GETTING CLOSURE ON CLOZE:
A VALIDATION STUDY OF THE "RATIONAL DELETION" METHOD

A SCHOLARLY PAPER SUBMITTED TO THE DEPARTMENT OF
SECOND LANGUAGE STUDIES OF THE UNIVERSITY OF HAWAI'I
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF ARTS

IN

SECOND LANGUAGE STUDIES

November 23, 2005

by

Treela McKamey

Advisor:

Dr. James Dean Brown

ABSTRACT

The present study was designed to (a) test the notion that the cloze procedure is related to the idea of closure in Gestalt psychology and (b) investigate the degree to which specific language skills (grammar knowledge, reading ability, and vocabulary knowledge) contribute to cloze test performance for second language learners and (c) investigate the degree to which cloze tests may require nonverbal reasoning skills. Forty-eight second language students from an Intensive English Language program sat for a battery of nine tests: a cloze test, a reading test, a vocabulary test, a grammar test, two tests of Closure 1, two tests of Closure 2, and the Raven's Progressive Matrices (Plus version). Test scores were submitted to a factor analysis and multiple regression analysis. The factor analysis showed two clear factors, one verbal and one nonverbal. The multiple regression analysis indicated that grammar and reading are good indicators of cloze test performance but nonverbal abilities such as closure and higher order reasoning are not. Implications for teaching and research are discussed.

ACKNOWLEDGEMENTS

This paper would not have been possible without the encouragement and support of my friends, peers and professors. I would especially like to acknowledge the support and advice that I received from my good friend Yuki Watanabe throughout the duration of this project. I would also like to express my gratitude to Dr. Richard Day for first encouraging me to pursue this idea, Dr. Chaudron for supplying me with input and resources in the early stages of the project, and Dr. Schmidt for taking time in his summer months to read an earlier draft of this paper. I am grateful, as well, to the ELI faculty and staff who cooperated with me on this project; in particular, thanks to Kenton Harsch, Priscilla Faucette and Yao Zhang for answering questions and making the ELI accessible. Finally, I'd like to thank my advisor, Dr. James Dean Brown, for his guidance, advise, feedback, suggestions, interest and encouragement throughout the entire process, from beginning to end. Any errors that remain are my own.

INTRODUCTION

In 1953, Wilson L. Taylor introduced the "cloze procedure" as a method of measuring the "readability," or difficulty, of a text, and later (1956) as a measure of reading comprehension for native speakers. The method was simple—systematically or randomly delete words from a particular passage and ask the student to restore the missing words—and so, it quickly gained in popularity and its momentum has not stopped. The popularity of the cloze procedure is evident in the literature even today. Since the millennium alone, there has been countless research, in L1 and L2, involving cloze procedures[i], including the use of cloze to measure specific constructs, to pilot it for new purposes, or specifically to question or improve its validity in testing. The popularity and wide-spread use of the cloze procedure is perhaps ironic considering the wide-spread debate over its use, especially in second language testing. As a test, researchers and language testers disagree on exactly what skills are brought to bear on filling a blank in a cloze test and, thus, what constructs exactly can be measured by cloze tests[ii]. For example, some researchers believe that traditional cloze tests—created by deleting every n^{th} word—are not capable of measuring "macro-level" or "higher-order" reading skills, but measure mostly only sentence level knowledge (for examples, Alderson, 1978, 2000; Bachman, 1982, 1985; Kibby, 1980; and Shanahan, Kamil & Tobin, 1982). Others have maintained that traditional cloze tests are sensitive to both sentence level knowledge and intersentential knowledge (for examples, Brown, 1983a; Chihara, Oller, Weaver, & Chavez-Oller, 1977; Cziko, 1983; Henk, 1982; Jonz,

[i] The Academic Search Premier database, for example, contains 54 studies since the year 2000 that use the cloze procedure in one way or another. This number is probably an underestimate since other databases exist and not all research can be accounted for in such databases. In all, there are 1,644 studies cited in EBSCOhost that address the cloze procedure.

[ii] The terms "cloze procedure" and "cloze test" are closely related, the difference being that cloze procedure is more general, demarking the use of an activity that follows "cloze procedures", whereas, a cloze test is a specific application of the cloze procedure to the testing situation. The two terms are used interchangeably in this paper.

1990; McKenna and Layton, 1990; and Oller, 1983). Also, a few researchers (Brown, 1983b); Hinofotis, 1980a; Oller, 1972; Oller & Conrad, 1971) apply cloze tests to the construct *general language proficiency* rather than to *reading proficiency* alone, which further confuses its use. More recently, Brown (2002 and earlier in 1983b) has pointed out that each cloze test, in fact, each cloze item, may function very differently for different language groups, requiring different skills for individual test-takers, depending on their proficiency level and other individual differences. Finally, the origin of the cloze procedure, discussed below, suggests that at least one of the skills required to "cloze" the gaps created by deleted words is not a language skill at all, but rather a kind of non-verbal reasoning skill, known in Gestalt psychology as "closure". If it is demonstrated empirically that cloze tests measure abilities that are language independent, even if partially so, then there would be further reason to exercise caution when applying cloze procedures to language assessment.

Despite the disagreements and ambiguity surrounding cloze tests, language professionals maintain their interest in cloze and are apparently unwilling to abandon it as a useful tool in second language teaching and assessment, citing its ease of construction, administration, and scoring as the primary reason (e.g., Bormuth, 1967; Brown, 1980, 2002; Gamaroff, 1998; Hinofotis, 1980a; Rankin & Culhane, 1969) . But we must exercise caution here. Responsible professionals need to garner greater certainty as to what cloze tests actually measure before they can take advantage of the benefits they may offer. It is probably not sound practice to continue "letting the cards fall where they will" (Brown, 2002, p. 110) as traditional cloze tests seem to do, especially when high-stakes, or even moderate, decisions may be made based on cloze test scores.

The present study was undertaken to address the knowledge gaps that exist concerning what cloze tests measure, thus "bringing closure to cloze". Specifically, this paper will begin

with a review of relevant research investigating the validity of cloze tests as measures of second language learner reading proficiency. This will be followed by an examination of the theoretical precursor to cloze tests, the idea know as "closure". And, finally, this paper will present an empirical study designed to determine the factors that contribute to cloze test performance.

REVIEW OF THE LITERATURE

Developments In Cloze Testing

Taylor (1953) first suggested the cloze procedure for determining the difficulty, or "readability," of a text—reasoning that if several people could reproduce the missing words of a "mutilated" passage than the text must be easy to read, but if they could not supply the missing words the text must be difficult. Early studies validating cloze for this purpose reported correlations between cloze scores and other measures of readability as high as .68 (Rankin & Culhane, 1969), .95 (Bormuth, 1967), and .96 (Bormuth, 1968). However, it was not long before Taylor himself saw the potential to use cloze as a measure of reading comprehension. In a 1956 paper, he reasoned that "if the statement that a passage is 'readable' means that it is 'understandable,' then the scores that measure readability should measure comprehension too" (p. 44). Later research employed criterion related validation procedures to find out if cloze could be an appropriate measure of reading proficiency (for example, Bormuth, 1967, 1969; Rankin & Culhane, 1969; Ransom, 1968; Taylor, 1956; Weaver & Kingston, 1963). The findings of these studies were generally favorable for the use of cloze as a measure of reading comprehension with native speakers, yielding correlation coefficients ranging from moderate (0.54-0.71) to good and quite good (0.80-0.95).

However, further research has demonstrated that cloze tests tend to correlate well with almost every kind of language test (as pointed out in Bachman, 1982), thus subduing the usefulness of criterion measure studies as far as cloze tests are concerned. For example, in two criterion related studies (Hinofotis, 1980a; Oller & Conrad, 1971), traditional cloze tests were found to correlate not only with reading comprehension tests (.80 in Oller & Conrad and in Hinofotis) but also with dictation (.82, Oller & Conrad), listening comprehension (.71, Hinofotis), structure/grammar (.63, Hinofotis; .58, Oller & Conrad), and vocabulary tests (.59 in both Hinofotis and Oller & Conrad). Results such as these are what led Oller and Conrad (1971) and others to mark cloze procedures as integrative tests and as good measures of *general language proficiency* rather than the more specific *reading comprehension* alone.

Rather than accepting cloze as a cure all for general language proficiency, however, it seems more likely that correlations such as those just described are indications that we do not fully understand the cloze procedure and what it measures. Oller and Conrad (1971) admit as much, stating, "We believe it is not only possible to make good use of tests which require the student to perform *little-understood skills* [emphasis added], but that it is absolutely essential…The cloze procedure is one of those tests" (p. 187). I disagree, especially if it turns out that non-verbal skills may be required to complete cloze procedures. A test cannot be validated for any purpose if we do not know what skills are required to complete it.

In an attempt to find out, at least in part, what skills are utilized to complete cloze tests, studies of a different design were conducted by a number of researchers, starting in the mid-70's. A cut and scramble procedure was first implemented in 1975 by Oller and again in 1977 by Chihara, Oller, Weaver, and Chavez-Oller. These studies were designed to test the sensitivity of cloze tests as measures of the ability to synthesize information across sentence boundaries, a skill

thought to be crucial in reading. The method was to cut and scramble the sentences of a cloze passage, thereby destroying the natural cohesion created by organized sentences. In the 1977 study, two such passages were created, one having every 7th word deleted, the other every 6th word, and both sequential and scrambled versions were given to native and non-native speakers. Because both groups had more difficulty with the scrambled cloze passages, the authors concluded that the cloze procedure is "sensitive to discourse constraints ranging across sentences" (p. 68). These results were supported by similar research conducted by Cziko in 1978 with native and non-native speakers and by McKenna and Layton in 1990 with native speakers. In contrast, however, Kibby (1980), Shanahan and Kamil (1982), and Shanahan, Kamil, and Tobin (1982) both used a similar sentence scrambling procedure, also with native speakers, but found no significant differences between scrambled and sequential passages; thereby raising doubt as to whether test-takers utilize information across sentence boundaries in order to complete cloze procedures.

Other researchers have been interested in whether intersentential information is utilized to any substantial degree in order to complete traditional cloze tests, but have approached the question from another standpoint. Alderson (1978, 1983, 2000) and Bachman (1982, 1985), for example, have questioned the arbitrary selection of words for deletion via an every nth word deletion method. They claim that the semi-random deletion procedure tends to result in the deletion of a large quantity of function words because they are the most frequent type in the English language. These function words can normally be restored using sentence-level (syntactic) knowledge alone and attention to the broader text is thus not required for successful completion of the clozed items. Levenston, Nir, and Blum-Kulka (1984) argued that "semotactic"[iii] cloze

[iii] Defined in Levenston et al. as words whose meaning can be understood based on "the meaning and occurrence of other words within the context" (p. 204). Examples are collocations like "brothers and sisters" or "hummed to life".

gaps can also be filled by lexical knowledge alone without reference to the wider text. Such claims would suggest that random or semi-random deletion methods do not make good reading tests. As a result of these observations[iv], these researchers have suggested that cohesive devices and content words that rely on intersentential relationships should be targeted for deletion in order to tap into a reader's ability to make associations across sentences and between paragraphs. Alderson and Bachman have both suggested a *rational deletion* method (called a *gap-filling procedure*[v] by Alderson), whereby words are selectively chosen for deletion based on theoretical concepts of reading.

