Semantically Acceptable Scoring Procedures (SEMAC) Versus Exact Replacement Scoring Methods (ERS) For 'Cloze' Tests: A Case Study

David R. Litz & Allison K. Smith
UAE University

Bio Data

David Litz has completed a BA, B.Ed, and an M.A. in TESL/TEFL. He is presently working towards a doctorate in Educational Administration and Management. He has taught in South Korea, Canada and the UAE and his professional interests include testing, assessment and higher education administration.

Allison Smith has completed a BA, Bed, and an MA in TESL/TEFL. She is also working towards a doctorate in Educational Administration and Management and she has taught both math and ESL/EFL in South Korea, Canada, and the UAE. Her professional interests include testing, assessment, statistics and educational administration.

Abstract

It has been suggested by a number of theorists that a semantically acceptable scoring procedure (SEMAC) is preferable to an exact replacement scoring method (ERS) for 'Cloze' tests. They argue that the SEMAC procedure is fairer for testees and that it provides a better indication of communicative competency and the overall linguistic abilities of ESL/EFL students. While this may be the case, it has also been noted that SEMAC scoring requires considerable effort on the part of the scorers. Moreover, ERS and SEMAC scores have been shown, in some instances, to correlate highly when compared to one another. This demonstrates that it is possible, in some instances, for testees to rank similarly, whether one uses either the exact word or the SEMAC scoring procedure. This paper will report on an experiment that was conducted with a group of forty-five Korean EFL students in order to confirm the truth of the claim that SEMAC scoring and exact word scoring do, in fact, correlate very highly. The findings from this study show that these two scoring methods did correlate with high statistical significance. Thus, the decision of which method to use can be based on language learning objectives.

1.0 Introduction

1.1 Integrative Testing and Cloze Tests

Integrative testing involves the testing of language in context. It is concerned, therefore, with overall meaning and proficiency, the total communicative effect of discourse, and

the underlying linguistic competence of which it is argued that all learners possess (Oller, 1979). Proponents of integrative testing suggest that natural language processing and production requires making a highly complex series of decisions, which will inevitably involve knowledge of a number of important components such as grammatical structure, lexis, discourse structure and conventions, pronunciation and intonation. As such, they contend that tests should not seek to separate specific language skills into neat and ordered divisions, but should instead seek to assess the testee's ability to use two or more skills simultaneously (Cohen, 1980; Oller, 1979).

One of the most common types of integrative tests is the Cloze test. The principles behind Cloze testing are based on *Gestalt Psychology* and the *Information Processing Theory* of 'Closure' "which refers to the tendency of individuals to complete a pattern once they have grasped its overall significance" (Weir, 1998, p.46). Cloze tests are intended to measure the testee's ability to decode interrupted or mutilated messages by making the most acceptable substitutions from all the contextual clues available. Methods of deleting words on Cloze tests vary. Some researchers delete words randomly and others selectively, but Cloze tests have traditionally consisted of the regular or systematic deletion of words from a text (usually every 5 to 10 words) and their replacement by even-length blank lines. The testee's task is then to guess the words that have been deleted from the passage (Alderson, 1979a and 1979b; Brown, 1993; Oller 1972; Soudek and Soudek, 1983). Deciding upon the right type of Cloze test for a certain purpose is often difficult. Thankfully, Pikulski and Tobin (1982) have provided some useful guidelines for the construction and application of Cloze tests. They suggest that when teachers want to diagnose their students' ability to use various types of contextual clues, the words should be deleted on a rational basis rather than a mechanical basis. Alternatively, when teachers want to assess students' overall comprehension of a passage, the words should be deleted systematically.

Proponents of Cloze tests have claimed that they provide an excellent overall picture of proficiency since they reflect the degree to which language skills are used in a meaningful context, but a number of researchers have also found them to be particularly

useful tools for the measuring of reading comprehension. Brown (1978) as cited in (Brown, 1993), DeSanti (1989), Harris and Hodges (1995), Oller and Jonz (1994) and Sampson and Briggs (1983) have all suggested that the primary reason for this is the fact that the Cloze procedure assumes that reading is an interactive process and these tests are designed in such a way as to show whether the reader is familiar enough with the author's language and content to interact with the text in a way that preserves that author's meaning. In addition, Cloze tests measure the reader's ability to use contextual clues to derive meaning. Contextual clues can be ideas in the passage, words within the sentence or grammatical structures, and theorists have pointed out that the ability to use contextual clues in order to derive meaning is an important step in the development of overall reading comprehension (Pikulski and Tobin, 1982).

The Cloze procedure has several advantages over other types of reading assessments. For example, Cloze tests are very easily created and administered. They are also based on silent reading, which is the predominant and most natural form of reading. Moreover, they can be constructed from materials that teachers use for instructional purposes or from authentic texts and they do not require the writing of specific comprehension questions. Finally, and perhaps most importantly, Cloze tests often exhibit a high degree of consistency; though this consistency may vary considerably depending on the text selected, the deletion starting point and gap rates that are employed (Alderson, 1979b; Sciarone and Schoorl, 1989; Soudek and Soudek, 1983; Weir, 1998).

1.2 Scoring Procedures for Cloze Tests:

One method for scoring Cloze tests is known as exact replacement scoring or exact word scoring (ERS) in which only the words that have been deleted from the text are counted as being the correct responses for replacement. Heaton (1997) and Pikulski and Tobin (1982) suggest that this method is particularly quick and efficient and since there is no question as to whether a given answer is correct, scorer inter-reliability is guaranteed. They also suggest that most researchers currently use ERS to establish functional reading levels in order to determine the grade or level of difficulty of texts (readability) or to evaluate content materials.

Another type of scoring method for Cloze tests that has evolved is called semantically acceptable scoring (SEMAC). Methods for coding correct answers with SEMAC vary. Traditionally, only exact replacements as well as direct synonyms to the deleted words were counted as being correct responses - known as the 'synonymic' approach to scoring. More recent methods, on the other hand, tend to accept other answers that preserve the meaning of the deleted word even if the answers are not always direct synonyms. Essentially, this latter method compares the students' answers with the context to see if the supplied word makes sense in the sentence and the overall passage, thus preserving the author's intended meaning. Bormuth (1975) reports that it takes between two to four times as long to score tests when semantically acceptable answers are scored. Nevertheless, a number of other theorists have argued that the acceptance of all semantically acceptable answers for scoring purposes is a very effective means of diagnosing students strengths and weaknesses and assessing how much each student has learned over a set period of time (Oller, 1972; Pikulski and Tobin, 1982).

