (following Messick 1989a) should be combined to gain access to a better picture of validity. Depending almost exclusively on figures as a result of quantitative analysis seems to show little of the real entity of the variables being studied and a lot of care should be taken in interpreting such data. 'Researcher research', which 'refers to the researcher's investigation of his/her own internal thought processes at the same time as he/she is taking a test' has been proposed as an alternative validation tool (Sadeghi 2004, p. 85). Through such a validation procedure, the researcher, involving himself/herself in the

actual test-taking process, directly experiences what others can only observe. The application of such a technique to cloze tests reveals that different cloze items make different demands on the reader (ibid).

To sum up, combining quantitative analysis with qualitative research seems to be more promising and is hoped to provide us with a better understanding of the nature of the unknown attributes being researched. The need for qualitative investigation has been voiced and practiced in different forms by a good number of researchers including (Gefen 1979), (Shohamy 1983), (MacLean 1984), (Lado 1986), (Markham 1987, Markham 1988), (Jafarpur 1995, 1996), (Connelly 1997), (Storey 1997), (Sasaki 2000), (Babaii and Ansary 2001) and (Mackey and Gass 2005) among others. (Markham 1988: 48), for example, asserts that 'purely quantitative techniques do not necessarily mirror the internal thought processes of the subjects.' Similarly, (Storey 1997, p. 214) claims that using qualitative techniques of introspection paves the way for understanding the 'vexed question of what cloze actually measures'.

**Competing interest**
The authors declare that they have no competing interests.

**Acknowledgments**
I would like to thank Professor Dan Douglas of Iowa State University and two anonymous reviewers who provided me with constructive comments which helped in revising the manuscript.

Received: 9 March 2013 Accepted: 3 May 2013
Published: 11 June 2013

**References**

Alderson, JC. (1979a). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly, 13*, 219–228.
Alderson, JC. (1979b). Scoring procedures for use on cloze tests. In CA Yorio, K Perkins, & J Schachter (Eds.), *On TESOL '79: The learner in focus* (pp. 193–205). Washington, D.C: TESOL.
Alderson, JC. (1983). Author's response. In JW Oller (Ed.), *Issues in language testing research* (pp. 218–223). Newbury House: Rowley, MA.
APA. (1985). *Publication manual of the American Psychological Association*. Washington, D.C: APA.
Babaii, E, & Ansary, H. (2001). The C-test: a valid operationalization of reduced redundancy principle? *System, 29*, 209–219.
Bachman, LF. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
Baggaley, AR. (1964). *Intermediate correlational methods*. New York: John Wiley.
Baker, D. (1989). *Language testing: A critical and practical guide* (1989th ed.). London: Edward Arnold.
Bensoussan, M, & Ramraz, R. (1984). Testing EFL reading comprehension using a multiple-choice rational cloze. *Modern Language Journal, 68*, 230–239.
Bormuth, JR. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading, 10*, 291–299.
Brown, JD. (1983). A closer look at cloze: validity and reliability. In JW Oller (Ed.), *Issues in language testing research* (pp. 237–250). Newbury House: Rowley, MA.
Brown, JD. (1988). *Understanding research in second language learning*. Cambridge: Cambridge University Press.
Brown, JD. (1994). A closer look at cloze validity. In JW Oller & J Jonz (Eds.), *Cloze and coherence* (pp. 189–196). London: Associated University Press.
Brown, JD, & Rodgers, T. (2002). *Doning second language research*. Oxford: Oxford University Press.
Carlisle, JF. (2007). Fostering morphological processing, vocabulary development and reading comprehension. In RK Wagner, AE Muse, & KR Tannenbaum (Eds.), *Vocabulary Acquisition: Implications for Reading Comprehension* (pp. 78–103). New York: The Guilford Press.
Carroll, JB, Carton, AS, & Wilds, C. (1959). *An investigation of cloze items in the measurement of achievement in foreign languages* (pp. 021–513). Cambridge: Cambridge, MA: Graduate School of Education, Harvard University, Laboratory for Research in Instruction.
Chapelle, CA, & Abraham, RG. (1990). Cloze method: What difference does it make? *Language Testing, 7*, 121–146.
Choi, IC, Sung Kim, K, & Boo, J. (2003). Comparability of a paper-based language test and a computer-based language test. *Language Testing, 20*, 295–320.
Connelly, M. (1997). Using C-tests in English with post-graduate students. *English for Specific Purposes, 16*, 139–150.
Cranney, AG. (1972–73). The construction of two types of reading tests for college students. *Journal of Reading Behaviour, 5*, 60–64.
Davies, A. (1990). *Principles of language testing*. London: Blackwell.
Farhady, H. (1983). The disjunctive fallacy between discrete point and integrative tests. In JW Oller (Ed.), *Issues in language testing research* (pp. 311–322). Rowley, MA: Newbury House Publishers.
Fotos, SS. (1991). The cloze test as an integrative measure of EFL proficiency: A substitute for essays on college entrance examinations? *Language Learning, 41*, 313–336.
Fulcher, G. (1999). Assessment in English for Academic Purposes: Putting content validity in its place. *Applied Linguistics, 20*, 221–236.
Garrett, HE, & Woodworth, RS. (1958). *Statistics in psychology and education* (5th ed.). London: Longman.