Alderson first suggested the *gap-filling* procedure in his 1978 dissertation when he found that, for second language learners, creating closure on traditionally deleted passages seems to be sentence or clause bound. As a means of validating a rationally deleted cloze procedure, Bachman (1982, 1985) created a rationally deleted cloze test and used item analysis to show whether certain types of words were more difficult to replace than others. The missing words were each classified according to the type of information that was presumably needed to replace them—for example, clause-level, sentence-level, intersentential, text level, and beyond the text. The item difficulty of each item was then used to determine which types of words were more difficult to fill in. In both studies, Bachman concluded that a rationally deleted cloze test can be used to measure discourse-level comprehension as well as sentence level comprehension, but can be tailored to do more of one or the other because of the selective process. The drawback to Bachman's research is that he was the only judge as to which skill(s) is required to fill-in the

[iv] Bachman (1985) was also concerned that deleting words arbitrarily may result in the loss of words central to the meaning of a text, since not all words carry the same amount of information, making some items difficult or impossible to restore.

[v] The "gap-filling" procedure suggested by Alderson and the "rational deletion" method suggested by Bachman are the same. Alderson feels that the gap-filling procedure should be separate from cloze procedure whereas Bachman views rational deletion as simply another type of cloze procedure.

missing word for each item.

Two researchers, Sasaki (1996) and Yamashita (2003), improved on Bachman's research a great deal by using participant think-aloud protocols. Sasaki administered a 294 word cloze test to eight ESL students and recorded the participants think-aloud reports as they took the test. She proceeded to classify each cloze item according to Bachman's classification system (with some modifications), but by referring to the recorded reports of the participants rather than by using her own judgment about what information would be used to fill in the missing words. Yamashita (2003) also used think-aloud verbal reports, administering a rationally deleted cloze test to twelve participants. By classifying the information that subjects reported using to fill in the blanks, Yamashita and Sasaki increased the validity of the item classification system created by Bachman.

However, there were some differences in the results of these two studies. While Sasaki (1996) found that clause level information was required to complete the majority of items on the cloze test in her study, Yamashita reported that "text-level information [that is information within the text but across sentences and clauses] was found to be the source of information most frequently referred to by the group as a whole" (p. 285). The reason for this difference is probably due to the different types of cloze tests used in the two studies. Sasaki's cloze test was created by an every 6^{th} word deletion method whereas Yamashita used a rationally deleted cloze test targeting items that would require intersentential information. Yamashita's method and results seem to support the use of rationally deleted cloze procedures to measure reading proficiency as Bachman predicted.

So far, the research has focused on discovering which language specific skills are utilized in completing cloze tests. There is evidence that a range of language skills are evoked during

cloze testing, such as word knowledge, grammar knowledge, discourse knowledge, and even world knowledge[vi] (also called background knowledge or schema) and that individuals may evoke these skills and knowledge differently depending on the test and the individual differences among the test takers too (see Brown, 2002). But it is not enough to know that vocabulary or grammar or intersentential information is utilized when completing a cloze test. Responsible testing requires us to know the degree to which each of these skills is evoked. Only by obtaining such clarity, can we make sound decisions about cloze test applications.

The present study was designed with this question in mind. It will ask, not only, "what skills are used to complete a cloze test?" but also, "to what degree is any given skill used?" Based on the rationale presented by Alderson, Bachman, and others, and the results from Bachman (1982, 1985) and Sasaki (1996), the rational cloze procedure is used in the present study as the more likely operationalization for reading comprehension.

Closure: A Precursor To Cloze

The other issue taken up in the present research is the relationship, if any, of closure to the cloze test. Taylor (1953) originally conceptualized the cloze procedure based on the idea of "closure", which itself originates in Gestalt psychology. Early Gestalt psychologists identified a human capacity to close gaps in a "familiar but not-quite-finished pattern" (Taylor, 1953, p. 415) such as the incomplete "B"[vii] below (a rather simplified example).

[vi] For interesting research that details the role of schema in reading comprehension see Alderson & Urquhart (1988), Caroll (1983 & 1984), Johnson (1981 & 1982), and Mohammed and Swales (1984).
[vii] Reproduced from http://www.ship.edu/~cgboeree/gestalt.html, retrieved May 5, 2005.

Thurstone identified and labeled this ability as the "first closure factor"[viii] and defined it as "the ability to perceive an *apparently* [emphasis added] disorganized or unrelated group of parts as a meaningful whole" (Thurstone & Jeffrey, 1996). A good example of this ability is demonstrated by Gestalt completion tests (for an example, see Appendix A), which have loaded heavily on the first closure factor in a number of studies (such as Botzum, 1951; Thurstone, 1944, 1949; Pemberton, 1952). The distinctive feature of Gestalt completion tests is that a figure appears only in part and the subject must supply the missing parts in his/her mind in order to complete the "gestalt,"[ix] or make meaning out of the picture (Street, 1931).

Although this ability does not seem to be language dependent, Taylor (1953) applied the principle of closure to language problems. For example, Taylor says, "Given 'Chickens cackle and _____ quack,' almost anyone can instantly supply 'ducks'" (Taylor, 1953, p. 416). He explains that in solving the cloze "one must guess what the mutilated sentence means as a whole, then complete its pattern to fit that whole meaning" (p. 416). While this adaptation of closure to language problems makes sense superficially, Taylor never tried to empirically determine whether cloze procedures in fact work the way he suggested that they do.

Treela 6/26/05 6:55 PM
Comment: A colleague pointed out that Gestalt completion tests may not measure the same thing as cloze because in Gestalt completion tests, the testee only needs to identify the whole, but does not have to draw in the individual parts, whereas, in a cloze test, the testee has to fill in the individual parts. Thus, my colleague suggested that Gestalt completion tests would be more analogous to reported the main idea or topic of a closed passage without providing the missing words. The point is well taken. However, according to the ideas articulated by the Gestalt psychologists, in order to achieve closure, the testee must supply the missing parts in his or her mind. Thus, even though they are not required to write them on the page, the assumption is that these parts are in fact supplied in order to make meaning out of the whole. In this way, cloze tests can be considered analogous to Gestalt completion tests.

It is worth investigating whether the cloze test does involve closure principles, as Taylor suspected, for two reasons. First, Taylor's hypothesis has only been tested in two studies that I am aware of (Kohler, 1966; Ohnmacht, Weaver, & Kohler, 1970), neither of which were conducted in second language (L2) contexts. Second, if cloze tests do require closure abilities, which are not language skills per se, than we have some basis for questioning the appropriacy of cloze procedures as tests of language proficiency. The studies that have explored the relationship between cloze and closure are described below.

[viii] Thurstone identified at least three "closure factors" but the third did not seem to bear out in later research. The second closure factor will be described later in this paper.
[ix] The word Gestalt means "a unified or meaningful whole": http://www.ship.edu/~cgboeree/gestalt.html.

Kohler (1966). Kohler was the first to investigate the relationship between cloze and closure in his 1966 unpublished dissertation "An Investigation of Cloze Scores in Terms of Selected Cognitive Variables". Using a multiple regression analysis, Kohler examined the relationships between eight different cloze passages and 11 different cognitive "factor-pure"[x] tests, including Hidden Patterns, the only test of closure in this study. The 11 cognitive tests were all taken from the *Kit of Reference Tests for Cognitive Factors* developed by French, Ekstrom, and Price (1963). For a detailed description of these tests, the reader is referred to the test manual. The eight cloze procedures were created from passages of approximately 250 words each, selected from college-level text books, covering the four topics of Biology, Chemistry, American Government, and World History. The cloze tests were created using an every 5th word deletion pattern and were scored by an exact answer scoring method.

The total battery of 19 tests was administered to 257 tenth grade native speakers of English. The main question, "What are the relationships between scores obtained on the selected factor-pure cognitive ability measures and scores on the cloze tests used in this study?" was answered through a step-wise multiple regression analysis. The dependent variables were four composite cloze tests and the independent variables were the 11 factor-pure tests. In every comparison, Wide Range Vocabulary, Logical Reasoning and Inference were consistently and significantly loaded on all four composite cloze tests. These findings provide evidence that cloze tests require word knowledge, reasoning and inferencing, all skills which are usually associated with reading.

[x] A factor-pure test is a test designed to measure a pure (single) factor or cognitive ability. Such tests will load heavily and consistently on a particular factor and are said to "mark" that factor. In comparison, most language tests are not "factor-pure" because several different cognitive abilities are brought to bear in completing them. For example a typical multiple choice reading test would not be considered factor pure. On the other hand, the Gestalt completion test mentioned earlier is thought to represent, in it's purest form, the perceptual ability speed of closure, and so is considered "factor-pure".

The relationship between cloze and closure was less clear. The only test representing closure in this study was the Hidden Patterns test and it loaded significantly on the "cloze" factor in only three out of four comparisons. As the only test in the battery directly connected to closure, it warrants some consideration here (for an example, see Appendix A). The Hidden Patterns test represents Thurstone's second closure factor, *flexibility of closure*, (see Thurstone, 1938, 1944, 1949 ; but also Botzum, 1951; Pemberton, 1952), and was described by Kohler (p. 52) as "the ability to discard given solutions in favor of better ones." Furthermore, he claimed that it "required the subject to search through irrelevant or distracting material" (p. 52). Oddly, it is not immediately obvious how the second closure factor could be related to cloze tests. In a cloze test, the testee must provide, from the mind, the missing components (words), not search for them among distracting material on the page. In the Hidden Patterns test, the testee searches among distracting information on the page to find a hidden figure or pattern that is present on the page as well, not missing as in the cloze test. Furthermore, it is impossible to explain how the text around a cloze gap could ever be considered "irrelevant". Quite the contrary, it is usually argued that the text surrounding a cloze item is the most relevant information for the solution. Kohler provides this explanation for the apparent relationship between cloze and *flexibility of closure*.

> It is possible for a given completed cloze unit to appear meaningful within the limited context of a phrase or sentence but to be incorrect of a larger unit such as a paragraph or entire passage. Therefore, this ability to discard early solutions in favor of better solutions found later would appear to be important in the cloze task (Kohler, 1966, p. 52).