Both SEMAC and ERS have been found to be reliable scoring procedures in earlier studies (Brown, 1978; Oller, 1972) but a number of researchers have reported that SEMAC has a better dispersion of scores and may, in fact, be more reliable, a fairer scoring method, and the superior scoring procedure of the two. Alderson (1979a), Brown (1978), Culhane (1973) DeSanti (1989) and Oller (1972) have all demonstrated that ERS fails to accept creative meaningful responses or other words supplied by the students that are perfectly acceptable synonyms and they argue that it is unfair to treat all non-exact answers in the same manner. They also suggest that SEMAC probably provides a better measure of communicative proficiency as this procedure provides testees with more possibilities to demonstrate their understanding of the text, recognise and understand ideas and to grasp the significance or importance of ideas.

Although the arguments in support of SEMAC seem well founded, a number of researchers have shown that SEMAC may not actually be preferable to ERS and that it

does not really justify the additional work entailed in defining what constitutes an acceptable answer for each item (Henk, 1981; Ruddell, 1964).

> "[R]esearch seemed to indicate that in scoring Cloze tests, credit shouldbe allowed only for exact word replacements, because accepting synonyms did not improve the validity of the procedure but only increased the inefficiency and subjectivity in scoring." (Sampson and Briggs, 1983, p.177)

The most compelling evidence of the similarities between results produced by these methods stems from the fact that the correlation between SEMAC and ERS scoring procedures is actually very high when the results of tests are compared to one another (McKenna, 1976; Hadley & Naaykens, 1999). A high correlation suggests that both scoring systems may actually be providing language teachers with the same information, measuring the same qualities and placing students precisely in the same manner relative to each other whether a scorer uses the SEMAC method or the ERS method.

The goals of this study were to investigate the claim that SEMAC scoring and ERS scoring methods of Cloze passages do actually correlate very highly when compared with one another. In this study, SEMAC refers to a testee's response that indicates a total understanding of the author's intended meaning and provides a reconstruction of the sentence without altering the meaning, while ERS refers to exact replacements by the testees of the original words from the passage and credit is not awarded for any other types of responses. Typical correlation coefficient studies comparing the two scoring methods have used a rank correlation for the actual comparison of the scoring techniques (McKenna, 1976) or undertaken a correlational study between both SEMAC and ERS on the one hand, and the results of a concurrently validating proficiency test of the other (See Alderson, 1979a; Hadley & Naaykens, 1999; and Oller, 1972 for examples of this type of an experiment). As we were not interested in finding out which method compared more favourably with scores from a standardised test of ability, the tests in this study were simply scored using both methods and then the testees' raw scores as well as their final rankings were applied statistically to check for a significant and meaningful correlation between the two methods.

It is likely that a scoring correlational study of this nature will have important ramifications for language teaching pedagogy. If, for example, the correlation between

SEMAC and ERS scores is statistically significant and meaningful then this implies that the ranked scores from either method are comparable and that teachers should feel free to use the method that requires less effort - exact replacement scoring (ERS). Conversely, if they are found to correlate slightly or the results of the experiment are either statistically insignificant or meaningless then it would seem as though a scoring approach that gives students credit for exact answers as well as all semantically acceptable answers might be a truer indicator of a student's ability and represent a more comprehensive measurement procedure.

2.0 Methodology

2.1 Subjects

A group of 45 Korean EFL students who were enrolled in Sungkyunkwan University's (SKKU; a South Korean university located in the city of Suwon) EFL programme volunteered to participate in this study. Two semesters of English is a graduation requirement for all SKKU students and a standardised placement test is conducted before they begin their English studies. All of the subjects in this study were volunteers from a 'Level 1B Intermediate' class. It was agreed that it would be beneficial to use students of comparable ability to decrease the likelihood of varying proficiency levels having an undue influence on the outcome of the experiment. Prior to undertaking the study, all volunteers were also asked to read and sign a ‘permission statement’ in Korean and English. This permission statement was based on The University of Michigan’s Human Subjects Consent Form.

2.2 Materials and Design

A 25-gap summary Cloze test was constructed and administered for the purpose of this study. The students were familiar with this technique and had been exposed to it on a few occasions in the past. The passage was approximately 500 words in length. The decision to use a summary Cloze test was prompted by the belief that it would be easier to select pertinent test items from a relatively short passage. In light of recent negative findings on the mechanical deletion of Cloze tests, as described by Hughes (1996) and Weir (1998), it was also decided that it would be advantageous to employ a selective-deletion gap

filling procedure for this test. Selective-deletion gap filling on summary Cloze tests allows test constructors more freedom to select items for deletion based on what is known about the language in general, about the difficulty of the text and about the way language works in a particular text (Hughes, 1996; Weir, 1998). It is easier, therefore, to specifically state what each test is intended to measure and to determine "where deletions should be made and to focus on those items which have been selected *a priori* as being important to a particular target audience." (Weir, 1998 p.48)

The deletions on the test were all drawn from the students' Level 1B course curriculum. They were all generally nouns and verbs for it was felt that most of them would have a wide assortment of potential synonyms or other semantically appropriate choices, and it was assumed that many of the students would be relatively familiar with most of the original words selected by the author as well as a few possible synonyms that could be used in each instance.

The selected passage was pre-tested on a total of 100 students in several other Level 1B EFL classes that were not included in this study. It was also pre-tested on six faculty members in the Sungkyunkwan University English Department. After the initial range of potential acceptable responses was recorded, the deletions were checked again for additional acceptable collocates with the Collins Cobuild Concordance software. While specific test-item facility and discrimination (item analysis) were not undertaken for the pre-selected reading passage and cloze items, it was expected that the thorough pre-test process would assist in the development of a full range of acceptable responses that could be used for SEMAC scoring purposes. It was also hoped that the scope of the pre-test process would ensure that an appropriate and efficient measurement device had been produced for the particular purposes of this study. The final Cloze test that was administered to the experimental group is located in Appendix I. All of the exact word replacements as well as the semantically acceptable alternatives can be found in Appendix III.

