TOWARD A PREDICTIVE MEASURE OF L2 PROFICIENCY: LINKING PROFICIENCY AND VOCABULARY IN SPANISH AS A FOREIGN LANGUAGE

by

Rebekah Hoy

Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Arts

in the

English

Program

YOUNGSTOWN STATE UNIVERSITY

August 2011

TOWARD A PREDICTIVE MEASURE OF L2 PROFICIENCY:
LINKING PROFICIENCY AND VOCABULARY IN SPANISH AS A FOREIGN LANGUAGE

Rebekah Hoy

Signature:

_____

Rebekah Hoy, Student Date

Approvals:

_____

Dr. Steven Brown, Thesis Advisor Date

_____

Dr. Kevin Ball, Committee Member Date

_____

Dr. Servio Becerra, Committee Member Date

_____

Peter J. Kasvinsky, Dean of Graduate Studies and Research Date

ACKNOWLEDGEMENTS

This project would not have been possible but for the tremendous outpouring of generosity on the part of many people. I would first like to thank the English department at Youngstown State University for a very rewarding two years of graduate study among a talented and supportive administration and faculty. Most importantly, I wish to thank Dr. Steven Brown, the chairperson and advisor for this project. Throughout all stages of this project, Dr. Brown has remained among the most accessible, approachable, and agreeable faculty members I have encountered. His remarkable patience and his insight into my work kept the gears of this process turning even in the many times I hesitated and stumbled. This thesis was greatly informed by his expertise and real-world experience with language teaching and learning, and I have been very fortunate to have had him as a teacher, an employer, and a mentor during my time at YSU.

I would also like to thank members of the Department of Foreign Languages for their interest in my study and helpful facilitation. Foremost, thanks to Dr. Servio Becerra for generously serving on my committee across departments; his input regarding the Spanish components of this project were invaluable, and provided me with crucial access to authentic language. Additionally, Dr. Becerra very helpfully aided in the logistics of the study administration, along with John Sandru, also of the Department of Foreign Languages. Thanks to the reception personnel in that department as well, for their welcoming spirit.

Next, many thanks to Dr. Kevin Ball, who provided extremely valuable suggestions on several drafts of the project, and whom I'd like to remind that literally everything is a process and therefore subject to revision—including post-it notes attached to drafts. Thanks also goes to Dr. G. Jay Kerns of the Department of Mathematics and Statistics for his help with the split-half reliability calculation, and to Ellen Wakefield of Maag Library for her patient assistance with source retrieval. Additionally, I am grateful to Monica Turenne for taking the time to offer important native speaker input during the early stages of test design.

Finally, I want to extend very special thanks to my mother and father for their constant love, support, and encouragement during the year this project was underway, and always. Dad, you even managed not to over-praise… Deepest thanks goes to Marc Turenne for his input and assistance in the data analysis ("GOBBLE"!), and his presence throughout my life—my earliest role model.

For my fiancé and partner in life, Daniel
Thank you for showing me that sometimes even words are inadequate and unnecessary.

ABSTRACT

Research in vocabulary size and depth in a foreign language has gained popularity since the 1980s, and efforts at developing reliable measures of FL vocabulary continue today as placement tests such as the Eurocentres vocabulary tests (Meara and Jones, 1990, 1998) remain subject to revision and improvement. In an attempt to investigate a possible relationship between vocabulary size and language ability, this study developed two measures: a cloze-passage test (recognized as a valid predictor of overall proficiency) and an adaptation of Meara and Buxton's (1987) Yes/No vocabulary test, which purports to estimate learners' vocabulary size. Scores on the two tests were correlated, yielding a weak positive correlation of .36; however, this finding was not statistically significant. Further testing with a larger sample is necessary for improving the behavior of the tests.

TABLE OF CONTENTS

INTRODUCTION: LANGUAGE ASSESSMENT

The past three decades have seen an abundance of work in the field of second language testing, particularly in English. Research in the development and use of tests designed to measure proficiency in the English language has lead to advances in test construction and a better understanding of both the potential and the limitations of what a language test can achieve, and how tests can impact not only the student, but also language education as a whole. Bachman and Palmer (1996) remark that language testing is needlessly shrouded in "mystique" and "misconceptions," often yielding negative affective consequences (p. 4). Certainly, the mythologizing of language testing as infallible or omniscient is beneficial to no one, least of all to those who are most impacted by the testing process: the students. Responsible test research, design, and administration to a great extent involve consideration of how the language test will affect a learner's life: educationally, professionally, economically, and personally. To that end, continuous reevaluation of the methods we employ in the construction and use of language assessment tools deserves a space of high disciplinary priority. Language testing is neither a magical art nor a guaranteed means of irrefutable or complete knowledge about a learner's ability; test construction requires conscientious scholarship and controlled application for continued refinement of assessment tools.

As stated, testing English as a second language (ESL) has been an academically and professionally rigorous field for decades. However, only since about the 1980s has vocabulary testing become a major area of interest for both theorists and pedagogues. One question that the following literature review seeks to explore is the degree of influence that learners' vocabulary ability might have on their total language competence. Untested intuition would suggest that vocabulary is necessarily the primary foundation upon which other language abilities are built.

Linking the significance of vocabulary skills to testing situations, Meara and Buxton (1987) argue that "all language tests are to some extent tests of vocabulary: without some knowledge of what the words appearing in a test mean, it is extremely difficult to perform at all, let alone well" (p. 142). Various aspects of the nature of vocabulary, its acquisition, and its assessment set it apart as a unique aspect of inquiry, and, despite having reached its zenith over twenty years ago, the topic deserves revisiting as a still-vital space of investigation. Although the study of ESL and EFL (English as a foreign language) testing has been extensively examined, other foreign language (FL) vocabulary research has been less prolific. The present study recognizes this gap and sets out to incorporate much of the established work on ESL vocabulary testing into a foreign-language vocabulary testing scenario. Meara (2005) acknowledges the preponderance of ESL vocabulary research, but also states his interest in other languages, specifically that of the increasingly important world language, Spanish (p. 271). Likewise, this thesis is specifically interested in the development and implementation of a test of the Spanish vocabularies of students enrolled in first-semester SFL (Spanish as a Foreign Language) courses. To this end, the study will probe the ways in which past research has spoken about the quantification of vocabulary, and what that work may suggest about testing the relationship between vocabulary size and proficiency.