This explanation seems plausible, but it does not exactly describe the task carried out in the Hidden Patterns test as I understand it (the testee must identify patterns hidden within *distracting*

material). Thus, theoretically, we are at a loss to explain the slight relationship found between flexibility of closure and cloze in this study.

Furthermore, the more obvious of the two closure factors to be used to complete a cloze procedure, speed of closure (the first closure factor), was not included in Kohler's study. As a result, his study does not take us very far in explaining the hypothesis suggested by Taylor (1953) that cloze is similar to closure (by which I believe he was referring to the first closure factor which measures a person's ability to unify a complex situation into a meaningful whole). However, the study does offer some preliminary evidence that such a relationship may exist.

Ohnmacht, Weaver, and Kohler, (1970). A few years after his dissertation, Kohler coauthored another study titled "Cloze and Closure: A Factorial Study" which used factor analysis to study more exclusively the possible relationship between cloze and closure. This study included four cloze tests and four tests of closure—two each of the first and second closure factors. In addition, the study included four verbal tests. The total test battery of 12 tests was administered to 113 high school students.

In their analysis, Ohnmacht et al. (1970) present three different factor solutions using two cessation methods (eigenvalue > 1 and the scree test) and two rotational analyses for each solution (varimax and maxplane). In all six solutions, a separate closure factor is clearly distinguished from other factors in the solution; that is, all four tests of closure load on one common factor, while the cloze and verbal tests load on other factors, sometimes together. These results indicate that closure abilities may not make a significant contribution to cloze test scores after all. The authors state, "The strength of this relationship is not such…that one would be led to accept the classical gestalt explanation of visual closure phenomena as descriptive of the process of completing cloze blanks" (p. 214). However, the diversity of tests used in this

study may not have been large enough to make the relationship manifest. Cattell (1979) points out that a factor is not only defined by the variables which load on it but, perhaps even more so, by those that do not; and, for this reason, a considerable number of diverse variables are needed in factor analysis. While the study included eight tests, the tests overlapped such that only three distinct variables were represented. In any case, the results obtained have never been confirmed by additional research.

An earlier study by Weaver and Kingston (1963) also used factor analysis to analyze the cloze procedure, but did not include any test of closure. The study is mentioned here, though, because of its implications about what cloze may be measuring. Comparing eight cloze tests to a variety of other language tests and to various subsections of the Modern Language Aptitude Test (MLAT), Weaver and Kingston obtained three factors defined by their loadings as “verbal comprehension", “redundancy utilization” (seven out of eight cloze tests load here), and “rote memory, flexible retrieval” (all sections of the MLAT load here). Initially, it seems strange that the cloze tests load together on a single factor, considering the general belief that cloze does measure some language ability and so should load on the verbal factor. However, there are two possible explanations for the strong presence of a cloze factor in this study. First, there were probably too many cloze tests (eight) in proportion to the number of other kinds of tests used in this study. Forty-four percent of the tests in the battery were cloze tests, as compared to 5-11% for any other construct/test represented. The researchers may have stacked the odds in their favor for obtaining a single cloze factor. Another interpretation of these results is also possible. That cloze did not load together heavily with other language tests may suggest that there is something else not represented in this study, something non-verbal, that is needed to complete cloze tests. Further research is warranted.

The research presented above is limited in what it can tell us about the possible role of closure in cloze testing. Kohler (1966) found word knowledge, logical reasoning and inferencing to load highly on cloze test scores; and closure less so, but he did not include more than one measure of closure in his study. In Ohnmacht, Weaver, and Kohler (1970), cloze and closure did not load on the same factor at all, however, the diversity of tests used in that study may not have been large enough to make such a relationship manifest. The third study, though it does not include any test of closure, offers some evidence that cloze may be measuring something non-verbal since the cloze tests in that study did not contribute to verbal test scores. These results combined indicate that further research in this direction is warranted. Furthermore, the relationship between cloze and closure among second language learners is virtually unstudied.

Clearly, the relationship between cloze, closure, and the various reading skills could withstand greater scrutiny and from a different perspective than has been taken up to now. Most studies have looked at cloze through item analysis, criterion-related validity or scrambled passage comparisons. Comparatively few have used factor analysis to isolate pure factors which may predict cloze test performance. There is a clear need for research of this kind.

Purpose of Research

Through factor and multiple regression analyses, the present study will address the general question: What does cloze measure? Previous research and theoretical considerations lead to five more specific questions.

1. Do human closure abilities contribute to cloze test scores, and if so, to what degree?
2. Do abilities of higher reasoning (non-verbal) contribute to cloze test scores, and if so, to what degree?

3. Does reading proficiency contribute to cloze test scores, and if so, to what degree?
4. Does grammar knowledge contribute to cloze test scores, and if so, to what degree?
5. Can word knowledge be confirmed to predict cloze test scores, and if so, to what degree?

METHODS

In order to address the five research questions, a battery of nine tests, as listed below, were administered to a group of English language students at the University of Hawai‘i, Mānoa. A description and rationale for selecting each test is provided in the sections that follow.

1. a cloze test (rational deletion)
2. a reading comprehension test
3. a vocabulary (synonyms) test
4. a grammar test
5. Gestalt completion, *Ekstrom, French, Harman, and Dermen (1976)*
6. Snowy Pictures, *Ekstrom, French, Harman, and Dermen (1976)*
7. Hidden Patterns , *Ekstrom, French, Harman, and Dermen (1976)*
8. Copying, *Ekstrom, French, Harman, and Dermen (1976)*
9. Raven’s Progressive Matrices (*SPM Plus* version), *JC Raven Ltd. (1998*)

Three of the nine tests were administered to incoming students at the English Language Institute (ELI) at the University of Hawai‘i, Mānoa as a part of its regular placement test battery: the cloze test, the reading test and the vocabulary test. ELI students were asked to voluntarily sit for the remaining six tests and were compensated for their time. The set of scores for the nine tests was submitted to a factor analysis and multiple regression analysis and results are presented below.

Participants

ELI students range in both native language and educational background—the ELI student body includes undergraduate freshman up to graduate doctoral students with TOEFL scores ranging from 500 to 600[xi]. Since the students participating in this study were all new students, their language performance varied to an even larger degree than continuing students might be expected to do since they will not have had a chance to assimilate.

Fifty-three students who had taken the ELIPT volunteered to participate in the study. The sample is considered moderately heterogeneous based on the variety of native languages represented (see Table 1) and the distribution of ages ranging from 18 to 46 years with a mean of 25.32 years. English proficiency was also heterogeneous based on reported TOEFL scores and age began learning English. Forty participants were able to report TOEFL scores and the scores

Table 1

Native Language

L1	Frequency	Percent	L1	Frequency	Percent
Cantonese	3	5.7	Persian	1	1.9
Chinese	7	13.2	Polish	1	1.9
Chamorro	2	3.8	Samoan	1	1.9
Indonesian	6	11.3	Sinhalese	1	1.9
Japanese	14	26.4	Tamil	1	1.9
Korean	5	9.4	Thai	4	7.5
Mandarin	1	1.9	Vietnamese	5	9.4
Myanmar	1	1.9			
			Total	53	100.0

[xi] All students who receive a TOEFL score between 500 and 600 are referred to the ELI to take the placement exams. Though, many students who score above 60 on subsections of the placement exams may be exempted from ELI classes as a result, even though their TOEFL score is below the standard university requirement (600). In fact, an average of 26.49% of students who took the ELI placement exams over the last three semesters were exempted from classes (Kenton Harsch, email communication, Nov. 1, 2005). This system recognizes that one test score (such as TOEFL) can be an inaccurate representation of a student's abilities, and it is therefore possible for students with wide ranging English proficiency to take ELI placement exams.

ranged from 510 to 600 with a mean of 555. Students began learning English as young as one year old and as late as 24 years with a mean of 11 years.

Tests

As listed above, nine tests were administered to the sample. A description and a rationale for selecting each test is provided here.

The cloze test. The cloze test currently in use at the ELI was developed in 1995 by Scott Todd. The 819 word passage was taken from a University level textbook. The topic was chosen because it was not believed to "favor any group of people". Todd intended that this topic would only "require backgrounds [sic] knowledge that is equal for all students" (Todd, 1995)[xii]. In developing the cloze test, 78 words—focusing on cohesive devices and content words—were deleted from the passage. The test was piloted and submitted to an item analysis to find 50 "good" items to be retained for the final version. The test is scored by acceptable word scoring[xiii] and a list of acceptable words is maintained and updated any time a new possibility arises. The reliability of this cloze test has been reported in Clark (2002) at .82 (KR-21, Fall 2001) and .85 (KR-20, Spring 2002). A more recent reliability coefficient of .85 (Cronbach Alpha) was obtained for the Fall 2004 test data also (calculated by present researcher). Because the cloze test is part of the ELI's placement battery, the test cannot be printed here. However, an example is provided on the ELI website and so does not pose a security threat. The example is given in Appendix B.

[xii] A great deal of research underscores the importance of background knowledge in reading comprehension. For comprehensive discussions of this issue, see Alderson & Urquhart, 1988; Carroll, 1983, 1984; Johnson, 1981, 1982; and Mohammed & Swales, 1984.

[xiii] For a comparison and evaluation of scoring methods see Alderson, 1983; Brown, 1980; Hinofotis, 1980a; Oller, 1972b.

The reading comprehension test. A 25 item multiple choice test, including six different passages taken from university level textbooks, makes up Section I of the ELI Reading Comprehension Test (RCT). One of the six passages is a pilot item with five questions that accompany it. Only the 20 operational items were used in the present analyses. No example of this test can be provided because of test security, however, a brief description of the test instructions are provided in Appendix B. The reading test is regularly submitted to item analysis and descriptive statistics to ensure quality. Past reliability coefficients are reported below.

The vocabulary test. A 25 item multiple choice synonyms test is also part of the ELIPT battery. Section II of the RCT includes 24 operational items and one pilot item. New items are piloted at every administration of the RCT and included in an item bank for use in future administrations of the test (Clark, 2003). The pilot item was omitted from the present analyses. An example of the vocabulary test (Section II of the RCT), taken from the ELI website, is given in Appendix B.

All ELI tests are regularly submitted to item analyses and descriptive statistics to ensure quality. The reliability of the RCT (Sections I and II combined) has been reported in Clark (2002) at 0.87 (KR-20, Fall 2001) and .85 (KR-20, Spring, 2002).