2.3 The Reading Passage

The process of selecting a suitable reading passage for testing purposes can be difficult and complicated. In this case it required the careful screening of a variety of expository authentic text types that had been drawn from journals, books and newspapers. Passages that were potentially offensive were avoided and any that discussed modern political issues, religion, or sex were discarded. Moreover, passages that contained any other type of content-based issues that might have created emotional stress for some testees, and thereby increase their potential for poor test performance, were not used. It was felt that any score the candidates obtained should be a true measure of their respective ability, and as such, it was undesirable for a text topic to potentially influence their scores in any way.

Another important issue relating to the impact of text content on test scores that surfaced during the text selection process is the role of background knowledge or 'schemata' in the facilitation of textual comprehension. Johnson (1982), for example, has shown that in EFL reading comprehension, syntactic and lexical simplification is generally far less important than familiar content schemas. In addition, Carrell (1987) and Nunan (1985) have demonstrated that more than the provision of systemic knowledge, what makes a foreign language text easier to comprehend and process is actually the students' background knowledge and their degree of familiarity with the text's content and formal schemas. These findings, therefore, would seem to indicate that familiar schematic knowledge allows students to make more use of their knowledge of the world, logic and reasoning skills and inference competencies while processing textual information for comprehension. Unfortunately, if students in a testing situation are faced with a text for which they have considerable background knowledge of the topic then they will be likely to comprehend it more easily, make better inferences and answer the questions without completely understanding the overall meaning or the true contextual and linguistic nature of the text. Subsequently, topics that specifically dealt with South Korean or Asian issues were not used. Similarly, other passages that contained information or subject matter that might have favoured some candidates' background knowledge capacities over others were excluded and a great effort was made to select a passage that was equally unfamiliar or obscure to all candidates but also as general as

possible [Alderson and Urquhart, 1988; McEldowney, 1976 as cited in (Owen, Rees and Wisener, 1997)].

The text that was selected for this test was thought to be appropriate and stylistically acceptable for the testees. It was also believed to be of an adequate level of difficulty for the testees' ability. The text was adapted and summarised from an on-line educational journal designed for children to practice their reading skills. It was probably suitable for sixth to eighth grade (Middle School) native speakers, and it was felt that the reading level required to comprehend the chosen text would be comparable to the reading ability of the subjects of this study. The original text is a fairly straightforward article about the history of house cats and it is located in Appendix II.

2.4 Data Collection & Scoring Procedures

A protocol for the administration of all instruments was designed to insure that procedures would be standard throughout data collection. Data for the study was collected through the administration of the reading passage cloze test. All of the testees were volunteers from the English Program. They were told that they were participating in a study and that their scores on the test would have no effect on their overall course grade. All pre-selected volunteer students were present in the pre-determined classroom at the designated time and once everyone was seated they were provided with the test instructions. The student volunteers were told that they would be reading a short passage about cats and undertaking a cloze exercise (i.e. filling-in-the-blanks; an example of this test-type was displayed at the front of the classroom). They were asked to fill-in each blank in the passage with the best possible answer. They were not provided with several answer choices and they were not informed about the nature or purpose of the experiment.

Following the test, the answer sheets were collected and scored using both the SEMAC and ERS scoring methods and a statistical analysis was then performed to interpret the findings accurately. The two categories of score responses were

differentiated as follows: (a) by the restoration of original words/exact replacement scoring (ERS), and (b) by semantically acceptable fill-ins (SEMAC).

3.0 Results and Discussion

As mentioned previously, the Cloze test was scored using both SEMAC and ERS scoring methods and a statistical analysis was then performed to interpret the statistical significance and accurateness of the findings. The raw and ranked test scores obtained via the two scoring methods can be found in Appendix IV, Figures 1 and 2. *Pearson's Product Moment Correlation Coefficient* and *Spearman's Rank-Order Correlation Coefficient* were calculated and tested for significance and meaningfulness in the hopes of supporting the claim that SEMAC and ERS scoring methods correlate highly.

Frequency distribution histograms for both sets of test scores were manufactured first to observe if the scores were distributed normally. The graphs for both sets of test scores are relatively exemplary of *normal distributions* (See Appendix IV, Figures 4 and 5). The fact that the means, medians, modes and midpoints obtained from each set of data have fairly similar values is another indicator of normal distribution (For exact values see Appendix IV, Figure 6). The normal distribution of data is a crucial factor for many statistical measures and subsequently is one of three basic assumptions that underlie Pearson's Product Moment Correlation Coefficient - along with *independence* and *linearity* (Brown, 1996). Independence requires that each score pair is distinct from all other score pairs, and care was taken in inputting and charting the data to prevent any systematic association between pairs of scores during the statistical analysis. A *scatterplot,* which provides a representation of two sets of scores for visual comparison, was produced to determine if the relationship between the two sets of scores was in fact linear, and the resultant graph demonstrates suitable linearity (See Appendix IV, Figure 3).

The *mean* gives information about the central tendency of a set of scores, the *standard deviation* is a measure of dispersion that gives information on the variance of all scores in relation to the mean, and *standard error* places a single observed sample mean

in relation to its true population mean. All of these statistics are necessary for calculating a *confidence interval* which reveals whether or not scores from two different samples have actually come from two different populations (Bachman, 1997; Nunan, 1992). The means, standard deviations, and standard errors for the ERS and SEMAC sets of scores are 13.6/2.86/0.4312 and 16.3/3.54/0.5337 respectively. With a 95% level of confidence, the ERS confidence interval is 12.69-14.42 and the SEMAC confidence interval is 15.27-17.40, and thus these two samples are almost certainly drawn from two distinct populations. There is, however, a 5% chance that the true population mean from either sample will lie outside its established range. Another appropriate and relatively simple procedure that compares two sample means is the *t-test* (Nunan, 1992). A 'paired' t-test was conducted on the results since the samples contained identical subjects whom were measured under different treatment conditions (i.e. ERS vs. SEMAC) (Nunan, 1992). The t-test value (t = -16) was tested for significance at 44 degrees of freedom (df) and the critical value for 't' at this level indicates that the null hypothesis (that the results occurred due to chance alone) can be rejected absolutely. *Effect size* is yet another useful statistical tool that allows for the interpretation of the size of the difference between two means in terms of standard deviation units (Cohen, 1960). In this case, the effect size value indicates how much larger the SEMAC mean is than the ERS mean, and since the effect size value is 0.94, the SEMAC mean is nearly one entire standard deviation larger than the ERS mean. A relatively large difference between the sample means was to be expected. As the students had more opportunities to respond correctly with SEMAC scoring, the overall average of scores obtained from this method was presumed to be higher.