Test Design

The process of designing effective vocabulary tests has been shown to be predicated heavily on construct validation, as put forward by Read and Chapelle (2001), who emphasize the importance of considering the test's purpose when designing and validating a vocabulary test. Vocabulary tests must present learners with tasks that are directly "relevant to the inferences to be made about their lexical ability" (p. 1) in order to be valid as indicators of vocabulary ability.

In other words, care must be exercised during test construction and implementation to isolate and control for what is being measured, and to ensure that vocabulary is in fact what is actually assessed in a given language test. The best test-designing practices are those which take into consideration every design decision before making any claims about the efficacy of a given test. Vocabulary test design has experienced much attention from the field of measurement theory, and what follows is a brief background on some general concepts, namely validity and reliability. The first issues to be discussed here are the various types of test validity. In its most basic form, validity refers to the degree to which a test measures what it sets out to measure, and is referred to as *construct validity*. "Construct," in recent years, has come to refer to the separate, largely abstract language abilities that tests attempt to measure: in the field of language testing, these constructs could be "reading ability," "writing ability," or, for the purposes of this discussion, "vocabulary knowledge" and "global proficiency." Subsumed under this broad definition of validity are two narrower categories which are more closely specified to particular aspects of internal validity, that is, content and criterion validity.

Content validity

As the name suggests, content validity refers to the degree to which a test's content reflects the construct it claims to measure. For example, a test of grammar must in fact test grammar. A better articulation of this idea is offered by Hughes (2003) "A test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which it is meant to be concerned" (p. 26). The "representative sample" element of Hughes' statement is important: even a test of grammar whose content includes only grammar items will not have a high degree of validity if the particular items tested are not reflective of the type of items that need to be assessed. Therefore, it is essential that *what* is to be

tested is set forth clearly before the design and/or administration of any assessment. The importance of content validity is directly related to the larger idea of construct validity, as articulated by Hughes (2003): "…the greater a test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure, i.e., to have construct validity" (p. 27).

It stands to reason, then, that in order to achieve high levels of content validity, the constructs to be measured must be clearly defined. Bachman and Palmer (1996, p. 117-120) identify two distinct approaches to construct definition, known as the *syllabus-based* and the *theory-based* approaches. Defining the construct in question based on the syllabus approach is most applicable to tests designed for assessment within the domain of the classroom. Theory-based construct definitions, on the other hand, are often more frequently generated for use in research capacities, and differ from syllabus-based definitions in that they are "based on a theoretical model of language ability rather than the contents of a language teaching syllabus" (p. 118). Syllabus-based construct definitions in language test design are likely to be implemented in "high stakes" testing situations, that is, situations in which the student's performance will influence major decisions such as university admission, course placement, and even employment. Conversely, tests designed and administered for research (laboratory) purposes can be thought of as generally low-stakes, since the outcomes of such tests will not affect students in any tangible way. The present thesis is an example of a theory-based study, in that it uses existing research to inform the construction of an artificial assessment scenario in order to explore the process of test construction and behavior, rather than to provide data about students' performance for the purpose of making decisions such as course grades or placement. For an extended discussion of theory-based construct definitions, see Bachman and Palmer (1996).

In further emphasis on the importance of test validation, Hughes stresses that the presence or absence of content validity can, and does, have a direct impact on the language classroom in the form of either positive or negative "backwash." If students continuously perform poorly on tests because the tests are not actually testing the skill under scrutiny, students may suffer negative affective reactions such as inhibited motivation or anxiety. Backwash, in general, refers to the effects that language tests have on the instructional setting; in other words, how does a certain test influence, for better or for worse, the process of teaching and learning outside of the laboratory, in the actual language classroom? Because backwash is largely a practical and a pedagogical concern, we will not deal extensively with it here. However, in "real-world" test design and administration, the possibilities of both positive and negative backwash must be considered for the sake of maximizing the efficacy and preserving the integrity of the learning environment.

Criterion-related validity

A second type of test validity is related to how closely test results resemble an external, independent measure of the construct in question, which serves as "the criterion measure against which the test is validated" (Hughes, 2003, p. 27). Examples of criterion measures, according to Hughes, include lengthier, more "established" tests, as well as teacher assessments of student performance or course outcome (grade). Integrated in this larger concept are yet two narrower types of criterion-related validity: concurrent and predictive. Concurrent validity can be measured when both the test under construction as well as the independent criterion assessment are administered at about the same time. When the results of both measures yield approximately the same results, the test itself can be said to have a high level of agreement with the independent criterion—and a high degree of concurrent validity, expressed mathematically as a *validity*

coefficient between 0 and 1. Measures of concurrent validity provide information about a test at the present time. On the other hand, ascertaining a test's *predictive* validity provides information with regards to potential future outcomes, such as the likelihood of a student's success in a given language course. Predictive validity, therefore, "concerns the degree to which a test can predict candidates' future performance" (Hughes, 2003, p. 27), and is appropriate to ascertain specifically in placement testing situations. The often lengthy process of validating a given test is necessary in order to obtain results that will inform its revision prior to implementation.

Reliability

The concept of reliability in language assessment is the idea that similar scores (or outcomes) should theoretically obtain regardless of when, where, and on whom the test is administered. Achieving a high degree of reliability is difficult due to external factors that influence test takers on any given occasion. Variables, including environmental and affective variables on any particular day, will influence a test's outcome to some extent, and this reality must be acknowledged and allowed for. As articulated by Hughes (2003):

> We know that the scores [will be] different if the test [is] administered on the previous or the following day. This is inevitable, and we must accept it. What we have to do is construct, administer and score tests in such a way that the scores actually obtained on a test on a particular occasion are *likely* to be very similar to those which would have been obtained if it had been administered to the same students with the same ability, but at a different time. The more similar the scores would have been, the more reliable the test is said to be. (p. 36, my emphasis)

Like validity, reliability can be quantified in terms of a coefficient, allowing for objective comparisons among different test formats. Perfect reliability registers at 1, and lesser degrees of reliability decrease incrementally to zero. What should be noted is that not all language tests will yield comparable reliability coefficients; in fact, typical reliability coefficients vary based on the

specific construct, or language ability, being measured. For instance, Lado (1961) claimed that a good reliability coefficient for vocabulary (and structure and reading) tests is somewhere in the range of .90 to .99, a significantly higher expectation than the .80-.89 range for listening tests and .70-.79 range for tests of speaking.