Gefen, P. (1979). An experiment with cloze testing. *English Language Teaching Journal, 33*, 122–126.
Glass, GV, & Hopkins, KD. (1984). *Statistical methods in education and psychology* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.
Glass, GV, & Stanley, JC. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice-Hall.
Goehring, HJ. (1981). *Statistical methods in education*. Arlington, Virginia: Information Resources Press.
Greene, BB. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading, 24*, 82–98.
Guilford, JP, & Fruchter, B. (1978). *Fundamental statistics in psychology and education* (6th ed.). Tokyo: McGraw-Hill Kogakusha.
Hale, GA, Stansfield, CW, Rock, DA, Hicks, MM, Butler, FA, & Oller, JW. (1989). The relation of multiple-choice cloze items to the Test of English as a Foreign Language. *Language Testing, 6*, 47–76.
Hanania, E, & Shikhani, M. (1986). Interrelationships among three tests of language proficiency: Standardized ESL, cloze and writing. *TESOL Quarterly, 20*, 97–110.
Hinofotis, FB. (1980). Cloze an alternative method of ESL placement and proficiency testing. In JW Oller & K Perkins (Eds.), *Research in language testing* (pp. 121–128). Newbury House: Rowley, MA.
Hopkins, KD, & Glass, GV. (1978). *Basic statistics for the behavioral sciences*. Englewood Cliffs, NJ: Prentice-Hall.
Ilyin, D, Spurling, S, & Seymour, S. (1987). Do learner variables affect cloze correlations? *System, 15*, 149–160.
Irvine, P, Atai, P, & Oller, JW. (1974). Cloze, dictation, and the test of English as a foreign language. *Language Learning, 24*, 245–252.
Jafarpur, A. (1995). Is C-testing superior to cloze? *Language Testing, 12*, 194–216.
Jafarpur, A. (1996). Native-speaker performance validity: In vain of for gain? *System, 24*, 83–95.
Johnson, P. (1981). Effects on reading comprehension of language complexity and cultural background of a text. *TESOL Quarterly, 15*, 169–182.
Jonz, J. (1976). Improving on the basic egg: The M-C cloze. *Language Learning, 26*, 255–265.
Klein-Braley, C. (1983). A cloze is a cloze is a question. In JW Oller (Ed.), *Issues in language testing research* (pp. 218–228). Newbury House: Rowley, MA.
Lado, R. (1986). Analysis of native-speaker performance on cloze tests. *Language Testing, 3*, 130–146.
Lee, YP. (1985). Investigating the validity of the cloze score. In YP Lee, ACYY Fok, R Lado, & G Low (Eds.), *New directions in language testing* (pp. 137–147). Oxford: Pergamon Press.
Lutji Spelberg, HC, De Boer, P, & Van den Boes, KP. (2000). Item type comparisons of language comprehension tests. *Language Testing, 17*, 311–322.
Mackey, A, & Gass, SM. (2005). *Second language research. Methodology and design*. Mahwah: Lawrence Erlbaum Associates, Inc.
MacLean, M. (1984). Using rational cloze for diagnostic testing in L1 and L2 reading. *TESL Canada Journal, 2*, 53–63.
Markham, PL. (1987). Rational deletion cloze processing strategies: ESL and native English. *System, 15*, 303–311.
Markham, PL. (1988). The cloze procedure and intersentential comprehension in college-level German. *IRAL, 26*, 44–51.
Mauranen, A. (1989). Can gaps measure comprehension? Modifications of cloze as tests of reading. In C Lauren & M Nordman (Eds.), *Special language: from humans thinking to thinking machines* (pp. 337–346). Clevedon: Multilingual Matters.
McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly: An International Journal, 3*, 31–51.
Messick, S. (1989a). Validity. In RL Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: Macmillan.
Messick, S. (1989b). Meaning and values in test validation: Science and ethics of assessment. *Educational Researcher, 18*, 5–11.
Messick, S. (1990). *Validity of test interpretation and use*. Princeton, N.J: Educational Testing Service. ETS-RR-90-11.
Messick, S. (1994a). *Alternative modes of assessment, uniform standards of validity*. Princeton, N.J: Educational Testing Service. ETS-RR-94-60.
Messick, S. (1994b). *Standards-based score interpretation: Establishing valid grounds for valid inferences*. Princeton, N.J: Educational Testing Service. ETS-RR-94-57.
Messick, S. (1994c). *Validity of psychological assessment: validation of inferences from person's responses and performances as scientific inquiry into score meaning*. Princeton, N.J: Educational Testing Service. ETS-RR-94-45.
Messick, S. (1996). *Validity and washback in language testing*. Princeton, N.J: Educational Testing Service. ETS-RR-96-17.
Oller, JW. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning, 23*, 105–118.
Oller, JW, & Conrad, CA. (1971). The cloze technique and ESL proficiency. *Language Learning, 21*, 183–195.
Oller, JW, Jr., & Jonz, J. (1994). *Cloze and coherence*. London: Associated University Press.
Porter, D. (1978). Cloze procedure and equivalence. *Language Learning, 28*, 333–341.
Rankin, EF. (1970). The cloze procedure – its validity and utility. In R Farr (Ed.), *Measurement and evaluation of reading* (pp. 237–253). New York: Harcourt, Brace and World.
Rankin, EF, & Culhane, JW. (1969). Comparable cloze and multiple-choice comprehension scores. *Journal of Reading, 13*, 193–198.
Richards, JC, Platt, J, & Platt, H. (1992). *Longman dictionary of language teaching and applied linguistics*. Harlow: Longman.
Roever, C. (2006). Validation of web-based test of ESL pragmalinguistics. *Language Testing, 23*, 229–256.
Sadeghi, K. (2002a). *The judgmental validity of cloze as a measure of reading comprehension*. Ankara, Turkey: METU. Paper presented at the 7th METU International ELT Convention,23–25 May.
Sadeghi, K. (2002b). *The criterion validity of cloze as a measure of EFL reading comprehension*. Exeter, UK: University of Exeter. Paper presented at BERA Research Student Symposium, 11–12 September.
Sadeghi, K. (2002c). *Is correlation a valid statistical tool in second language research?* Basel, Switzerland: University of Basel. Paper presented at 12th EUROSLA Conference, 18–21 September.
Sadeghi, K. (2002d). *Self-research: A new technique for validation in language testing*. Edinburgh, Scotland: University of Edinburgh. Paper presented at the Scotland TESOL 20th Annual Conference, 2 November.
Sadeghi, K. (2003). *An investigation of cloze procedure as a measure of EFL reading comprehension with reference to educational context in Iran*. University of East Anglia, Norwich, UK: Unpublished PhD Dissertation.