The grammar test. The grammar test was constructed for the purpose of this study. It was not a part of the ELI placement test battery. The test includes 50 items. In order to avoid overlap with skills required in cloze testing, the frequently used method of fill in answers was avoided. Instead, the methods of multiple choice error identification and error correction were utilized. For error identification, the testee was given a sentence with four underlined portions labeled A, B, C and D. One of these underlined portions is grammatically incorrect and the testee must identify and circle the error. An example of such an item is provided in Appendix B.

The error correction items are identical in form to the error identification items, except that the testee was required to correct the mistake in addition to identifying it. This second task is thought to be more challenging and appropriate for the more advanced level of ELI students. This test was piloted before the main study to ensure that a good distribution of scores could be achieved among the population using this test. The 80 item pilot test realized a Cronbach alpha reliability of .96. The test was modified through an item analysis to include 50 items for the final version.

Speed of Closure. Two tests were selected to represent speed of closure in this study for the main reason that, if only one test were used, we would be unable to make a strong argument for a closure factor, if one exists. On the other hand, with two tests, if they should both load on the same factor, we will have greater evidence for the closure factor. The two tests chosen were *Gestalt Completion* and *Snowy Pictures*, both taken from the *Kit of Factor-Referenced Cognitive Tests* (Ekstrom, French, Harman, & Dermen, 1976) distributed by ETS. Gestalt Completion tests have loaded prominently and consistently on the speed of closure factor in previous research (Botzum, 1951; Thurstone, 1944, 1949; Pemberton, 1952). Compared to Gestalt Completion, Snowy Pictures is a more recently developed test of the first closure factor (Ekstrom et al. 1976). Both of these tests require the least amount of culture and language specific knowledge. The test-taker is required only to identify and name an obscured or incomplete figure that they perceive in the field. The images are selected carefully so as not to require special cultural experience, and, furthermore, participants were allowed to answer in any language so that second language vocabulary knowledge would not confound their scores on this measure.

Gestalt Completion tests are created by removing parts of complete images or figures in a picture. The partial figure is then shown to the test-taker and the test-taker must supply the

missing parts in his/her mind in order to complete the "gestalt," or make meaning out of the picture (Street, 1931). The figure in the picture is normally white on a black background or vice versa (see Appendix A for an example). The test has two parts with ten items in each part. The test-taker has two minutes to complete each part. The final score is the combination of correct answers from parts one and two (maximum of 20). Ekstrom et al. report reliabilities of .82 and .77 for this test in their 1963 studies and .85 in the '76 revised version (Ekstrom et al., 1976).

Snowy Pictures tests are created by obscuring an object with snow-like spatters (see Appendix A for an example). The test has two parts with 12 objects in each part. The test-taker has three minutes to complete each part by identifying the object and writing its name on the line below the picture. The final score is the combination of correct answers from parts one and two (maximum of 24). A reliability coefficient of .86 is reported in Ekstrom et al. (1976).

Flexibility of Closure. As with speed of closure, flexibility of closure will be measured by two tests for the same reason given above. Both tests were taken from the *Kit of Factor-Referenced Cognitive Tests* (Ekstrom, French, Harman, & Dermen, 1976) distributed by ETS. The first is the *Hidden Patterns* test which is based on a test Thurstone himself developed in 1938. Hidden Patterns was chosen to represent flexibility of closure because it loads prominently and consistently on the closure flexibility factor (Botzum, 1951; Thurstone, 1944, 1949; Pemberton, 1952) and involves the least amount of culture and language specific knowledge. In this test, the test-taker must identify a hidden figure within the lines of a larger figure as quickly as possible. There are 200 figures in the test and the test-taker is given three minutes to identify the model figure as many times as possible. An example is shown in Appendix A. The final score is number correct minus the number incorrect. Incomplete items

are counted as incorrect (maximum score of 200). This test obtained a reliability of .81 in the '63 version (reported in Ekstrom et al. 1976).

The second test chosen to represent flexibility of closure was a copying test which loaded heavily on the flexibility of closure factor in Thurstone's 1949 study. In *Copying*, the test-taker is asked to copy a pattern onto a field of dots so that it matches the model as closely as possible. The participant is given three minutes to copy as many items as possible in each part. There are two parts in the test with 32 items on each. The final score is the combination of correct answers from parts one and two (maximum of 64). The reliability coefficient obtained from the '63 version is reported in Ekstrom et al. (1976) at .84.

These tests are said to measure "the ability to keep in mind a configuration despite distraction" (Thurstone, 1949, p. 17). Although this skill does not seem to be clearly related to the demands required in the cloze procedure, it may be related to skills required in reading; although, that relationship is not obvious given the definition above. Thurstone (1949) conjectured that flexibility of closure could be generalized outside the perceptual domain as the ease with which a person can keep the essential features of a complex situation in mind despite distraction (p. 17). In reading, a reader does need to keep individual ideas in mind while he/she continues to read and needs to avoid being distracted by irrelevant or ambiguous details that may be present in the prose. There is no evidence that this relationship exists, but it will be interesting to see how reading, cloze and flexibility of closure pan out in the analysis.

Raven's Progressive Matrices. The Progressive Matrices tests, originally developed by J. C. Raven in 1936, were designed to measure "a person's ability to form perceptual relations and to reason by analogy independent of language and formal schooling" (Raven, 2004). In this study, scores on the RPM are interpreted as a measure of a general, language-independent

cognitive ability and it is believed that the problem solving and analogies problems are similar to processes involved in reading comprehension. The test contains five sets of 12 problems each. Each problem contains a diagram or design with a part missing. A number of possible solutions are printed beneath the diagram and test-takers are expected to select the correct part to complete the design. An example of such an item, taken from J. Raven (2000) is presented in Appendix B for illustration. Items in the RPM get progressively more difficult as one proceeds through the test, hence the name. The Standard Progressive Matrices *Plus* (SPM) version, developed in 1998, was selected for this study because it was designed to compensate for the observed inflation in scores in the original Standard version. Dobrean, Raven, Comsa, Rusu, and Balazsi (2005) recently obtained a reliability coefficient of .88 for this version of the test. Other reliability coefficients have been reported at .90, .83, and.93 (Schuhfried).

Procedures

Three of the nine tests discussed above were administered to 233 second language students at the ELI in the Fall of 2005 as a part of their regular placement test battery. Fifty-three volunteers sat for the additional six tests within a month of the ELI placement tests. Several testing sessions were scheduled and participants signed up at their convenience. All tests were administered in groups. During these testing sessions, the following procedures were followed. Participants completed the consent form shown in Appendix D. The four closure tests, all timed tests, were completed as a group. Together these four tests took approximately 30 minutes to administer. Immediately following the closure tests, the grammar test was administered and participants were given 35 minutes to complete the test, including time for reading the instructions. A short break followed which was in turn followed by the administration of the RPM (Standard *Plus* version). For practical reasons, the participants were

given 60 minutes to complete the RPM, though some were not finished at that time. The test manual indicates that the test should take between 20 and 40 minutes to complete and that testees should be stopped once they miss five in a row. Because the test was group administered it was impossible to monitor the progress of every participant, so the 60 minute time marker was used to signal when the test had gotten too difficult for the testee instead of five wrong in a row. Testing sessions took approximately two hours in total and participants were reimbursed for the contribution of their time at the end of each session.

Results from the ELI placement exams (cloze, reading and vocabulary) were obtained and compared with the additional six tests (the grammar test, the four closure tests and the Raven's Progressive Matrices). Test scores were analyzed by applying two different techniques: an exploratory principle factor analysis and a multiple regression analysis.

RESULTS

The results are presented in three broad categories below: (a) descriptive statistics, (b) the factor analyses, and (c) the multiple regression analyses. First, a brief statistical description of the nine variables will be presented. Second, the results that define the relationships between the variables will be presented. And finally, the results that indicate the predictive power of the independent variables for cloze test performance will be presented.

Descriptive Statistics

After all the data had been collected, there were 48 complete cases. While there were 233 students who took the ELI placement exams this fall, the additional tests used in this study

were taken voluntarily by only 53 participants. Four of these volunteers, as it turned out, had been exempted from the ELI placement exams and were taking ELI classes of their own volition. Since ELI test data was not available for these four, they were excluded from the analyses. Also one student participated in the voluntary study twice and so her second set of data was discarded. In the end, a complete set of data was only available for 48 participants. This is admittedly a low *N* size for factor and multiple regression analyses. Kline (1994) sets 50 as a minimum.

The descriptive statistics of the nine tests used in this experiment ($N = 48$) are shown in Table 2 and Table 3. Table 2 shows the results for the nonverbal tests (the RPM, Gestalt, Snowy Pictures, Hidden Patterns, and Copying). Table 3 shows the results for the verbal tests (grammar, vocabulary, reading, and cloze). By examining the means, medians, modes, standard deviations, and ranges for each test, it can be seen that the distribution of scores for each of the tests is approximately normal. To save space, skew and kurtosis data are not reported. However, it should be noted that none of the tests were skewed except for the Gestalt test which was slightly negatively skewed. To adjust for that skew, the Gestalt scores were transformed and the transformation is described in Appendix C. The transformed data were used in the main analyses.

The reliability estimates[xiv] for each test are also given in Tables 2 and 3. The reliabilities of the nonverbal tests are all quite high accept for the Snowy Pictures test which only has moderate reliability (.47). The reliabilities of the verbal tests were also reasonably high, with the exception of the Reading Comprehension test which had a moderate reliability of .60.

[xiv] $N = 48$ for all reliability calculations except for the grammar test. Three participants misunderstood the instructions and completed the test incorrectly. Therefore, they were excluded from the reliability calculation. However, the mean score was replaced for these three cases in the subsequent analyses.

Table 2

Descriptive Statistics of the Nonverbal Tests

N = 48	RPM	Gestalt	Snowy Pictures	Hidden Patterns	Copying
k	60	20	24	200	64
Mean	41.42	11.42	14.23	7.17	40.60
Std. Error of Mean	.85	.64	.47	12.13	2.05
Median	42.00	12.00	14.00	6.00	39.50
Mode(s)	42, 45	16	12	-26[b]	64
Std. Deviation	5.86	4.41	3.28	84.01	14.23
Variance	34.38	19.44	10.78	7057.68	202.54
Range	30	17	13	348	54
Minimum	25	1	8	-162	10
Maximum	55	18	21	186	64
Reliability[a]	.82	.77	.47	.998	.91

[a] Cronbach alpha reliability unless otherwise stated
[b] Multiple modes exist. Other modes include -24, 16, 34

Table 3

Descriptive Statistics of the Verbal Tests

N = 48	Grammar[b]	Vocabulary	Reading	Cloze
k	50	24	20	50
Mean	28.87	13.27	12.50	24.21
Std. Error of Mean	1.23	.74	.46	1.06
Median	27.00	13.50	13.00	24
Mode(s)	27	18	14	23, 27
Std. Deviation	8.22	5.09	3.18	7.33
Variance	67.48	25.9	10.09	53.66
Range	35	19	12	31
Minimum	8	4	5	10
Maximum	43	23	17	41
Reliability[a]	.86	.82	.60	.81

[a] Cronbach alpha reliability unless otherwise stated.
[b] *N* = 45; Three participants did not complete the grammar test according to instructions so their data has been omitted from the descriptive statistics analyses. Mean scores are used in later analyses however to keep the *N* size constant across variables.