The manner in which the reliability coefficient is obtained for a given language test relies on the acquisition of two sets of scores for comparison. The *test-posttest* method involves administering the same test twice, with an amount of intervening time between each administration. Those scores are then correlated (compared for similarity) for the purpose of quantifying the degree of reliability. Second, the *alternate forms* method requires the administration of two different formats of a test, identical to one another except for superficial elements. Both tests must feature precisely the same content, but present that content slightly differently. If scores on both test forms are comparable, they can be said to be reliable. Finally, the *split-half* method of reliability measurement yields a coefficient of *internal consistency*. In this method, a completed test is divided into two equal parts, and scored twice: once for the first half and another for the second half. If the two scores are comparable, the test can be said to have a high level of internal consistency, and therefore reliability. Finally, scorer reliability is another important aspect of language test design, and speaks to the objectivity of the test. If a test's scorer reliability coefficient is 1, the test is demonstrated to be truly objective. Hughes (2003, pp. 46, 50) suggests that test-designers not only employ multiple, independent scorers, but also allow knowledgeable colleagues to attempt provision of alternative responses (different than those anticipated by the designers) that could be construed as correct in order to eradicate any ambiguity that might skew the test's outcome.

Varieties of language tests

Because language tests are typed by the sort of information they provide, they can be chosen to fit the needs of the specific purpose determined by the testing situation. We will consider briefly four types of language tests, keeping in mind the differences among them in terms of their contextual purposes. What follows is a short description of the purposes of proficiency, achievement, placement, and diagnostic language tests. In general, *proficiency tests* can be regarded as those assessments designed to measure speakers' overall ability in a second or foreign language. The assumption is that learners will demonstrate proficiency levels irrespective of any formal or informal language education they may have had in the past. Hughes (2003) points out that a test measuring independent proficiency in a language begs a definition of what "proficiency" entails, and submits: "In the case of some proficiency tests, 'proficient' means having sufficient command of the language for a particular purpose" (p. 11). Thus, it is worthwhile to carefully examine any test claiming to measure proficiency; the abilities measured (and the purposes for their use) must be clearly defined. All proficiency tests attempt to measure speakers' abilities based on an independently decided-upon set of standards with regard to any or all of the four (or five, when cultural competence is included) language skills. Additionally, "all proficiency tests have in common the fact that they are not based on courses that candidates may have previously taken" (p. 12). Hughes makes special reference to what he believes to be a harmful backwash effect characteristic of many proficiency tests; he argues that such tests have a tremendous role in dictating both the method and the content employed in the classroom. It would seem that a degree of danger lies in allowing standardized proficiency tests to stipulate what content is presented in the classroom, and which method is used to deliver that content, particularly in high-stakes situations.

Unlike proficiency tests, *achievement tests* are usually directly related to the instructional content of a language course in which the speaker is enrolled. Hughes (2003) names the purpose of achievement tests as "establish[ing] how successful individual students, or the courses themselves have been in achieving objectives" (p. 13). Achievement tests can take one of two forms: the *final* achievement test measures ultimate success at the end of a course, while the *progress* achievement test measures success during the course's progression. Hughes mentions some complicated questions that are actively debated with regard to exactly what elements of course content and/or objectives achievement tests are meant to assess, and also calls attention to the similarity, at times, between proficiency and final achievement tests. Indeed, "if a test is based on the objectives of a course, and these are equivalent to the language needs on which a proficiency test is based, there is no reason to expect a difference between the form and content of the two tests" (p. 14). Although the two tests may superficially appear to measure the same construct in the same way, the issue is in fact more complicated than what initially seems like simply different names for essentially the same type of test. Hughes insists on the separate nature of the achievement and proficiency test, citing reasons dealing mainly with course objectives, and goes on to describe other controversial elements with regard to achievement tests, concluding with the argument that "it is better to base the content of achievement tests on course objectives rather than on the detailed content of a course" (p. 15), suggesting a larger, more inclusive and long-range view of the testing process.

The third type of language test is the *diagnostic test*, which lives up to its name: diagnostic tests are designed to "diagnose" both the strong and the weak aspects of learners' ability. Diagnostics should also serve as means by which to "prescribe" any additional instruction that might be necessary. Hughes argues that proficiency tests accomplish essentially

the same goals as diagnostic tests and could be employed toward the same purposes; this argument likely stems from his perception that there are few well-designed diagnostic tests (p. 16). Finally, *placement tests* are used specifically for the purpose of assigning students to the level of language study appropriate to their level of ability. Educational institutions have a choice between constructing their own placement tests in-house or purchasing them from independent test designers. It would seem that the best language placement tests are those created in-house, with a specific student population in mind. Customized placement tests are perhaps more likely to address the specific needs and characteristics of the students, the language program, and the institution.

Vocabulary Assessment

In order to begin developing an understanding of vocabulary assessment, it is necessary to discuss some issues related to the construct of vocabulary itself. As will be explored in the present study, the notion of what it means to "know" a word complicates the process of test design, from the test's instructions themselves ("check every word that you know"), to how to best create scenarios in which learners can provide evidence of this knowledge. Does "knowledge" of a word refer to the ability to simply *recognize* it? Or does truly knowing a word require the ability to *use* it, and if so, must this use be receptive (as in listening and reading) or productive (as in speaking and writing), or both? Furthermore, does the somewhat reductive receptive/productive dichotomy truly aid us in describing the complex processes involved in language learning? Oller (e.g. 1983), advanced a theory of "expectancy grammar" which effectively closes the perceived gap between receptive and productive ability; this model will figure prominently here in a later discussion of global language proficiency.

Most of these questions represent fodder for other studies and cannot be discussed exhaustively here. However, "what it means to know a word" will be seen to present as a particularly salient problem during the execution of the Yes/No Vocabulary test in the present study, since "knowing" a word is not an all-or-nothing ability, but can be described in terms of a continuum ranging from sight recognition to fluent use. Richard's (1976) proposed definition of word knowledge is an often-cited conceptualization of the multidimensional nature of word knowledge. Richards suggested a conception of word knowledge that includes the following aspects of that knowledge: knowing the chances of encountering the word in print or in conversation; knowing the ways in which a word may or may not be used, given its function and the situation; knowing how the word behaves syntactically; knowing both the root and derivations of a word; knowing how the word interacts with other words in the given language; and knowing its basic semantic value (Richards, 1976, p. 83).