The final two statistical measures that were calculated - Pearson's Product Moment Correlation Coefficient (Pearson's r) and Spearman's Rank-Order Correlation Coefficient (Spearman's rho) - approximate the degree of association between two variables. Coefficient values may range from +1.0 (a perfect positive correlation) to -1.0 (a perfect negative correlation) (Brown, 1996; Nunan, 1992). In this study, the variables refer to the ERS and SEMAC test scores and the resultant coefficients provide information regarding the extent to which the two sets of scores vary in relation to each other. The ordinal-scale

data was inputted into the Spearman's rho formula to provide an estimate of similarity between the test scores' two sets of ranks. The interval-scale data was inputted into the Pearson's r formula to show if the two scoring methods were dispersing the students in relatively the same way. Both of these measures were then tested for significance to determine whether or not the results occurred purely by chance or by other factors. Finally, the Pearson's r coefficient was tested for meaningfulness. The *coefficient of determination* represents the proportion of overlapping variance between two sets of scores and thereby is an indicator of a correlation's meaningfulness. This coefficient can be calculated easily from Pearson's r, but unfortunately the process is rather arduous and complicated for Spearman's rho and since Spearman's rho should be interpreted carefully anyway - because it is a rather weak estimate of tendency - it was not computed. Moreover, the two correlation coefficients should be similar in any case since the Spearman's rho formula was developed to approximate the Pearon's r formula (Brown, 1996). The two correlation coefficients obtained from the data were indeed similar: Pearson's r = 0.9325 and Spearman's rho = 0.9553. Both correlation coefficients were also found to be statistically significant using a 'directional' decision at 99% certainty ($p<0.01$). The coefficient of determination for Pearson's r was equal to 0.87; this indicates a fairly high percentage of overlap between the two sets of scores. The culmination of this statistical analysis, along with evidence previously propounded, demonstrates that there is a very high correlation between the SEMAC and ERS scoring methods.

4.0 Pedagogic Implications

If it is merely a question of wanting to know which scoring system is more reliable for Cloze testing, this study lends credence to the view that ERS and SEMAC are similarly effective scoring methods as they correlate very highly with one another. As a diagnostic tool, ERS is certainly quicker and easier to design and score than SEMAC. It is important to consider, however, that when using ERS deleted items are often chosen in terms of their limited or minimal number of acceptable alternatives. As such, ERS scoring methods may only assess a partial area of language proficiency and simply test learners' knowledge of "key words" such as those found in textbook glossaries. If the goal is to assess a broader view of a learners' L2 proficiency or insight into learners' progress

through stages of a course, complementary and alternative forms of testing such as SEMAC should be considered. Other recent alternatives such as Clozentrophy may also be examined. This is a procedure in which a cloze test is first given to a group of native speakers, and their responses are listed in frequency order. Afterwards, the test is given to non-native speakers and a testee that responds with a high-frequency word would score higher than one who responds with a low frequency word (Richards, Platt and Platt, 1992). Similarly, the role of language corpuses (i.e. Collins Cobuild) in the design of SEMAC and other forms of Cloze testing might be expanded. Higher scores, for example, could be awarded for more commonly used collocates of relevant words in the immediate context of the deletion. Consequently, score results might show the degree of proximity of a learner's proficiency to native speaker linguistic competence.

Additional factors may also influence a teacher's decision of which particular scoring system to use. Hadley & Naaykens (1999) suggest that teachers may need to consider how the advantages and disadvantages of the ERS and SEMAC systems might compliment or clash with the values of their students' culture of learning. They point out that Japanese students prefer the ERS system because it gives the testee the impression that there is only **one** true correct answer whereby SEMAC scoring methods are distrusted by students who tend to resist the possibility that there can be communicative variation in test answers. Alternatively, teachers in other countries or regions may encounter different responses to tests based on their respective students' culture, situation, and individual learning styles and strategies. As such, there will not be any easy answers to these issues and researchers and teachers must make informed decisions based on careful considerations of the teaching and learning context as well as our students' individual needs as language learners.

5.0 Implications for Further Research

This particular study attempted to corroborate earlier correlational studies on ERS and SEMAC scoring methods / systems with a specific type of EFL learner – the Korean university student. It should be acknowledged, however, that it also brought forth some entirely new questions and possible avenues towards additional research.

First, a pre-test was carried out but an item analysis was not conducted on the test instrument, and as such, the selection of deletions and possible alternatives may have had an impact on the results of the study. More rigorous item facility and discrimination analyses in future studies might provide more reliability and validity to similar experiments or even establish divergent or conflicting results. In addition, the inclusion of information from a language corpus might have also strengthened the test instrument in this particular study and could be included in the design of test instruments for use in research that is designed to identify stages of a student's L2 development on the basis of calculated statistical information about collocation with respect to testees' cloze test answers. Last, the goal of the study was to provide an overall comparison of the test scores and as such comparisons and illustrations of responses to specific test items that demonstrated the possible compatibility between SEMAC and ERS were not undertaken in this study. This would have undoubtedly supported to the correlational results of the study and similar research could address this shortcoming in the future.

Second, a correlation coefficient is not necessarily the best indicator of similarity between two scoring methods such as ERS and SEMAC. Appropriate statistical measures depend upon the types of decisions being made about the testees. For example, the Pearson's r and Spearman's rho correlation coefficients are used to make 'relative' as opposed to 'absolute' decisions about testee ability. These measures, therefore, describe a testee's relative position within a group and the evaluation of one's standing is made in relation to the performance of other testees in the comparison group. Relative testing decisions might involve adjusting program levels to suit students' abilities or making comparisons between various programs. In contrast, test scores may be interpreted absolutely; giving an indication of whether each testee's score falls above or below some pre-specified criterion. Testee performance in this instance is assessed on an individual basis - regardless of the performance of others. Absolute decisions might pertain to the testees' eligibility for admission into a program (Shavelson and Webb, 1991). If Cloze tests were used solely for making relative decisions, then the use of a correlation coefficient would be an acceptable means of summarising the similarity between scores

produced by ERS and SEMAC scoring procedures. But assuming that Cloze tests are also sometimes used to make absolute decisions, other studies must be conducted to resolve whether ERS and SEMAC scoring methods are equally applicable for making these types of decisions. The *agreement coefficient* would serve this purpose since it estimates the consistency of decisions that classify subjects as either 'masters' or 'non-masters' of an ability or skill and as such would describe the degree of classification agreement between ERS and SEMAC scoring procedures (Brown, 1996).