Factor Analysis

Before examining the results of the factor analysis, a preliminary look at the correlation matrix for the nine variables will be informative. The nine variables have been grouped in Table 4 such that a pattern starts to emerge. As indicated by the correlations, the cloze, grammar, reading and vocabulary tests each have moderate correlations with each other but not with the RPM or the four closure tests. Also, the RPM, Gestalt, Snowy Pictures and Copying tests have moderate correlations with each other but not with the other tests. Hidden Patterns does not correlate very well with either group, but fits better with the latter than the former. This correlation matrix seems to indicate two major groupings of the variables in this study: a verbal group and a nonverbal group.

Table 4

Correlation Matrix

	1	2	3	4	5	6	7	8	9
1. RPM	1.00								
2. Gestalt[a]	.47**	1.00							
3. Snowy Pictures	.45**	.65**	1.00						
4. Copying	.48**	.52**	.51**	1.00					
5. Hidden Patterns	.25	.33*	.32*	.49**	1.00				
6. Grammar	.11	.30	.20	.23	.26	1.00			
7. Cloze	.13	.07	.05	.02	.23	.58**	1.00		
8. Vocabulary	-.23	.08	-.19	-.12	.17	.57**	.54**	1.00	
9. Reading	.26	.21	.20	-.00	.30*	.43**	.56**	.58**	1.00

[a] Transformed data were entered into the analysis (see Appendix C) and all further analyses.
* $p < .05$. ** $p < .01$

To explore the relationships between the nine variables more exactly, a factor analysis was carried out using the correlation matrix in Table 4. Principle axis factoring with varimax rotation was used. Before choosing the varimax rotation, however, a preliminary analysis using

an oblique rotation method was carried out and the resulting factor scores were correlated to find out whether they were oblique or orthogonal. The initial factor scores were highly orthogonal (r = .19) so an oblique rotation was not necessary and the orthogonal rotation method was applied. Varimax is the most common orthogonal rotation in the area of factor analysis, so it was chosen for these analyses in order that the data could be compared to other research. In order to determine the number of factors to extract, two methods were used. Eigenvalues greater than 1 were chosen for extraction and verified by an examination of the slope of the scree plot of the eigenvalues. Both methods indicated that only two factors could be extracted.

The results of the first factor analysis are shown in Table 5. As with the correlation matrix, the factor analysis indicates two clear groups of variables. Factor 1 seems to be clearly defined by the five variables loading heavily on it as a nonverbal factor whereas Factor 2 can be clearly defined as a verbal factor by the four verbal tests loading rather heavily on it. The nonverbal factor is best defined by Snowy Pictures with a factor loading of .77 and almost

Table 5

Rotated Factor Matrix

Variable	*Factor 1*	*Factor 2*	h^2
RPM	.64	.02	.40
Gestalt	.73	.16	.56
Snowy Pictures	.77	.01	.59
Copying	.74	-.00	.55
Hidden Patterns	.46	.28	.29
Grammar	.23	.69	.53
Cloze	.06	.73	.54
Vocabulary	-.20	.85	.76
Reading	.18	.69	.51
Eigenvalue	2.41	2.32	
Proportion of variance	.27	.26	.53

equally as well by Copying and Gestalt Completion with factor loadings of .74 and .73 respectively. The verbal factor is best defined by Vocabulary with a factor loading of .85 and second best by Cloze at .73. Grammar and Reading define the verbal factor equally well at .69.

The communalities in the fourth column of Table 5 indicate the amount of variance in each variable explained by the two factors. These communalities range from a low of 0.29 (Hidden Patterns) to the highest at 0.76 (Vocabulary). On average, the two factors do not account for more than half of the variance in any one variable, suggesting that some factor may be unaccounted for by this set of variables. The very low communality of the Hidden Patterns test indicates that its contribution to this model is negligible (.29) and therefore probably should not be included in the regression analyses. The proportion of variance accounted for by each factor in the model is given in the last row of Table 5. These proportions indicate that Factor 1 accounts for 27% of the variance in the model and factor 2 accounts for 26% of the variance. Together, this model can account for 53% of the total variation.

A surprising result of this analysis is that the expected separate factors for Closure 1 and Closure 2 (speed of closure and flexibility of closure) do not appear. In order to explore whether the verbal/nonverbal difference was overpowering other relationships between the tests, additional factor analyses were conducted on subsets of the variables. Table 6(A) shows a factor

Table 6

Factor Matrices: (A) Verbal, (B) Nonverbal

(A) verbal tests	1	h^2	(B) nonverbal tests	1	h^2
Grammar	.71	.50	RPM	.67	.45
Cloze	.77	.59	Gestalt	.80	.63
Vocabulary	.77	.59	Snowy Pictures	.74	.55
Read	.70	.49	Hidden Patterns	.47	. 22
			Copying	.77	. 59

analysis of just the verbal tests, while Table 6(B) shows a factor analysis of just the nonverbal tests. In both cases, only one factor could be extracted when the eigenvalue ≥ 1 rule was employed, so no deeper relationships were revealed by this secondary analysis. However, in the case of the nonverbal tests, when two factors were requested in a third analysis (ignoring the size of the eigenvalues) the four tests did divide into two factors in a manner consistent with expectations (see Table 7); That is, Gestalt Completion and Snowy Pictures loaded on Factor 1 and Hidden Patterns and Copying loaded on Factor 2. The reader may notice in this case the larger loading of the Raven's Progressive Matrices on the first factor, here interpreted as Closure 1 (speed of closure).

Table 7

Rotated Factor Matrix – Nonverbal Tests: 2 Factors

	1	2	h^2
RPM	.54	.36	.43
Gestalt	.73	.35	.66
Snowy Pictures	.80	.23	.70
Copying	.41	.79	.79
Hidden Patterns	.18	.56	.34

When a similar test was conducted on the verbal tests (that is two factors were requested) there was no meaningful difference in the result; that is all the variables continued to load more heavily on the same factor. Since the analysis did not provide any useful information, the results are not presented.

Multiple Regression Analysis

The multiple regression analysis was conducted in order to discover which subset(s) of the variables are most effective for predicting cloze test scores. With the cloze test set as the

dependent variable, six independent variables were entered into the regression as follows. First, the third factor analysis revealed that some of the data could be reduced. Hidden Patterns and Copying loaded together on one factor, indicating they more than likely represent the same factor, flexibility of closure. Gestalt Completion and Snowy Pictures also represent a single factor, speed of closure. Since it was found that Hidden Patterns is contributing very little to this model, however, Copying alone was chosen to represent flexibility of closure in the regression analysis. On the other hand, since Gestalt Completion and Snowy Pictures contribute equally to the model, a composite (average) score was computed to represent speed of closure in the regression analysis. The composite score was analyzed for normality, skewedness, and outliers. The data were normal and there were no outliers.

The results of a stepwise regression analysis are summarized in Tables 8, 9, and 10. Table 8 shows the independent variables that were included in each of the two steps of this regression, while Table 9 shows the variables that were excluded at each step. Table 10 summarizes the regression coefficients and part and partial correlation coefficients for the variables included in the model.

Table 8

Summary of S

Dependent Varia

Step	Independent Variables	*R*	R^2	ΔR^2	*df*	MS	*F*
1	Grammar	.58	.34	.34	1, 46	871.12	23.78**
2	Grammar, Reading	.68	.46	.12	2, 45	585.12	19.00**

[a] Criteria: Probability of F to enter <= .05, Probability of F to remove >= .10

** Significant at the *p* < .01 level

Table 9.

Summary of Stepwise Regression Analysis [c]:

Variables Excluded (N = 48)

Step	Independent Variables	*β*	t
1 [a]	Reading	.38	3.12**
	Vocabulary	.30	2.17*
	RPM	.07	.540
	Closure 1 (composite)	-.11	-.869
	Closure 2 (copying)	-.12	-1.01
2 [b]	Vocabulary	.14	.96
	RPM	-.02	-.16
	Closure 2 (copying)	-.08	-.74
	Closure 1	-.16	-1.36

[a] Predictors in the Model: Grammar

[b] Predictors in the Model: Grammar, Reading

[c] Dependent Variable: Cloze

* Significant at the p < .05 level ** Significant at the *p* < .01 level

Table 10

Summary of Regression Coefficients with Part and Partial Correlation Coefficients [a] *(N = 48)*

Model	Independent Variables	*B*	*SE* (*B*)	95% CI for *B* Lower Bound	95% CI for *B* Upper Bound	*β*	t	Correlations *r*	Correlations *pr*	Correlations *sr*
1	(Constant)	8.36	3.32	1.68	15.05		2.52*			
	Grammar	.54	.11	.32	.77	.58	4.88**	.58	.58	.58
2	(Constant)	1.73	3.72	-5.76	9.21		.46			
	Grammar	.39	.11	.16	.62	.42	3.47**	.58	.46	.38
	Reading	.88	.28	.31	1.45	.38	3.12**	.56	.42	.34

[a] Dependent Variable: Cloze
* p < .05 ** *p* < .01

It may be noticed straight away that the stepwise regression model (Table 8) does not include any of the nonverbal variables, confirming the results of the previous factor analysis and indicating that there is no relationship between the nonverbal variables and the cloze test in this study. Recall that in the factor analysis, a clear distinction between the verbal and nonverbal tests was revealed. However, in order to be sure that any possible relationships among the nonverbal tests and Cloze were not obscured by the greater relationships observed among the verbal tests, a second regression analysis was conducted, including only the nonverbal tests as independent variables. The results of this analysis are presented in Table 11. As shown in that table, none of the multiple correlation coefficients were significant at the .05 level and the R^2 was practically zero, confirming again that the nonverbal tests do not contribute to cloze test performance in this study and cannot be used to predict cloze test scores.