The above components indicate that the construct of vocabulary is not unidimensional, and suggests the presence of at least two rather distinct research interests in vocabulary assessment, namely the study of relative *depth* of word knowledge (e.g., Wesche and Paribakht, 1996) and the study of learners' vocabulary *breadth*, or size (e.g., Nation, 1990; Laufer and Nation, 1995, 1999). In an analysis of the distinctness between tests of vocabulary depth and breadth, Laufer et al. (2004) suggest that vocabulary size and strength are linked but distinct, because word recognition is consistently easier than word recall[1] (p. 223). They conclude, furthermore, that tests of size are in fact related to proficiency, and that "vocabulary size…may suffice as a surrogate measure of overall proficiency or as a predictor of overall performance" (p. 224). Since depth tests can assess only a relatively small number of words, it would seem that

[1] For a critical review of word recognition research, see Koda (1996).

they are better suited for specific, targeted investigations. The interests of the present study lie in exploring the relationship between vocabulary size and overall language proficiency.

Word frequency

Fundamental to attempting an assessment of vocabulary size is an understanding of word frequency. Frequency refers to the occurrence of a word relative to other words in a language, and is generally established by assembling large corpora of texts presumed to be representative of the frequency of words in a given language. High-frequency words are those that occur most often in a language, and tend to be shorter, whereas low-frequency words tend to be longer, and include specialized and academic jargon.

An important distinction that must inform the construction of vocabulary tests is that of *content* versus *function* words. Words that carry little semantic meaning, such as *a, the, and*, and other grammatical functors such as prepositions, are known as function words, and are among the most frequently occurring words. *Content* words, on the other hand, are substantive words (verbs, nouns, adverbs and adjectives) that carry semantic weight. Content words are generally regarded as the target items of vocabulary assessment. The debate surrounding whether function words are truly part of the vocabulary of the language or merely aspects of its grammar is an interesting avenue of inquiry for other studies.

Using and compiling frequency lists necessitates an awareness of the *type-token* ratio. A token is one individual occurrence of a type within a text and may occur in any form, whereas a type is the base, non-inflected, non-derived form of the word. Individual words (tokens) are counted on each occurrence, whereas types represent different words counted one time (Read, 2000, p.18). Inflectional endings obviously add massive amounts of words to a language's lexis. Using Read's example, this idea can be illustrated with the word *wait,* which also carries with it

the associated *waits, waited, and waiting*; similarly, the noun *society* is complicated by inflections that create the words *societies, society's* and *societies'"* (p. 18). Base and inflected forms of words are known collectively as *lemma*, which, for research purposes, simplifies the process of counting words in a text. By lemmatizing tokens, "inflected forms are counted as instances of the same lemma as the base form" (p. 18). In addition to inflectional endings, further complexity in word counting is represented by the numerous possible derivations stemming from base word forms. Read provides the examples *leaky, leakiness and leaker*, all derivatives of the base word *leak*. Although each of these words is different orthographically, they each share the common semantic characteristic "loss of fluid." Words such as these, which are closely related semantically, are known as *word families*.

The counting of single words, while carrying its own set of problems during the compilation of frequency lists and dictionaries, is also accompanied by the counting of what Read calls "larger lexical items" (p. 20). These multi-word items include phrasal verbs, compound nouns, and idioms, and are generally recognized as discrete units within the lexis, comprising what he terms the body of "prefabricated language" (p. 21). Notoriously difficult for second-language learners to acquire, these multi-word lexical units present learners the challenge of distinguishing between the meanings of not only the individual words, but also the meaning of the phrase as a whole, which is often rooted in cultural phenomena associated with the given language. In other words, and to use Read's example, knowing the meanings of the individual components *put, up,* and *with* does not necessarily indicate an understanding of the phrase *put up with*. Read points out two schools of thought with regard to the problem of multi-component lexical units in his reference to scholarship that has challenged Chomsky's (e.g. 1965) theory of Universal Grammar. Specifically, Pawly and Syder (1983) argue that fluency represents varying

levels of lexicalization of thousands of memorized sentences and phrases. This view of language (phrases and sentences) as "building blocks" (p. 208) is counter to Chomsky's emphasis on grammatical rules as mechanisms that make possible the acquisition of an infinite set of structures. In an effort to reconcile these two views, Sinclair (1991) uses corpus research to propose a marriage between the Chomskyan *open-choice principle* and the newer *idiom principle*, and suggests that "A language learner has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (p. 110).

Further insight into the question regarding the treatment of multi-word lexical items is Nattinger and DeCarrico's (1992) concept of the *lexical phrase,* which, like the label suggests, is "a group of words that looks like a grammatical structure but operates as a unit, with a particular function in spoken or written discourse" (Read 2000, p. 22). Nattinger and DeCarrico delineate four different categories of the lexical phrase: polywords, institutionalized expressions, phrasal constraints, and sentence builders. In general, these types of multi-word lexical items, or lexical phrases, occupy less space in vocabulary research and assessment than do single items.

The frequency dictionary and tests of vocabulary size

In his review of the published work on dictionary-based vocabulary size estimates, Nation (1993) indicts most of the work as having provided "misleading" estimates due to biased methodology (p. 27). Nation cites Edward Thorndike's (1924) review of nine studies estimating the vocabulary of school-age English speaking children as the earliest work to illuminate four of the fundamental problems plaguing previous vocabulary studies, all of which are bound up in sampling problems. First, Thorndike asserted that using a sample source (dictionary) that was too small would tend to underestimate the vocabulary size of the sample population. Secondly,

he found that the criteria for what would be included in a word family (i.e., base form only; base and derived forms; subsequent homograph entries) must be clearly stated by the researcher; failure to delineate precise guidelines for word inclusion would amount to "leaving the decisions to the dictionary makers" (Nation, 1993, p. 28), thereby resulting in inflated estimates of vocabulary size. Third, and, according to Nation, the most important principle to consider when using a dictionary to construct a lexical size test, is the necessity of using a sampling procedure that does not bias word selection toward high frequency (more frequently-occurring, or common) words. One such pitfall in sampling procedure included choosing the first word on a page, "regardless of whether it was the first full entry and whether it was a subsequent homograph" (p. 28). This procedure introduces bias due to the fact that the samples derived will contain a preponderance of high-frequency words, known to most or all of the people taking the test. Regarding such biasing sampling methodology, Nation states:

> This bias occurred simply because high-frequency words occupied more space per entry and had more entries than low-frequency words. The greater the size of the dictionary, the more the space given to high-frequency words, and thus the greater the overestimation of vocabulary size. (p. 28)

Although Thorndike himself offered several solutions to these biasing effects, including the numbering of each dictionary entry and choosing every *n*th word on every *m*th page, Nation remarks that Thorndike's important 1924 paper, because it was published not in an easily-accessible journal but in an obscure collection of working papers, went largely unread, and much of the subsequent work in this area failed to make use of the study's contributions. Indeed, according to Nation, the work that followed Thorndike's breakthrough continued to commit the same types of sampling errors that he decried even as late as the 1940's.