Sadeghi, K. (2004). Researcher research: An alternative in language testing research. *The Reading Matrix: An International Online Journal, 4*, 85–95.
Sadeghi, K. (2006). *Rethinking correlational validation*. Kermanshah, Iran: Razi University. Paper presented at 3rd TELLSI Conference, 3 February.
Sadeghi, K. (2010). Cloze validation against IELTS reading paper. *Journal of English Language Teaching and Learning, 53*, 130–153.
Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple-data source approach. *Language Testing, 17*, 85–114.
Sciarone, AG, & Schoorl, JJ. (1989). The cloze test: Or why small isn't always beautiful. *Language Leaning, 39*, 415–438.
Shohamy, E. (1983). Interrater and intrarater reliability of the oral interview and concurrent validity with cloze procedure in Hebrew. In JW Oller (Ed.), *Issues in language testing research* (pp. 229–236). Newbury House: Rowley, MA.
Stansfield, C, & Hansen, H. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly, 17*, 29–38.
Storey, P. (1997). Examining the test-taking process: a cognitive perspective on the discourse cloze test. *Language Testing, 14*, 214–231.
Stubbs, JB, & Tucker, GR. (1974). The cloze test as a measure of English proficiency. *Modern Language Journal, 58*, 239–241.
Taylor, WL. (1953). Cloze procedure – a new tool for measuring readability. *Journalism Quarterly, 30*, 415–433.
Taylor, WL. (1957). 'Cloze' readability scores as indices of individual differences in comprehension and aptitude. *Journal of Applied Psychology, 41*, 19–26.
Trites, L, & McGroatry, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing, 22*, 174–210.

doi:10.1186/2229-0443-3-15
**Cite this article as:** Sadeghi: **Doubts on the validity of correlation as a validation tool in second language testing research: the case of cloze testing.** *Language Testing in Asia* 2013 **3**:15.