Table 11

Summary of Regression Analysis: Nonverbal Tests Only [a] *(N = 48)*

Independent variables	R^2 = .02 Intercept = 2.17	Correlations			
	β	Zero-order	pr	sr	*t*
RPM	.146	.13	.12	.12	.80
Closure 1	.023	.06	.02	.02	.12
Closure 2	-.064	.02	-.05	-.05	-.35

[a] Dependent Variable: Cloze
* p < .05 ** *p* < .01

Returning to the main regression analysis presented in Table 8, Grammar and Reading were the only two variables that made significant contributions to the model. Vocabulary, surprisingly, did not make it into the model, indicating that it does not contribute anything unique to cloze test scores after grammar ability and reading proficiency have been accounted for. The stepwise regression indicates that among the three independent verbal constructs, grammar knowledge, reading proficiency, and word knowledge, grammar knowledge is the most important predictor of cloze test performance, explaining 34% of the variance in cloze test scores, and reading proficiency is the second most important predictor, explaining the remaining 12% of the variance beyond that which grammar can explain (R^2 = .34 + .12 = .46).

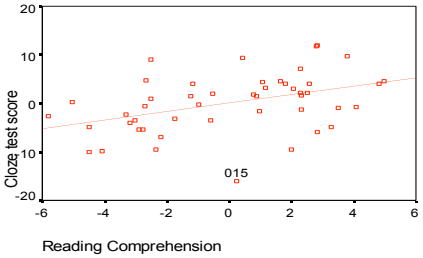
The semipartial correlation coefficients (*sr*) for Grammar and Reading shown in Table 10 indicate the degree to which either variable contributes uniquely to Cloze. The semipartial correlation of Grammar with Cloze was .38 while the semipartial correlation of Reading with Cloze was .34.

Further examination of the standardized scatterplots for grammar and reading predicting cloze give more information about the relative effectiveness of each. Figures 1 and 2 show the standardized scatterplots of each independent variable against the dependent variable. The

steeper slope shown in Figure 1 indicates, in fact, that the grammar test would be a more effective predictor of the cloze test score. On the other hand, however, the spread of the data points seems to be more even in the case of the reading test (Figure 2). Therefore, choosing the more effective predictor may depend on the particular application of the tests.

Figure 1. Partial Regression Plot of Grammar

Figure 2. Partial Regression Plot of Reading



An Additional Analysis: Vocabulary and Cloze

It was found that Vocabulary did not make a significant contribution to variation in cloze scores in the main regression analysis, but this result can be misleading. That word knowledge does not contribute to the variance in cloze test scores when grammar and reading are first considered does not indicate that word knowledge is not utilized at all when completing a cloze test. What it does indicate is that vocabulary knowledge probably overlaps with the constructs of reading proficiency and grammar knowledge as they were operationalized in this study; its contribution to Cloze is just not *unique*. To confirm that Vocabulary does make a significant contribution to Cloze when it alone is considered (and to establish the degree to which it does so), a simple regression was performed using Grammar and Vocabulary only as the independent variables and Cloze as the dependent variable. The results are shown in Tables 12 and 13. Table

12 gives the multiple R, R^2, and change in R^2 and shows that Vocabulary contributes 6% of the variance beyond what grammar contributes alone (ΔR^2 = .06) and that the contribution is significant at F (2, 45) = 19.00. Table 13 further indicates the degree to which Vocabulary can predict Cloze scores (sr = .25).

Table 12

Summary of Regression Analysis: Vocabulary and Grammar Only [a] *(N = 48)*

Model	Independent Variables	*R*	R^2	ΔR^2	*SE*	*df*	*MS*	*F*
1	Grammar	.58	.34	.34	6.05	1, 46	871.12	23.78**
2	Grammar, Vocabulary	.64	.40	.06	5.82	2, 45	515.10	15.19*

[a] Dependent Variable: Cloze
* p < .05 ** *p* < .01

Table 13

Summary of Regression Analysis: Vocabulary and Grammar Only [a] *(N = 48)*

Regression Coefficients with Part, and Partial Correlations

						Correlations			Collinearity Statistics	
Model	Indep Variables	*B*	*SE (B)*	*β*	*t*	*r*	*pr*	*sr*	*Tol*	*VIF*
1	(Constant)	8.36	3.32		2.52*					
	Grammar	.54	.11	.58	4.88**	.58	.58	.58	1.00	1.00
2	(Constant)	7.14	3.25		2.20*					
	Grammar	.38	.13	.41	2.95**	.58	.40	.34	.68	1.47
	Vocabulary	.44	.20	.30	2.17*	.54	.31	.25	.68	1.47

a Dependent Variable: CLOZE
* p < .05 ** *p* < .01

DISCUSSION

The major purpose of the present study was to ascertain what cloze measures. In order to achieve this purpose, five specific research questions were posed. In the section that follows, each of these questions will be discussed consecutively.

Q 1. Do Human Closure Abilities Contribute To Cloze Test Scores, And If So, To What Degree?

The results of this study indicate that, in fact, human closure abilities do not seem to contribute to cloze test scores. In the factor analysis, none of the closure variables included in this study loaded on the same factor as the cloze test to any significant degree. This result was confirmed in the regression analysis where all of the closure variables were excluded from the regression model because their contributions were negligible. This result confirms previous results obtained in Kohler (1966) and Ohnmacht, Weaver and Kohler (1970) who, despite their own conclusions, found very weak relationships between cloze and closure abilities.

Q 2. Do Abilities Of Higher Reasoning (Non-Verbal) Contribute To Cloze Test Scores, And If So, To What Degree?

The Raven's Progressive Matrices was used in this study to represent the construct of higher order reasoning (non-verbal). In the factor analysis, the RPM loaded moderately high on the first factor, defined as the non-verbal factor in this study, but did not load at all on the verbal factor where the cloze test was. This result indicates that the two constructs are not related to any substantial or significant degree. Furthermore, as with the closure variables, the RPM variable was excluded from the regression model because the contribution it made to cloze was too low to consider. This result implies that non-verbal reasoning skills are not utilized to complete the cloze procedure used in this study.

The only study, to my knowledge, which also compared the RPM to a cloze procedure was conducted by Flahive in 1980. Flahive found a much higher correlation between cloze and the RPM (R = .61) than was found here (R = .13; Table 4). However, his study differs from this one in three important ways. First, Flahive used a semi-random deletion method (every 7th word) for his cloze test whereas a rationally deleted cloze test (focusing on cohesive devices and content words) was used in the present study. Second, Flahive used the original Standard Progressive Matrices test whereas the *Plus* version was used in this test. Finally, the sample size in Flahive's study was less than half the size (N = 20) of the present study. Because of these differences, the results cannot be compared directly.

Q 3. Does Reading Proficiency Contribute To Cloze Test Scores, And If So, To What Degree?

Reading and Cloze both loaded substantially on the same factor in the factor analysis (factor loadings of .69 and .73, respectively) and Reading contributed significantly to the regression model with Cloze as the dependent variable, as well, indicating that reading proficiency does contribute to cloze test performance. The best indicator of the degree of this contribution is given by the semipartial correlation coefficient (Table 10), sr = .34. It should be kept in mind that 54% of the variance (100 - R^2 = 100 - .46 = .54; see Table 8) in the cloze test cannot be explained by the variables in this study. Based on these results, however, it appears that reading skills do contribute to the rationally deleted cloze test used in this study, though probably not to the degree we would hope for when cloze is employed as a reading test.

Q 4. Does Grammar Knowledge Contribute To Cloze Test Scores, And If So, To What Degree?

Grammar knowledge also appears to contribute to the cloze test used in this study and, in fact, was the highest predictor of cloze test performance. According to the semipartial

correlation coefficient, grammar scores uniquely predict 38% (sr = .38) of the cloze test scores. This result is somewhat surprising when we consider that the rationally deleted cloze test intentionally tries to avoid grammar dependent items in favor of contextually dependent items (reading). It would be a fruitful and interesting next step, therefore, to look at the cloze test used in this study anew and conduct an item analysis on the test to see whether the items are, in fact, contextually dependent as the procedure would suggest.

These results are further surprising since great care was taken to write a grammar test that avoids "cloze type" items so that the two tests could not be said to be very similar. That is, the grammar test used in this study did not include any fill-in-the-blank items that are often found on grammar tests and that are certainly required by the cloze procedure. Therefore, I feel more confident in suggesting that a real and substantial relationship exists between grammar knowledge and cloze test performance based on these results.

A caution is warranted when interpreting the results pertaining to questions three and four however. Though grammar appears to be the largest predictor of cloze test scores in this study—based on the regression model produced in Table 8—in reality, the difference between the contributions made by Grammar and Reading (as indicated by their semipartial correlation coefficients, .38 and .34, respectively) is not that great and may be an artifact of the particular characteristics of the present study. Small differences in sample size and populations could likely yield opposite results; that is, with reading proficiency predicting a greater portion of cloze scores than grammar knowledge. However, I would anticipate that the respective contributions of grammar and reading abilities to cloze test scores are at least relatively equal.

Q 5. Can Word Knowledge Be Confirmed To Predict Cloze Test Scores, And If So, To What Degree?

Surprisingly, word knowledge did not factor into the regression model for predicting cloze test scores in this study. However, there are some caveats in this regard that should be considered. First of all, Vocabulary correlated to almost the same degree with the cloze test as did Reading and Grammar. (.54, .56, and.58, respectively). Also vocabulary was the heaviest marker for the verbal factor (.85) in the factor analysis with cloze being a close second (.73). That reading was selected in the step-wise regression seems to have merely a statistical relic because reading happened to correlate higher with cloze than vocabulary did, just barely, and the stepwise process can only enter one—the highest—variable at a time. That vocabulary was left out of the regression model by chance probably suggests that reading and vocabulary are not doing very different things in terms of predicting cloze test performance. Once reading had been entered, there was nothing unique that vocabulary could contribute. In fact, when grammar and vocabulary only are entered into the model, vocabulary is a significant contributor (sr = .25, significant at $p < .05$) to the cloze test performance. Therefore, we can conclude, that, yes, vocabulary is a significant contributor to cloze test performance, but its contribution is not unique. Reading and vocabulary seem to overlap.

Potential Implications

This research has implications for language testers, researchers and teachers or anyone else who would be interested in applying the cloze procedure to language assessment problems. First of all, it has been shown that the cloze procedure appears to be a fair measurement of language ability as opposed to nonverbal reasoning or "intelligence". Therefore, teachers and other test administrators who wish to assess language students using a cloze test may do so without fear that they are unintentionally measuring language independent human abilities (at least, not of the type included in this study). On the other hand, the precise application of the

cloze test to certain language skill areas still seems to be open to debate. In this research, the rationally deleted cloze test did not prove to be a strong measure of reading comprehension as its designers intended, although it does seem to be measuring reading ability to some degree. Test administrators should be cautious when applying the rationally deleted cloze procedure to reading comprehension practices. Reading does not appear to be the most important contributor to rationally deleted cloze test performance.

Limitations

The research presented here is not without its limitations and the results should only be interpreted to the extent that the sample here can be generalized to a second language population. Of obvious concern is the small sample size used in this study. Kline (1994) suggested an *N* size of 50 as an absolute minimum for factor analysis, however, a ratio of 20:1 cases to variables is usually desired for multiple regression analyses (Tabachnick & Fidell, 1989).