The fourth and last of Thorndike's recommendations was the use of an outside check of the representativeness of the compiled sample; in his case, this check was a word-frequency list called "The Teacher's Word Book." Checking a sample against a frequency list enables the researcher to ascertain whether a given sample provides adequate coverage of all frequency levels.

Corpus research and Spanish frequency dictionaries

While research on English corpora, frequency lists, and vocabulary coverage has proliferated since Thorndike's time, less work has been invested in the assembly of representative corpora in Spanish. Davies (2005) laments the fact that, prior to 2001, the only publically-available Spanish corpus suffered from several limitations: the corpus consisted of only one million words; it was comprised only of written texts, thereby neglecting the spoken component of the language; the texts used in creating the corpus contained no Modern Spanish; and no Central or South American source materials were included. Then, in 2001, a database with two large online corpora was made available by the Real Academia Española. These new corpora (one of which contains only Modern Spanish) are large and very well-representative; however, Davies observes a lingering flaw in the corpora with serious implications for the construction of reliable Spanish frequency dictionaries, namely, that neither corpus is lemmatized. Non-lemmatized corpora prevent the creation of accurate frequency dictionaries in that the individual word forms appearing in the list cannot be grouped into one lemma, or headword. Consequently, Spanish verb forms such as *digo, diremos, and dijeran* cannot appear under the headword *decir*, the infinitive form (Davies, 2006: 133).

Thus, the frequency dictionaries that have been produced in Spanish are tremendously problematic and unreliable. Of the extant Spanish frequency dictionaries (Buchanan, 1927;

Eaton, 1940; Rodríquez Bou, 1952; García Hoz, 1953; Juilland and Chang- Rodríguez, 1964; Alameda and Cuetos, 1995; and Sebastián, Carreiras, and Cuetos, 2000), none escaped the effects of the corpora's deficiencies (cited in Davies, 2006: 134). Chang- Rodríguez (1964) has generally been regarded as the most complete, but even this widely-used dictionary is nonetheless subject to the methodological problems inherent in the available corpora of the time. In a practical sense, the limited and problematic nature of the research on Spanish vocabulary coverage also engenders serious pedagogical concerns. Since flawed corpora result in poor frequency dictionaries, it follows that Spanish textbooks and other teaching materials developed with these dictionaries will fail to engage learners in texts and vocabulary that are representative of the Spanish likely to be encountered in the real world. Instruction materials, argues Davies (2006), should above all prepare students for authentic encounters with Spanish vocabulary: "The goal of a textbook should be to prepare students for the 'real world'…the goal should not be that of simply mastering the domains and language functions of 'textbook language'" (p. 133). In order to cultivate SFL students' vocabulary acquisition, it is essential that textbooks and other materials provide learners with the most accurate and comprehensive sample of the language as possible.

In 2002, and in what represented a major step forward for Spanish vocabulary coverage interests, the *Corpus del Español* became publically available.[2] In short, the new corpus resolved all of the fundamental flaws inherent in previous efforts: it contains an impressive 100 million words, 20 million of which originate from the 1900s and beyond; it represents sources from both Spain and Latin America; and it contains sources from both the spoken and written (and fiction and non-fiction) registers (Davies 2006: 134). Davies names the corpus's lemmatization as the primary feature that makes possible the construction of a frequency

[2] www.corpusdelespanol.org

dictionary superior to all previous attempts. It was based on this corpus achievement that Davies (2005) produced *A Frequency Dictionary of Spanish: Core Vocabulary for Learners*. One facet of the superiority of this new frequency dictionary is evident in a comparison of its word inclusions with those of Chang- Rodríguez (1964). Unlike Chang-Rodriguez, Davies' dictionary includes high-frequency and extremely useful words including *película, control, television, oportunidad,* and *cruel* (p. 135). Other helpful features include information about entries' frequency, range, and the register in which the word is most commonly found; the dictionary also provides an English gloss and a sample sentence for each entry. The construction of valid and reliable vocabulary tests, particularly tests of vocabulary size, rests fundamentally on the accuracy and representativeness of corpora and the frequency dictionaries based thereon. Davies' (2005) *Frequency Dictionary of Spanish* is currently the superior resource for teachers and researchers of Spanish vocabulary, and was used in the construction of the lexical size test in the present study.

Conclusion: Toward an Examination of the Cloze Procedure and the Yes/No Vocabulary Test

The preceding general introduction has established the foundation on which we can now begin a more focused discussion about the two types of language tests under investigation in the current study. The impetus for this study was the broad empirical question regarding the relationship between vocabulary size and overall proficiency in a foreign language, and further, whether vocabulary size can be used as an predictive index of general language proficiency. To that end, it was necessary to design two tests: an efficient and reliable estimate of learners' lexical breadth, and a valid means of proficiency assessment. The two tests chosen for these purposes were the Yes/No Vocabulary test (e.g. Meara and Buxton, 1987), and a cloze passage, or gap-filling test. An important distinction to highlight here is the difference between *discrete*

point versus *integrative* assessment. Hughes (2003) defines discrete point tests as "testing one element at a time, item by item" (p. 19), thereby isolating one specific element of language (and therefore, ability). Discrete point tests assess learners' capacity to cope with a linguistic element in a largely de-contextualized situation. Examples of discrete point tests include multiple choice, matching, true-false test formats, and, under specific investigation in this study, the Yes/No Vocabulary Checklist test. Integrative testing, on the other hand, "requires the candidate to combine many language elements in the completion of a task," (p.19) and may include composition, dictation, and, as in the present study, completion of a cloze passage. The following literature review examines the Yes/No Vocabulary Checklist test (specifically Meara and Buxton, 1987, and later replications) and the cloze test for the purposes of understanding the basic methodology employed in their construction and what the empirical results suggest about their validity, reliability, and administration. Finally, we turn to a discussion of the Yes/No test and cloze test designed and implemented in the current study, with a view toward future improvements of the tests' correlative behavior.