Third, this particular study presented findings from a particular homogeneous Korean context. Further research needs to address similar issues in different EFL/ESL contexts and the degree to which culture has an impact on Cloze tests. These studies might examine the relationship between scoring methods or systems and specific learner needs, behaviours, strategies, expectations or even students' particular cultures of learning. Likewise, data for this study was conducted under experimental conditions in a simulated testing situation. Alternative studies could examine Cloze scoring methods in actual authentic large scale and/or high-stakes testing situations such as standardized English language proficiency exams.

6.0 Conclusion

This experiment was conducted in order to investigate the claim that SEMAC and ERS scoring procedures correlate very highly. The Pearson Product-Moment Correlation Coefficient statistical formula was used in this study to compare the raw scores that were obtained from each scoring method. The final calculation revealed that there is an extremely significant statistical correlation between the raw scores obtained by both scoring methods. In addition, the Spearman Rank-Order Correlation Coefficient statistical formula was used in order to compare the rankings or standing of the testees that were acquired via the two scoring methods. This measurement demonstrated that there is a very high statistically significant correlation between the final rankings obtained by each scoring method.

Admittedly, an item analysis was not conducted prior to the administration of the reading passage and this could have possibly had an impact on the results. Nevertheless, the statistical analyses that were undertaken on the test items in this study did, in fact, corroborate similar studies and demonstrated that SEMAC and ERS scoring procedures do, in some circumstances, correlate very highly. While we cannot assume that both scoring methods always provide the same type of information about our students or the same measure of our students' relative ability, the fact that both types of scoring methods typically correlate highly would suggest that teachers might feel free to choose ERS over SEMAC in quick diagnostic situations where a relative decision is being made from the test results, as it requires less effort, it is more time efficient and it probably gives a reliable picture of each student's comparative ability within a limited range. Caution should be exercised, however, as future research must still be conducted to examine if the same conclusion holds true in different or alternate EFL/ESL contexts, for absolute or authentic, high-stakes testing scenarios or for situations where a 'broader' view of a learners' L2 proficiency is required.

References

Alderson, J.C. (1979a). The Cloze procedure and proficiency in English as a Foreign Language. *TESOL Quarterly*, *13*(2), 219-28.

Alderson, J.C. (1979b). The effect on the Cloze test of changes in deletion frequency. *Journal of Research in Reading*, *2*(2), 108-19.

Alderson, J.C. and Urquhart, A.H. (1988). This test is unfair: I'm not an economist. In P.L. Carrell, J. Devine, and D. E. Eskey (Eds.), *Interactive approaches to second language reading*. (pp. 168-182). Cambridge: Cambridge University Press.

Allen-Figural, J. (Ed.). (1964). *Improvement of reading through classroom practice.* Newark: International Reading Association.

Bachman, L. (1997). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bormuth, J.R. (1975). The Cloze procedure: Literacy in the classroom. In W.D. Page, (Ed.), *Help for the reading teacher: New directions in research*. National Conference on Research in English (Illinois, USA).

Brown, J.D. (1978). *Correlational study of four methods for scoring Cloze test*s. Unpublished master's thesis, University of California, USA.

Brown, J.D. (1993). What are the characteristics of natural Cloze? *Language Testing, 10*(2), 93-116.

Brown, J.D. (1996). *Testing in language programs*. Upper Saddle River, NJ: Prentice Hall Regents.

Carrell, P.L. (1987). Content and formal schemata in ESL reading. *TESOL Quarterly, 21*(3), 461-81.

Carrell, P.L., Devine, J, and Eskey, D.E. (Eds.). (1988). *Interactive approaches to second language reading*. Cambridge: Cambridge University Press.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, A.D. (1980). *Testing language ability in the classroom*. Rowley, MA: Newbury House.

Culhane, J.W. (1973). *T*he use of an iterative research process to study the adaptation of Cloze for improving the reading comprehension of expository materials. In *Dissertation Abstracts International, 34*, 997A.

DeSanti, R.J. (1989). Concurrent and predictive validity of a semantically and syntactically sensitive Cloze scoring system. *Reading Research and Instruction, 28*(2), 29-40.

Hadley, G. and Naaykens, J. (1999). Testing the test: Comparing SEMAC & exact word scoring on the selective deletion test. *Korea TESOL Journal, 2*, 63-72.

Harris, T.L. and Hodges, R. E. (Eds.). (1995). *The literacy dictionary: The vocabulary of reading and writing*. Newark: The International Reading Association.

Heaton, J.B. (1997). *Writing English language tests*. New York: Longman.

Henk, W.A. (1984). Effects of modified deletion strategies and scoring procedures on Cloze test performance. *Journal of Reading Behavior, 13*, 347-356.

Hughes, A. (1996). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Johnson, P. (1982). Effects of reading comprehension of building background knowledge. *TESOL Quarterly, 16*, 503-516.

McEldowney, P. (1976). Test in English (Overseas), The position after ten years. Occasional Paper 36, Manchester: Joint Matriculation Board.

McKenna, M.C. (1976). Synonymic Versus Verbatim Scoring of the Cloze Procedure. *Journal of Reading*, *20*, 141-143.

Nunan, D. 1985. Content familiarity and the perception of textual relationships in second language reading. *RELC Journal*, *16*(1), 43-51.

Nunan, D. (1992). *Research methods in language learning*. Cambridge: Cambridge University Press.

Oller, J.W. (1972). Scoring methods and difficulty levels for Cloze tests of proficiency in English as a second language. *Modern Language Journal*, *56*(3), 151-157.

Oller, J.W. (1979). *Language tests at school*. New York: Longman.

Oller, J.W. and Jonz, J. (Eds.). (1994). *Cloze and coherence*. Cranbury, NJ: Bucknell University Press.

Owen, C., Rees, J. and Wisener, S. (1997). *Testing*. The University of Birmingham: School of English Centre for English Language Studies.