Furthermore, heterogeneous groups are desirable, yet the sample presented in this study may be somewhat homogeneous as compared to the entire population of ESL students. As compared to the entire population of ESL students, this sample is slightly homogeneous in terms of language proficiency as indicated by the narrow range of TOEFL scores (510-600), and they may be somewhat homogeneous in their general reasoning ability as well, all being college students.

Additionally, exploratory factor analysis in general is most successful when a wide range of variables are sampled. In this study, only 53% of the total variance in the factor analysis could be accounted for by the nine variables entered, which indicates that some factor has been unaccounted for. On the other hand, this study was not entirely exploratory, being based on the research and results of a few predecessors. Therefore, I had some precedence for choosing the

particular variables that I did. There is no clear reason that I can see, for example, why the cloze procedure, when administered in writing, would have any relationship with listening or speaking ability and because of that, tests of listening and speaking were not included in the design. Only those tests that were expected to contribute to cloze test performance were considered.

Finally, the extent to which the results presented here can be generalized to other cloze tests may be limited. According to Brown (2002), cloze tests based on different passages, and even based on deleting different words in the same passage, are not equivalent. Therefore, although grammar and reading were the only variables found to contribute to the present cloze test, it may not follow that the same would be true for other cloze tests. Those wishing to implement cloze tests for their own purposes are cautioned to consider the nature of the cloze procedure and their specific purposes for testing and to develop their own cloze test accordingly.

CONCLUSION

The present study was designed with two aims in mind: (a) to test the notion that the cloze procedure is related to the idea of closure in Gestalt psychology and (b) to investigate the degree to which specific language skills (grammar knowledge, reading ability, and vocabulary knowledge) contribute to cloze test performance for second language learners. The origin of the cloze procedure as explained in Taylor (1953) rests in the arms of Gestalt psychology where it was observed that humans vary in their ability to perceive an image that is not complete. That is, they vary in their ability to supply missing information from their minds. Taylor guessed that the same principle could be applied to language problems, that individuals would vary in their ability to supply words missing from a text by first perceiving (or understanding) the complete message

of the text. In fact, it is this argument that is utilized by researchers today, though in slightly different terms, when they talk about applying cloze tests to reading comprehension. Those who would use the cloze procedure to test reading ability suggest that readers must refer to a whole text in order to understand and fill in missing parts. However the results obtained in this study would seem to indicate otherwise.

The factor analysis indicated, first of all, that second language learners do not utilize the same psychological processes to complete cloze tests and closure tests. The multiple regression analysis supported this outcome. Therefore, we must seek another explanation to describe what happens when a second language learner completes a cloze test.

One possibility is that cloze test taking may be a skill unto itself, an idea suggested by Carroll as far back as 1959 and possibly supported by results obtained in Weaver and Kingston (1963). Although this possibility is only partially supported in the present study, since grammar and reading together accounted for 46% of the variance. Another possibility is that cultural knowledge, not included in this study, plays a greater role in cloze testing than was previously thought. Some evidence to this effect already exists in Sasaki (2000). She compared cloze test scores from a culturally loaded cloze test to scores obtained on a less loaded cloze and found that students performed better on the more culturally familiar cloze test.

Furthermore, one may conjecture that the biggest difference between the cloze test and the tests of closure used in this study was the amount of background knowledge required for completion. The incomplete images in the Gestalt Completion test, for example, consisted of rather simple items, like cats, trees, houses, etc. These are items that foreign language students could be expected to have experience with, and therefore, closure would be possible from the mind. In contrast, the cloze test, being in a foreign language, makes greater demands for cultural

knowledge, familiarity and understanding that the language students might have lacked. Considering that, it may be the case that all else being equal, closure and cloze processes are more similar than suggested here, but when cultural knowledge and familiarity of a language are taken away, cloze tests become a different sort of thing.

I might go even farther to suggest that the results obtained in Brown (2002) support this notion. As mentioned, Brown found that cloze test items operate differently for different individuals, some not functioning at all. Although he conjectured that language proficiency was the main construct contributing to these differences, cultural familiarity with the language could also be an explanation for the results he obtained. Of course, this is all speculation at this point, but a follow up study that administered the same set of tests to native speakers could shed some light on this issue.

REFERENCES

Alderson, J. C. (1978). A study of the cloze procedure with native and non-native speakers of English. Unpublished doctoral dissertation, University of Edinburgh.

Alderson, J. C. (1983). The cloze procedure and proficiency in English as a foreign language. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 205-217). Rowley, MA: Newbury House. (Reprinted from *TESOL Quarterly, 13*(2), 219-227, 1979)

Alderson, J. C. (2000). *Assessing Reading*. Cambridge University Press.

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

Bachman, L. F. (1982). The trait structure of cloze test scores. *TESOL Quarterly, 16*(1), 61-70.

Bachman, L. F. (1985). Performance on Cloze Tests with Fixed-Ratio and Rational Deletions. *TESOL Quarterly, 19*(3), 535-555.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading, 10*, 291-299.

Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement, 5*, 189-196.

Bormuth, J. R. (1969). Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly, 4*, 358-368.

Botzum, W. A. (1951). A factorial study of the reasoning and closure factors. *Psychometrika*, *16*(4), 361-386.

Brown, J. D. (1980). Relative merits of four methods for scoring cloze tests. *The Modern Language Journal, 64*(3), 311-317.

Brown, J. D. (1983a). A closer look at cloze: Validity and reliability. In J. W. Oller (Ed.), *Issues in language testing research* (pp. 237-250). Rowley, MA: Newbury House.

Brown, J. D. (1983b). A cloze is a cloze is a cloze? (ERIC Document Reproduction Service No. ED275145)

Brown, J. D. (2002). Do cloze tests work? Or, is it just an illusion? *Second Language Studies*, 21(1), 79-125.

Carroll, J. B.; and Others (1959). *An investigation of cloze items in the measurement of achievement in a foreign language*. Cambridge, MA: Laboratory for Research in Instruction, Harvard School of Education.

Chihara, T., Oller, J., Weaver, K., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language Learning*, *27*(1), 63-73.

Clark, M. (2002). *A Brief History of the ELI Placement Test*. Unpublished paper. University of Hawai‘i, Mānoa.

Clark, M. (2004). The development of an item bank for the English Language Institute Reading Comprehension Test. *Selected Papers from the Eighth College-wide Conference for Students in Languages, Linguistics, and Literature*. University of Hawai‘i at Mānoa: College of Languages, Linguistics, and Literature.

Cziko, G. (1978). Differences in first- and second-language reading: The use of syntactic, semantic and discourse constraints. *The Canadian Modern Language Review, 34*, 473-489.

Cziko, G. (1983). Another response to Shanahan, Kamil, and Tobin: Further reasons to keep the cloze case open. *Reading Research Quarterly, 18*(3), 361-365.

Dobrean, A., Raven, J., Comsa, M., Rusu, C., & Balazsi, R. (2005). Romanian standardization of Raven's Standard Progressive Matrices *Plus*. *WebPsychEmpiricist.* Retrieved October 29, 2005, from http://wpe.info/papers_table.html

Ekstrom, R. B., French, J. H., Harman, H. H., & Dermen, D. (1976). *Kit of Factor-Referenced Cognitive Tests*. Princeton, NJ: Educational Testing Services.

Flahive, D. E. (1980). Separating the g factor from reading comprehension. In J. W. Oller and K. Perkins (Eds.), *Research in Language Testing.* (pp. 34-46). Rowley, MA: Newbury House.

French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *Kit of Factor-Referenced Cognitive Tests*. Princeton, NJ: Educational Testing Services.

Gamaroff, R. (1998). The cloze test as a measure of language proficiency: A statistical analysis. *South African Journal of Linguistics, 16*(1), 7-15.

Green, D. R., & Tomlinson, M. (1983). The cloze procedure applied to a probability concepts test. *Journal of Research in Reading, 6*(2), 103-118.

Henk, W. (1982). A response to Shanahan, Kamil, and Tobin: The case is not yet clozed. *Reading Research Quarterly, 17*, 591-595.

Hinofotis, F. B. (1980a). Cloze as an alternative method of ESL placement and proficiency testing. In J. W. Oller, Jr. & K. Perkins (Eds.), *Research in Language Testing* (pp. 121-128). Rowley, MA: Newbury House.

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *TESOL Quarterly, 24*, 61-83.

Kline, P. (1994). *An easy guide to factor analysis*. London: Routledge.

Kohler, E. (1966). *An investigation of cloze scores in terms of selected cognitive variables*. Unpublished Doctoral Dissertation. Florida State University: Tallahassee.

Levenston, E. A., Nir, R., & Blum-Kulka, S. (1984). Discourse analysis and the testing of reading comprehension by cloze techniques. In A. K. Pugh & J. M. Ulign (Eds.), *Reading for professional purposes* (pp. 202-212). London: Heinemann.

McKenna, M. C., & Layton, K. (1990). Concurrent validity of cloze as a measure of inter-sentential comprehension. *Journal of Educational Psychology, 82*(2), 372-377.

Oller, J. W., Jr.; And Others (1972). Cloze Tests in English, Thai, and Vietnamese: Native and Non-Native Performance. *Language Learning, 2*(1), 1-15.

Oller, J. W., Jr. (1972). Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in English as a Second Language. *The Modern Language Journal, 56*(3), 151-158.

Oller, J. W., Jr. (1975). *Cloze, Discourse, and Approximations to English*. (ERIC Document Reproduction Service No. ED107144)

Oller, J. W., & Conrad, C. A. (1971). The cloze technique and ESL proficiency. *Language Learning*, *21*(2), 185-195.

Ohnmacht, F. W., Weaver, W. W., & Kohler, E. T. (1970). Cloze and closure: A factorial study. *Journal of Psychology, 74*, 205-217.

Pemberton, C. (1952). The closure factors related to other cognitive processes. *Psychometrika*, *17*(3), 267-288.

Rankin, E. F., & Culhane, J. W. (1969). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading, 13*, 193-198.

Ravin, J. (2000). The Raven's progressive matrices: Change and stability over culture and time. *Cognitive Psychology*, *41*, 1-48.

Sasaki, M. (1996). *Second language proficiency, foreign language aptitude, and intelligence: Quantitative and qualitative analyses*. New York: Peter Lang Pub.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing,17*(*1*), 85–114.