CHAPTER 2: CLOZE PROCEDURE

Introduction

The body of literature on cloze procedure typically attributes the origin of cloze to Wilson Taylor (1953), who introduced the technique as a way of "approximating the readability [or difficulty] level of written materials in the field of journalism" (Brown, 1993). In simplest terms, the cloze procedure is a technique for developing tests in which a passage of text is subject to a deletion of certain words which the student is required to supply, thereby "restoring" the text. A minimum of fifty total deletions is generally recommended for sufficient reliability, although some studies have created tests with as few as thirty-five items. Taylor's initial use of cloze interpreted the mean cloze score of a large group of native English speakers as an index of the level of difficulty of a given passage of text. According to several retrospective discussions of the history of cloze procedure, Taylor borrowed the term "cloze" from Gestaltist psychological theory, which postulates that within the nature of human beings is the urge and tendency to fill in gaps in discernable patterns (e.g. Oller and Conrad, 1971). As articulated by Stansfield (1980):

> Gestaltists believe that learning follows a sequence through which one first understands the whole of broader issues, and then grasps the individual details. Similarly, the cloze procedure requires the student to perceive the whole by filling in the missing words as if they were not missing at all. (p. 30)

This gestaltist concept of processing input from the top-down, that is, digesting the input on a global level first, will recur later in this paper as an argument in favor of cloze as a measure of overall proficiency in response to the central debate regarding cloze. Appraised once as "nothing less than a stroke of raw genius" (Oller, 1973), Taylor's cloze procedure would

eventually become the subject of a substantial amount of research. Early work with cloze with regard to text readability would evolve into investigations of language assessment, and cloze became the subject of interrogation by theoretical and applied linguists. The question "what does cloze actually measure?" would lead to an effort to investigate the behavior and characteristics of cloze tests in various settings. Perhaps the fundamental research question is that of the cloze test's potential as a measure of either overall proficiency or discrete, lower-order skills. That is, much of the underlying debate regarding the cloze procedure is about the type of language ability it is capable of assessing: integrative, holistic language proficiency, discrete skills, higher-order processing, or lower-order abilities. This question will be shown to emerge from several decades of cloze procedure research.

The cloze procedure would eventually pique the interest of the second-language research community. Oller (1973) claims that Taylor himself put forward the idea of cloze use in second-language settings as early as 1956. As referenced by Brown (1993), early attempts to utilize cloze procedure in second language research included Allen (1968), who advocated cloze as both a teaching and a testing device in ESL situations; Estrada (1969), whose study represents an early use of cloze for the purposes of measuring varying degrees of sentence difficulty for Navajo children, and Crawford (1970) who used cloze technique in order to gauge English reading comprehension levels in Spanish-speaking American children. Devons (1969) used a multiple-choice cloze format for testing English reading proficiency among groups in Israel, and Bowen (1969) compared varying levels of English proficiency in different Ethiopian groups using cloze tests designed in both English and Amharic. Darnell (1968), using native English speaker norms as scoring criteria, tested cloze as an ESL proficiency measure and achieved high correlations with the TOEFL (Test of English as a Foreign Language). Panopoulos (1966),

Spolsky (1969) and Kaplan and Jones (1970) met with high correlations between cloze and tests of listening proficiency (cited in Brown, 1993).

However, not until Oller (1979) was cloze procedure first said to be valid as a test of second language *global proficiency*. Oller would later argue that scores on cloze tests could unequivocally be interpreted as indication of a "general language proficiency factor" (p. 3), or *expectancy grammar,* a pragmatic theory of mental processing that will be discussed in this paper in terms of its role in the global-local proficiency debate. Since then, subsequent studies have served to challenge and complicate the assertion that cloze reliably serves as a valid measure of overall proficiency. For instance, Alderson (1979) and Klein Braley (1981) reported a closer relationship between cloze and discrete, localized skills, (namely vocabulary and grammar, an instance in which the two are differentiated) rather than integrative proficiency. The debate regarding whether cloze indeed measures global, discourse-level textual comprehension or more localized ability remains one of the contested and still largely unresolved issues associated with cloze procedure. Embedded in this problem are other, necessarily related issues, such as: cloze procedure's psychometric behavior; its optimum format (traditional random-word deletion, multiple-choice, C-Test, etc.); the most effective method of deletion (rational, fixed ratio, etc.); the most appropriate scoring method (exact answer, any semantically acceptable answer, clozentropy); the effects of difficulty of closure; and with what external criteria cloze best correlates. This review of the literature on cloze procedure will highlight these and other controversies, with the aim of moving toward an understanding of how best to utilize cloze tests in the specific context of foreign language vocabulary research.

As mentioned in Chapter 1, Meara and Buxton (1987) assert that "All language tests are to some extent tests of vocabulary" (p. 142). They would also argue that, by the time of their

seminal introduction of the Yes/No vocabulary technique, the "neglect" once experienced by vocabulary acquisition as a specific aspect of language learning had come to an end. However, because much research on cloze procedure has been devoted to correlations between cloze performance and scores on standardized criteria such as the TOEFL exam and tests of language abilities such as reading and dictation, the specific relationship between SL/FL vocabulary and cloze performance is arguably still underexplored. This suggestion can be defended by examining vocabulary's general status in the second- and foreign-language fields, as articulated by Lee (2008): "Vocabulary (knowledge and use) has not been labeled as a skill; thus vocabulary instruction has not been the subject of as much research as instruction of the four skills" (p. 645). Indeed, vocabulary in cloze research has generally been subsumed into theories of "integrated" reading and writing (Nation 2001), and the category of grammar, and has not occupied its own arena of investigation. The present literature review and study attempt to address this gap in the literature with a view toward examining how a cloze test correlates with a criterion designed explicitly to measure vocabulary size.

Perhaps the only consensus with regard to the cloze procedure is that its behavior has been historically inconsistent. Researchers have warned against an overly-optimistic adoption of cloze without careful consideration of its design and cautious interpretation of its results. In fact, Alderson (1979) asserts that cloze "is in fact merely a technique for producing tests, like any other technique, for example the multiple-choice technique, and is not an automatically valid procedure" (p. 226). Brown (1993) would later support this argument, reporting highly inconsistent levels of reliability and validity; reliability estimates, he reports, have ranged from .31 to .96, and validity measurements have demonstrated correlation coefficients of .43 to .91.[3] Thus, the value of cloze as a test of language ability (whether global, local, or otherwise)

[3] For additional references on reports of reliability and validity estimates in cloze testing, see Brown (1993, p. 94).

depends largely on the quality of its design, and the resultant levels of reliability and validity. These psychometric properties are strongly affected by variables such as text difficulty (e.g. Abraham & Chapelle, 1992) and scoring procedure and deletion frequency (Alderson, 1979). Another criticism of cloze is the ambiguity of what exactly it measures in any given situation. Abraham and Chapelle censure those who would use cloze as a proficiency measure without knowing with certainty what precise abilities are required for its completion. This review of cloze research begins with an overview of the types of cloze procedure that have been developed since its emergence as a language measure, and the various research contexts in which the different types have appeared.