Page, W.D. (Ed.) (1975). *Help for the reading teacher: New directions in research*. National Conference on Research in English (Illinois, USA).

Pikulski, J.J. and Shanahan, T. (Eds.). (1982). *Approaches to the informal evaluation of reading.* Newark: International Reading Association.

Pikulski, J.J. and Tobin, A.W. (1982). The Cloze procedure as an informal assessment technique. In J.J. Pikulski and T. Shanahan (Eds.). *Approaches to the informal evaluation of reading.* Newark: International Reading Association.

Richards, J. Platt, J. and Platt, H. (1992). *Longman Dictionary of Applied Linguistics*. (2nd ed.). London: Longman.

Ruddell, R.B. (1964). A study of the Cloze comprehension technique in relation to structurally controlled reading material. In J. Allen-Figural (Ed.). *Improvement of reading through classroom practice*. (pp. 298-303). Newark: International Reading Association.

Sampson, M.R. and Briggs, L.D. (1983). A new technique for Cloze scoring: A semantically consistent method. *Clearing House*, *57*(4), 177-179.

Sciarone, A.G. and Schoorl, J.J. (1989). The Cloze test: Or why small isn't always beautiful. *Language Learning*, *39*(3), 415-38.

Shavelson, R.J. and Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, NJ: Sage Publications.

Soudek, M. and Soudek, L. (1983). Cloze after thirty years: New uses in language teaching. *ELT Journal*, *37*(4), 335-40.

Weir, C. (1998). *Communicative language testing*. Hemel Hempstead: Prentice Hall.

Appendix I

MODIFIED 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 around 3000 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 storehouses.

The Greeks were probably the first _____7_____ to recognise 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 DaVinci summed it up best when he referred to the cat as 'Nature's Masterpiece'.

ANSWERS:

1. _____ **11.** _____ **21.** _____

2. _____ **12.** _____ **22.** _____

3. _____ **13.** _____ **23.** _____

4. _____ **14.** _____ **24.** _____

5. _____ **15.** _____ **25.** _____

6. _____ **16.** _____

7. _____ **17.** _____

8. _____ **18.** _____

9. _____ **19.** _____

10. _____ **20.** _____

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 II

THE FELINE: From Goddess to Pet

The cat has a history as fascinating and mysterious as the creature itself. The true beginnings of the domestic cat are unknown, but the cat may have first appeared around 3000 B.C. in a country called Nubia, which bordered Egypt. Egypt later conquered Nubie, and by 2500 B.C., the cat was domesticated in Egypt. The cat's first name in Egypt was Myeo or Mau. The mau's status in Egypt grew rapidly; she was eventually considered guardian of the temple and was worshipped as a goddess. Ancient Egyptians believed the cat was the daughter of Isis and Goddess of the Sun and the Moon. They also believed that the glow from a cat's eyes held captive the light of the sun. The goddess Bast, who controlled the life giving sun's heat, had the head of a cat. Besides being worshipped as goddesses, cats also had a practical function: they kept mice from overrunning the Egyptian grain storehouses. Ancient Egyptians carved wooden figures of cats and crafted furniture and jewellery in the shops of cats. Most of the larger art museums today have at least one life-size sculpture of a cat from ancient Egypt.

Killing a cat in ancient Egypt was a crime punishable by death. Cats were mummified when they died and saucers of milk along with mummified rats and mice were placed in the cat's tombs. Cat cemeteries existed along the Nile River and cat mummies have also been found in the tombs of ancient Egyptians.

Although law prevented exporting cats from Egypt, cats were frequently stolen by Phonecian traders and taken to Mediterranean countries. Egyptian soldiers were encouraged to take any cats they saw during their foreign travels and bring them back to Egypt, their true home.

When Persia attacked the Egyptian city of Pelusium, the Persian king was aware of the Egyptians' devotion and loyalty to the cat and he devised a plan: he ordered his soldiers to search the city and take any cat they found. During the next attack, his soldiers each held up a live cat as they came near the Egyptian borders. Rather than harm the cats, the Egyptians surrendered their city to Persia.

The Greeks were probably the first Europeans to recognise cats for their mouse-catching talents. When Egyptians refused to sell or trade any of their cats, the Greeks stole several of the Egyptian cats and sold the kittens of these stolen cats to Roman, Gauls, and Celts. The cat became the symbol of liberty in ancient Rome. Roman artists often showed Libertus, the Goddess of Liberty, with a cat lying at her feet.

Cats were domesticated in India and the Far East later than in Egypt, around 2000 B.C. People in India and China praised cats for their ability to catch rats and mice. Ancient Chinese believed that cats brought good fortune and that the glow from a cat's eyes frightened away evil spirits.

In Burma and Siam it was believed that the souls of departed people lived in the bodies of sacred cats before moving on to the next life. Cats lived in the temples and palaces of Siam) which is now Thailand). In Japan religious ceremonies were sometimes held for the souls of departed cats. Cats in Japan were required to be kept on leashes until 1602; a low was then passed to set the cats free to kill the rats, which were hurting the silkworms.

By the end of the eleventh century cats were popular among sailors because of their rat-catching skills. Sailors admired cats because they destroyed disease-infested rats, which lived on ships. Many sailors believed that cats possessed special powers that could protect them at sea.

Although the cat was held in high regard and fancied during ancient times, the cat didn't fare will in Europe during the Middle Ages. Leaders of the Christian church began a campaign against cats. They were slaughtered in masses in just about all of Europe, which led to the near extinction of cats in Europe by 1400. Cats were associated with evil, witchcraft, voodoo, and black magic. Many people believed that witches regularly transformed themselves into cats. Men and women were tortured or even killed for helping a sick 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.

Why were cats persecuted during Europe's Middle Ages? Possibly because they travelled at night, and often bonded with elderly women living alone who were frequently victims of witchcraft accusations by suspicious neighbours. The aloofness and independent nature of cats may have added to the belief that they were evil. This terrible period of persecution of cats lasted about 400 years, but never spread to India or the Middle and Far East. By the eighteenth century, cats were once again looked upon favourably. In 1835, a low was passed in England forbidding the mistreatment of any animals.

Traders from the Orient, explorers, and colonists brought cats to the Americas in the 1700's. Cats even travelled to North America with the Pilgrims on the Mayflower. Most of these immigrant cats are ancestors of cats living here today.