Schuhfried, G. (n.d.) *Raven´s Standard Progressive Matrices Plus (SPMPLS)*. Retrieved October 29, 2005 from http://www.schuhfried.at/eng/wts/spmpls.htm

Shanahan, T., & Kamil, M. (1982). The sensitivity of cloze to passage organization. In J. Niles & L. Harris (Eds.), *New Inquiries in Reading Research and Instruction: Thirty-first yearbook of the National Reading Conference* (pp. 204-208). Rochester, NY: National Reading Conference.

Shanahan, T., Kamil, M., & Tobin, A. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229-255.

Street, R. F. (1931). *A Gestalt completion test: A study of a cross section of intellect*. (Contributions to Education No. 481). New York: Teachers College, Columbia University.

Tabachnick, B. G., & Fidell, L. S. (1989). *Using multivariate statistics* (2nd ed.). New York: Harper Collins.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly, 30*, 414-433.

Taylor, W. (1956). Recent developments in the use of "Cloze Procedure." *Journalism Quarterly*, *33*(1), 42-48, 99.

Thurstone, L. L. (1938). *Primary Mental Abilities*. (Psychometric Monograph No. 1). Chicago: University of Chicago.

Thurstone, L. L. (1944). *A factorial study of perception*. Chicago, IL: University of Chicago.

Thurstone, L. L. (1949). *Mechanical aptitude III: Analysis of group tests*. (The Psychometric Laboratory No. 55). Chicago: University of Chicago.

Thurstone, L. L., & Jeffrey, T. E. (1996). *Closure Speed (Gestalt Completion): Information Guide*. Chicago: Pearson Reid London House.

Todd, S. (1995). *Coming out of the closet: Product and process in a cloze design.*. Unpublished paper. University of Hawai‘i at Mānoa.

Yamashita, J. (2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing, 20*(*3*), 267-293.

Weaver, W. & Kingston, A. (1963). A factor analysis of the cloze procedure and other measures of reading and language ability. *Journal of Communication, 13*, 252-261.

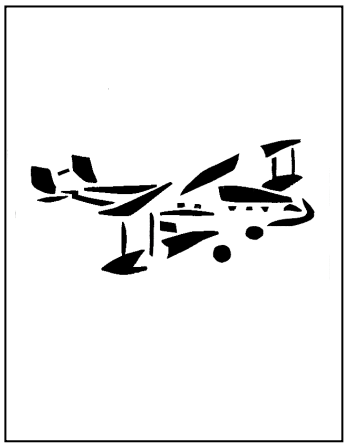
APPENDIX A

Gestalt Completion

- The test taker must identify the image represented by the partial picture. Example A is a man. Example B is a plane. Testee will be able to reply in any language. The examples demonstrate the task, but they represent rather easy items. The test items in the actual test get progressively more difficult.

A

B



Figures reproduced from Street (1931).

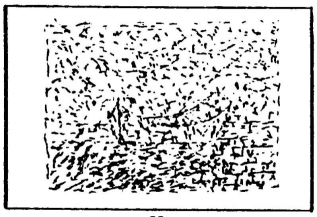
Snowy Pictures

- The test taker is asked to recognize "hard-to-see" objects which are obscured by snowy spatters. Sample item 1 is an anchor. Sample item 2 is a small boat on the water. For this study, respondents were allowed to use their native language to answer these items.

Sample Item 1:

1. anchor

Sample Item 2:



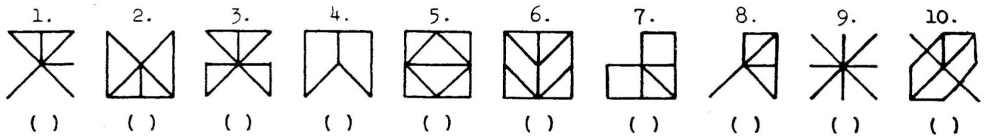
2. a small

Figures reproduced from Ekstrom et al. (1976)

Hidden Patterns

- The respondent is asked to search for the following pattern hidden among more complex patterns.

The respondent places an X in the space below each pattern in which the model appears and a 0 in the space below the pattern in which the model does not appear.



In this example, an X should appear below patterns 1, 2, 4, 8, and 10 and a 0 below 2, 5, 6, 7, and 9.

Figures reproduced from Ekstrom et al. (1976)

Copying

- Respondents are asked to copy a pattern onto a square of dots, testing their ability to keep a pattern in mind so that it can be quickly found in the dots.

Figures reproduced from Ekstrom et al. (1976)

APPENDIX B

EXAMPLE TEST ITEMS

Cloze[xv]

The elderly man was walking slowly down the street when he stepped on a piece of ice and fell _____. Luckily, he didn't hurt himself.

ACCEPTABLE: down, over, suddenly, accidentally

After trying to study most of the night, the young woman finally went to bed. She was so sleepy that she couldn't _____ in the morning, and she missed her test.

ACCEPTABLE: awaken, concentrate
NOT ACCEPTABLE: wake up (2 words)

NOTE: Spelling is not graded, but grammar (such as a past-tense "ed" or a plural "s") are graded.

Reading Comprehension Test

SECTION 1. In this section, there are 6 passages to read. Each passage is approximately half a page, and has 4-5 questions to answer. Your job is to read each passage and the questions that follow, then choose the best answer for the question (A, B, C, or D).

SECTION 2 (**vocabulary**): Given the word in the top row, choose the word or phrase below it that is closest in meaning (A, B, C, or D).

Here are some examples:

feverish	invisible	related to work
A. hungry	A. not seen	A. recreational
B. poor	B. not divided	B. occupational
C. hot	C. not allowed	C. sensational
D. evil	D. not wanted	D. rotational

In Example 1, the word nearest the meaning of the word feverish is the word hot. If this were a part of the test, you would mark the letter C on your answer sheet.

[xv] ELI Placement Test Examples (n.d.). Retrieved April 4, 2005 from http://www.hawaii.edu/eli/students/newstudents.html

Grammar Test

Eun Young has been working for three years when she decided to quit her job.
A B C D

The correct answer is A (that is, A is grammatically incorrect).

Raven's Progressive Matrices

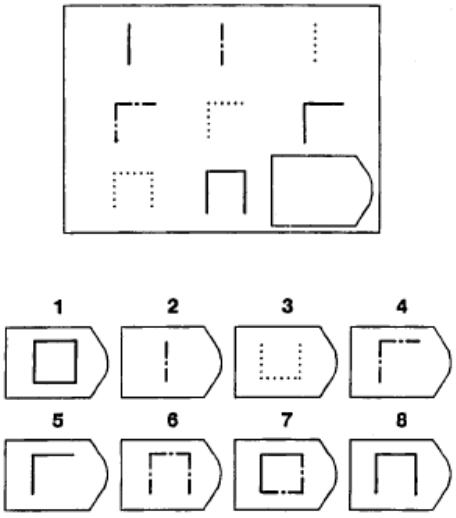


FIG. 1. Illustrative Progressive Matrices item. Respondents are asked to identify the piece required to complete the design from the options below. (The item shown here is not from the current range of tests.) [Reproduced from Raven, 2000]

APPENDIX C

Transformation of Gestalt Scores

The following transformation was performed on the raw Gestalt test scores.

gesttrans = SQRT(K-X)

where K is a constant and X is the Gestalt score. K was set to 19, one higher than the highest value in the data. The new transformed data was normal, but the sign of the new scores were all reversed because of the reflection about the mean. In order to keep the relationship positive, the data had to be *re*reflected. The following reflection formula was used.

gesttra2 = 5.3 - gesttran

5.3 is one higher than the highest score. The descriptive statistics of the newly transformed data (gesttra2) are shown in Table C1 below.

Table C1

Descriptive Statistics for "Gesttra2" – The Transformed Gestalt Scores

N	53
Mean	2.68
Std. Error of Mean	.12
Median	2.65
Mode	3.89
Std. Deviation	.86
Variance	.73
Skewness	-.11
Std. Error of Skewness	.33
Kurtosis	-.82
Std. Error of Kurtosis	.64
Range	3.36
Minimum	.94
Maximum	4.30

APPENDIX D

AGREEMENT TO PARTICIPATE IN RESEARCH
"Cloze Test Validity Study"

Investigator: Treela McKamey, M.A. Candidate
Faculty Supervisor: Prof. J.D. Brown
Department of Second Language Studies
570 Moore Hall, University of Hawai'i at Manoa
1890 East-West Road, Honolulu, HI 96822
(808) 956-8610

Purpose of this Research

The purpose of this study is to explore the relationship between seven different kinds of tests. The main goal is to discover whether the Cloze Test is a useful test for reading comprehension.

What You Will Be Expected to Do

If you participate in this research, you will be asked to do the following things:

1. Complete this form, giving me permission to access your ELI Placement Test scores.
2. Complete six tests -- four picture tests, a language test, and a computerized picture test.
    a. These tests will take a total of 80 to 120 minutes to complete.
    b. You will sit for all five tests in one session, move to another room, and take the final test.

Compensation

If you agree to participate in this study you will receive compensation of a $10 value.

You Have a Right to…

- **Confidentiality**

  During your participation in this study, YOUR PERSONAL INFORMATION, INCLUDING YOUR IDENTITY, WILL BE KEPT STRICTLY CONFIDENTIAL (SECRET). YOUR NAME WILL NEVER BE USED in any reference to this study. You will be assigned a number for record keeping purposes. At the end of the study, any record linking your name to a number will be destroyed and only the numbered data will be retained.

- **Ask Questions at Any Time**

  You may ask questions about the research at any time. Call the investigator at (808) 428-0208, or, if you prefer email, send an email message to treela@hawaii.edu. If the investigator cannot answer your questions, contact the faculty supervisor at (808) 956-2784.

- **Withdraw at Any Time**

  YOU MAY WITHDRAW FROM THE STUDY AT ANY TIME, and you may require that your data be destroyed without any consequences or loss of compensation.

Benefits

Although this study may not be of direct and immediate benefit to you, it is anticipated that the main benefits will be to future language learners like you. The results will help teachers and researchers understand how cloze tests work and how they can best be applied in language testing. This knowledge might help future teachers and administrators develop better tests.

Possible Risks

There are no known risks involved in this study.

Signature

I certify that I have read and understand the above, that I have been given satisfactory answers to any questions about the research, and that I have been advised that I am free to withdraw my consent and to discontinue participation in the research at any time, without any prejudice or loss of benefits or compensation.

I agree to be a part of this study with the understanding that such permission does not take away any of my rights, nor does it release the investigator or the institution (or any agent or employee thereof) from liability for negligence.

If I cannot obtain satisfactory answers to my questions, or have comments or complaints about my participation in this study, I may contact: Committee on Human Studies (CHS), University of Hawaii, 2540 Maile Way, Honolulu, HI 96822. Phone: (808) 956-5007.

_____ __/__/__

(print your name) (date)

_____

(signature)

CC: Signed copy to participant