Since the advent of cloze procedure in the early 1950s, several permutations have been developed from the original test format. Research has shown repeatedly that these tests are not created equal, nor do they all measure the same language constructs or demonstrate the same psychometric property levels. Types of cloze procedures now include not only the traditional cloze passage, but also the C-test (in which individual letters from words are deleted), the cloze elide test (in which subjects are themselves asked to delete superfluous words), and the open-ended multiple choice format. One important theoretical principle that has emerged from the body of work on various forms of cloze is that not every cloze test is appropriate in every research or instructional context. Furthermore, correlations of cloze tests with various language abilities vary widely depending on the peculiarities of a given text as well as researcher intercession and manipulation.

Cloze Text Selection and Researcher Manipulation

The text selection process involved in constructing a cloze passage is one example of researcher intercession that has important implications for the behavior and performance of

cloze. If it is assumed that different cloze tests necessarily measure different abilities, it is reasonable to attribute significance to the selection of the passage to be used for cloze construction. The text selection process typically involves choosing an authentic passage judged on the appropriateness of its difficulty level (readability) for the test group, as ascertained by either an instructor's knowledge of her student population. Other, more formal means include the Flesch-Kincaid readability test, a formula for ranking the ease or difficulty of a text in English from 0 (extremely difficult) to 100 (extremely easy). Readability scores around 60 are generally taken as representing texts in ordinary, non-academic English. Perhaps it is noteworthy that articles from *Scientific American* were chosen as texts for at least three classic studies in an ESL context (e.g. Irvine, Atai, and Oller, 1974; Bachman, 1985; Abraham and Chapelle, 1992).

It is worthwhile to look at two studies lying on either extreme on the continuum of researcher-imposed variables in terms of text selection. Kobayashi (2002) is an example of a study in which maximum researcher intervention was employed in the selection and construction of cloze passages. The results of the pilot study inspired deliberate manipulation to control for text length and difficulty in the following ways: topics were strategically chosen so as to be moderately familiar; both of the texts selected were used to produce four different texts with different "rhetorical organization[s]"; and the length and difficulty levels of each text were subject to homogenization (p. 574). Kobayashi's (2002) study, therefore, can be seen as an examination of cloze procedure founded upon calculated manipulation of variables. On the other end of this spectrum is the cloze procedure developed by Brown (1993), termed the "natural cloze test." Brown defines natural cloze procedures as those tests "developed without intercession based on the test developer's knowledge and intuitions about passage difficulty,

suitable topics, etc. (i.e., the criteria which are often used to select a cloze passage appropriate for a particular group of students)" (p. 93). The natural cloze experiment was an answer to what Brown considered the "inconsistent overall picture" (p. 94) of what an organic cloze test actually does without being subject to external manipulation. This study was an attempt to move toward a *totally random* method of text selection, and according to Brown, was the only one of its kind at that time. The intent, according to the researcher, was to merely *observe* the behavior of the cloze test in operation, rather than to influence it in any way—it was an attempt to detect naturally-occurring data patterns. Indeed, his process of materials selection does appear to achieve the completely random aim: passages were randomly selected from all of the adult books from a public library; a page was randomly selected from each randomly selected book; fifty passages were randomly assembled for modification into cloze tests; every 12th word was deleted, resulting in fewer instances of intercession and more intact text between blanks as compared to a test with more frequent deletions. Because it represented the method that would introduce the least amount of outside manipulation, the exact scoring method was used. Brown reports a significant amount of variation in the tests of his 50 "natural" cloze tests, and inconsistent levels of validity, which suggest that "a cloze test [unmanipulated by researchers] is not necessarily and automatically a sound overall ESL/EFL proficiency measure" (p. 110), and that, in order for sound cloze tests to be constructed, a degree of researcher intercession is required. Brown (1993) suggests that, although the majority of the natural cloze tests produced in this study performed poorly, perhaps a larger random sample of cloze tests would yield results of greater statistical and psychometric integrity. Each action by the researcher, therefore, can be interpreted as creating a slightly different test version.

Deletion rate and method

A defining difference between various types of cloze tests is the manner in which the text is distorted, or "mutilated," making deletion rates among the most discussed aspects of cloze design. Alderson (1979) argued that "it is misleading to ignore deletion rate differences to arrive at a composite score for any test" (p. 219). His study's major finding was the importance of the chosen deletion rate; in fact, manipulating the deletion rate seemed to produce altogether different tests which "unpredictably" measured different ability constructs (p. 225). Generally regarded as the most widely-used type of cloze test, the *fixed-ratio* cloze was designed to elicit a regular sample of various types of words occurring in a given test; in this format, test designers choose an interval of deletion (every 7^{th}, 11^{th}, or 12^{th} word, etc., is deleted), and blanks of uniform length are provided, in which the test-taker supplies the missing word to achieve "closure." True random deletion, on the other hand, while seldom encountered in the research, is a method in which deletion rate is based on an in-text percentage of word categories.

Because of his thorough investigation of cloze in a series of studies throughout the 1970s, it is worthwhile to discuss in some detail the theory that underscores John Oller's approach to the interpretation of cloze procedure. Oller (1973) considers this "pragmatic grammar of expectancy" at length; it is this theoretical schema that would crystallize from much of his work (1972-1979) on cloze. Oller (1972) conceives of an expectancy grammar as one that describes "the capacity to anticipate elements in sequence," a construct to which he ascribes enormous consequence, even to the extent that it is "the foundation of all language skills" (p. 151). From the expectancy perspective, the question of what exactly cloze tests measure can be interrogated and described by way of a view of language in which the student continuously posits, tests, and adjusts hypotheses about what she anticipates in the sequential discourse:

> It is interesting to note that the process of taking a cloze test involves more than "passive" reading. By sampling the information that is present the subject formulates hypotheses, or expectations, about information that is to follow. By sampling subsequent sequences, he either confirms or disconfirms these expectations. If the expectations are disconfirmed they must be revised and new hypotheses must be formed. (p. 114)

An expectancy angle would extend this framework and argue its applicability also to visual processing, speaking, and writing. It is through this notion of continual anticipation, hypothesis and adjustment that Oller is able to claim that language abilities such as speaking and writing are "not actually distinct processes" (p. 114), but rather that both abilities incorporate receptive *and* productive elements of language use. The model is succinctly described as the integration and negotiation between what can be expected to occur between *both* "atomic and molecular units of discourse" (Oller, 1972, p. 151). Similarly, cloze tests, in their demand that test-takers continually conjecture (sample input) and test (offer output), represent a measure of *global* language proficiency. The erosion of the rift generally constructed between receptive and productive ability leads to a situation in which cloze can convincingly be argued to serve as a measure of integrative language proficiency. To be successful on a cloze test, a learner must be able to maneuver with some precision across the continuum of productive and receptive abilities. Oller suggests that the successful negotiation of the cloze procedure indicates a certain category of competence, namely a *general language proficiency factor* that activates and depends on "memory constraints" (116), an avenue that is beyond the scope of the present paper.