Cats have lived with kings as well as presidents. White cats were popular among royalty, like Louis XV of France. Japanese and Chinese emperors owned white cats. Cats lived with Abraham Lincoln, Theodore Roosevelt, Calvin Coolidge, Herbert Hoover, Ronald Reagan, and Bill Clinton. Napoleon may have hated cats but his conqueror, the Duke of Wellington, was a cat lover, as was Queen Victoria. France's Cardinal Richelieu provided for the future care of his fourteen cats in his will. In India today the Hindu religion urges followers to provide food and shelter for at least one cat.

Some legends and superstitions 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 on every black cat there is a single hair that is white. If you remove it without the cat scratching, this white hair will bring you wealth of luck in love. One superstition states that a black cat crossing one's path brings bad luck, but in Great Britain black cats are thought to bring good luck.

Today the elegant, graceful cat has become a popular house pet throughout the world. The cat is one of the smartest of tame animals. They are independent and hard to train. Cats are valued for their gentle, affectionate natures. They have good memories; they remember who treats them well and who treats them badly. A cat's loyalty is earned; a cat won't stay where it is mistreated. 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 see in near darkness. Not only are the cat's eyes beautiful but their eyes are the largest in proportion to body size when compared to other animals. Cats are clean and have been praised for their mysterious, exotic looks. Perhaps Leonardo DaVinci summed it up best when he referred to the cat as "Nature's Masterpiece".

SOURCE: Coll, J. 2000. 'The Feline: From Goddess to Pet'. http://www.indiana.edu/~eric_rec/fl/pcto/feline.htm.

Appendix III

SEMANTIC KEY:

1. history, past, background, story, background, narrative, account
2. country, place, nation, locale, land, area, region, territory, realm
3. name, title, kind, breed, species, label
4. status, reputation, popularity, standing, fame, influence, importance, demand, population, numbers, stature, significance, prestige, prominence, eminence
5. function, use, purpose, role, ability, talent, skill, aspect, duty, mission, job, responsibility
6. mice, rats, rodents, vermin, pests, insects grasshoppers, bugs, birds
7. Europeans, Westerners, Whites
8. stole, took, captured, thieved
9. kittens, offspring, majority, litter, descendants, young, brood
10. symbol, sign, representative, representation, figure, image
11. century
12. destroyed, killed, ate, caught, captured, hunted, chased, eliminated, exterminated, ravaged
13. protect, preserve, guard, safeguard, shield
14. ancient, old, early, archaic
15. magic, sorcery, wizardry, occultism, devilry
16. witches, sorceresses, spirits, ghosts, wizards, hags, enchantresses
17. sick, dying, maimed, hurt, ill, wounded, cut, pregnant, lost, infected, hungry, starving, unwell, diseased, ailing, infirm
18. superstitions, myths, stories, legends, (folk) tales, proverbs, rumours, fears, fables
19. luck, karma, fate, happenings, events, fortune, fortuity, happenstance
20. pet, animal
21. world, globe, earth, planet
22. good, excellent, great, terrific, okay, large, special, accurate, outstanding, long, photographic, keen, sharp, spectacular, proficient, commendable, accomplished, efficient, competent, capable
23. remember, know, recognise , retain, recall
24. mistreated, unhappy, abused, disliked, unwanted, unwelcome, ignored, hated, unsafe, uncomfortable, maltreated, ill-treated, harmed
25. see

EXACT-WORD 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 IV

Fig. 1 - RAW CLOZE-TEST SCORES:

A = Testee Number B = ERS Score C = SEMAC Score

A	B	C
1.	8	9
2.	8	10
3.	9	10
4.	10	11
5.	9	11
6.	11	12
7.	10	12
8.	10	13
9.	11	13
10.	12	13
11.	13	14
12.	11	14
13.	11	14
14.	12	14
15.	13	15
16.	12	15
17.	13	15
18.	11	15
19.	14	16
20.	14	16
21.	14	16
22.	12	16
23.	14	17
24.	14	17
25.	13	17
26.	15	17
27.	15	17
28.	14	18
29.	15	18
30.	13	18
31.	15	18
32.	14	18
33.	16	19
34.	17	19
35.	16	19
36.	16	19
37.	16	20
38.	17	20
39.	15	20
40.	17	21
41.	16	21
42.	18	21
43.	18	22
44.	19	22
45.	19	23

Fig. 2 - RANK-ORDER OF CLOZE TEST SCORES:

A = Testee Number B = ERS Rank C = SEMAC Rank

A	B	C
1.	1.5	1
2.	1.5	2.5
3.	3.5	2.5
4.	6	4.5
5.	3.5	4.5
6.	10	6.5
7.	6	6.5
8.	6	9
9.	10	9
10.	14.5	9
11.	19	12.5
12.	10	12.5
13.	10	12.5
14.	14.5	12.5
15.	19	16.5
16.	14.5	16.5
17.	19	16.5
18.	10	16.5
19.	25	20.5
20.	25	20.5
21.	25	20.5
22.	14.5	20.5
23.	25	25
24.	25	25
25.	19	25
26.	31	25
27.	31	25
28.	25	30
29.	31	30
30.	19	30
31.	31	30
32.	25	30
33.	36	34.5
34.	40	34.5
35.	36	34.5
36.	36	34.5
37.	36	38
38.	40	38
39.	31	38
40.	40	41
41.	36	41
42.	42.5	41
43.	42.5	41
44.	44.5	43.5
45.	44.5	45

Fig. 3 - SCATTERPLOT OF ERS VS. SEMAC SCORES:

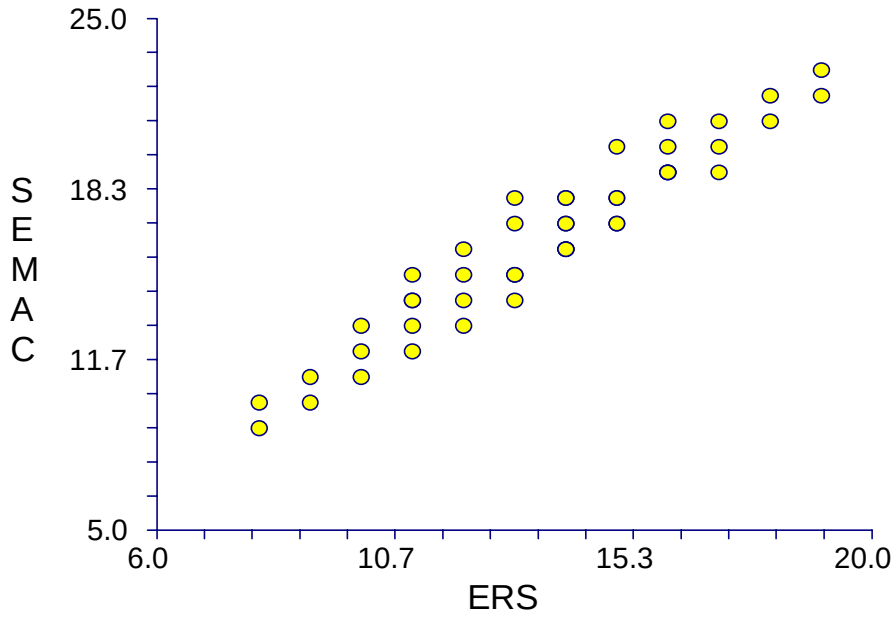

Fig. 4 - FREQUENCY DISTRIBUTION OF ERS SCORES:

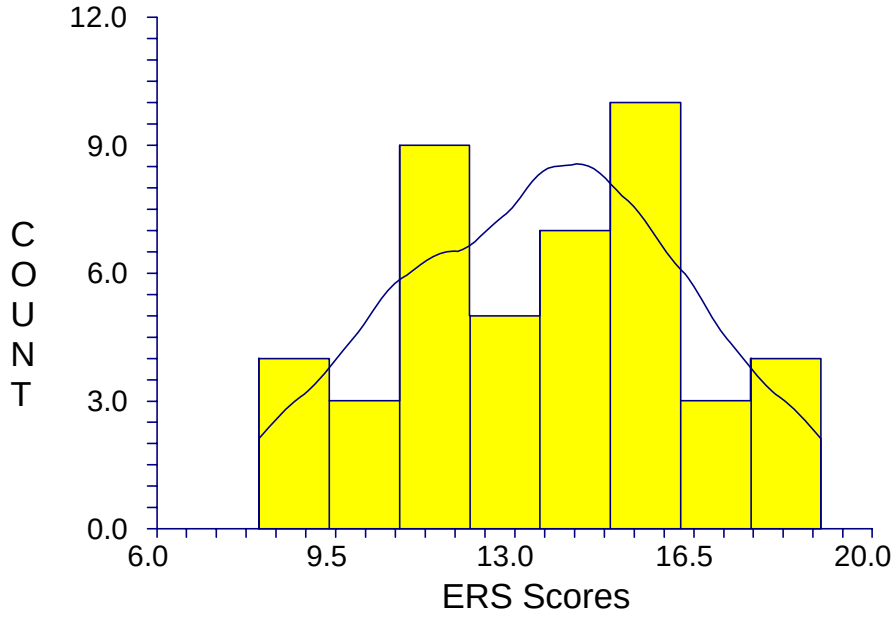

Fig. 5 - FREQUENCY DISTRIBUTION OF SEMAC SCORES:

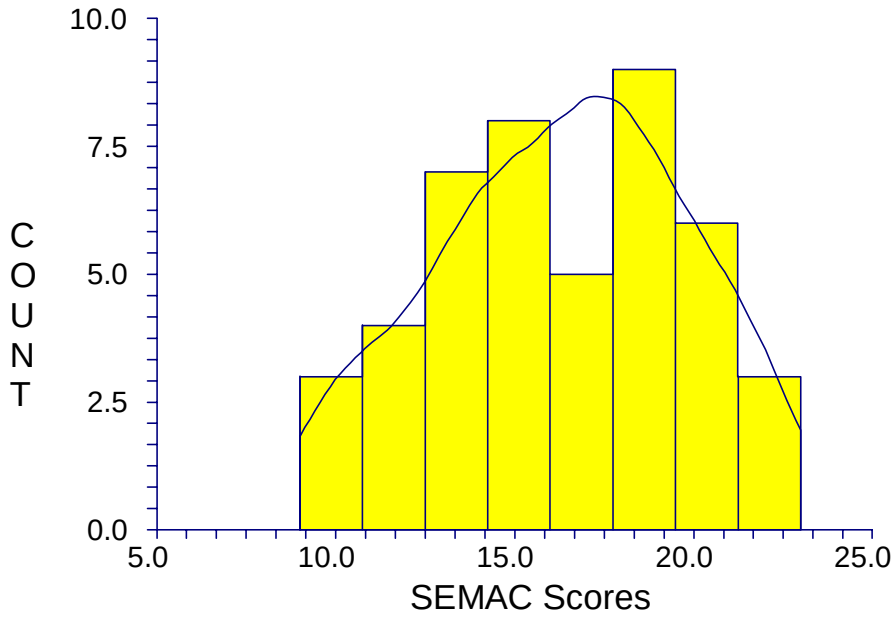

Fig. 6 - STATISTICAL ANALYSIS:

ERS Values	*SEMAC Values*
Mean: 13.6 (or 54.2%)	***Mean:*** 16.3 (65.2%)
Median: 14	***Median:*** 17
Mode: 14	***Mode:*** 17/18
Midpoint: 12.5	***Midpoint:*** 12.5
Range: 11	***Range:*** 14
Variance: 8.18	***Variance:*** 12.53
Standard Deviation: 2.86	***Standard Deviation:*** 3.54
95% Confidence Interval: 12.69 - 14.42	***95% Confidence Interval:*** 15.27 - 17.40
Standard Error: 0.4312 (z = 6.26)	***Standard Error:*** 0.5337 (z = 5.06)

Effect Size ***(Semac Mean - ERS Mean) ÷ ERS s.d.***
e = 0.94

t-test (paired):
t = -16.6
df = 44
* Probability of the result assuming the null hypothesis is zero

STATISTICAL ANALYSIS (CONT.)

Pearson's Product-Moment Correlation Coefficient:
r_{xy} = 0.9325, (p<.01)
t = 17
Coefficient of Determination: 0.87

Spearman's Rank-Order Correlation Coefficient:
r_s = 0.9553, (p<.01)
t = 21.198
df = 43 (significant at 0.05 level - directional test)