Carter's (1998) discussion of cloze procedure speaks of expectancy grammar as a way to conceptualize the idea of the "redundancy" of a given passage's message, and explains that "the redundancy of the message in most normal naturally occurring texts should be such as to allow

accurate insertions into the blanks" (p. 228). Perhaps best discussed in terms of the reading process, expectancy grammar assumes a two-way interaction between reader and text, for which the use of both "co-textual as well as contextual clues" is activated, thereby permitting Oller's perception of the cloze test as a measure of holistic language proficiency. Foley's (1983) review and analysis of cloze testing also incorporates a substantial discussion of expectancy grammar, placing particular emphasis on the idea that cloze data can be interpreted as indicating one of three levels: "frustrational, instructional, and independent." Complicating the discussion, Foley's review refers to cloze as merely a "blunt instrument" (p. 67) capable only of supplying imprecise estimates of criterion validity. Although cloze procedure would excite far less optimism for Foley and others, earlier studies (e.g. Oller and Conrad 1971), nonetheless continue to fuel the debate regarding the cloze test's potential as an indicator of global proficiency.

In addition to operationalizing the theory of reading that drove his conception of what cloze measures, Oller (1975), in response to an investigation by Carroll (1972) of discourse constraints on cloze procedure, employed a *cut-and-scramble* manipulation before deleting every 7th word in tests of both sequential and scrambled texts.[4] At least for native English speakers, closure of the scrambled text was significantly more difficult than that of the sequential texts, indicating a level of involvement of the context surrounding cloze blanks. This observation would bring into question Carroll's (1972) supposition regarding the dominance of "local redundancy," that is, intra-sentential context, in cloze testing: "Cloze scores are probably more dependent on detection of grammatical than of semantic clues" (p. 189). In other words, Oller's hypothesis championed the case for cloze as a measure of more macro-level ability, since

[4]For more on specific scrambling techniques see Markham's (1985:424-425) brief review.

semantic clues are argued to access global knowledge. Supporting this position in a later investigation of context were Chihara, Oller, Weaver, and Chavez-Oller (1977), who substantiated Oller's (1975) report that non-sequential texts tend to yield far inferior performance when compared with sequential fixed-deletion rate texts for both native and nonnative speakers of English. They interpreted this as evidence that "the cloze procedure is sensitive to discourse constraints ranging across sentences" (p. 68), bolstering the global-proficiency perspective. Admittedly, in a later study of "intersentential sensitivity," Markham (1985) would derive findings challenging the previously-mentioned studies. In what he interpreted as contrary evidence with regard to cloze procedure as a valid test of global L2 proficiency, Markham employed a rational deletion method in which "only content-word deletions (substantives, verbs, and modifiers) were scored. Deletions that involved function words (determiners, conjunctions, prepositions, and interjections) were not included" (p. 426). Based on the study's results, Markham found no evidence that content words contributed any more significantly to restoration across sentence boundaries; the lack of evidence results in his lack of support of rational deletion procedures. He concludes that, at least when reading comprehension is taken as a marker of global proficiency, cloze tests do not represent a valid measure of said proficiency.

Some of the early second-language cloze research advocated the use of a fixed-ratio deletion method, usually based on arguments for ease of test construction and high correlations with standardized criteria. Oller and Conrad (1971) made a case in support of fixed-ratio deletion based on their observations that other types of deletion methods will favor certain grammatical categories and will produce more difficult tests. Arguing for fixed-ratio (rather than truly random) deletion, they argue: "This mechanical method of selecting blanks to be filled in by the student can, in the long run, be expected to reflect the frequency of occurrence of

grammatical and lexical forms in the languages tested" (p. 187). Adequate textual coverage of words of varying frequencies is arguably an important element in cloze tests correlated with an external vocabulary measure. Oller (1973) discussed the issue of deletion rate with regard to the difficulty level of the resultant test, positing that deletions occurring more frequently than every fifth, sixth, or seventh word create overly difficult tests with less discriminatory capability. However, he goes on to offer a position which suggests that cloze is resilient to changes in difficulty level: "we might expect to get pretty much the same information out of cloze tests of quite different difficulty levels" (p. 110).

Finally, in the *rational* deletion method, items are chosen for omission on the basis of a specific theoretical objective—i.e., to isolate for investigation certain specific grammatical (e.g. prepositions) or discourse-level (content) constructs. The rational cloze, in other words, derives its deletions based on a pre-determined condition. An example of an investigation imposing the rational condition on cloze design is Bachman's (1985) comparison of performances on fixed-ratio versus rational-deletion cloze tests; in this study, the rationally-deleted words were selected based on the range of context required for closure. This study follows Bachman's (1982) earlier work in which three classifications of rational deletion types were shown to point to three differentiable language constructs: syntactic, cohesive, and strategic. Employing this same classification schema and arguing for the superiority of rational cloze, Bachman (1985) justifies his position on deletion rate as follows: "By far the majority of the studies have distorted a text according to a procedure which is generally based on the linear arrangement of words and thereby virtually ignored hierarchical, structural and semantic relationships" (p. 537). The problem with this method, he argues, is that it rests on an unsound assumption regarding word redundancy, namely, that it distributes evenly through a given text. Bachman postulates three

inaccuracies inherent in this assumption, best described by what the study itself demonstrated: some deletions cause greater loss of meaning than others (i.e., content words versus function words); words may function at different and multiple structural levels; and the proportion of words functioning at different levels may be unequal. When deletion was rationally chosen based on the range of context required for closure, that is, by "discourse hierarchy," the fixed-ratio cloze tended to be more difficult than the rational, a phenomenon Bachman claims can only be attributed to the specific words and their contexts, since both cloze forms contained the same deletion ratio. That is, as the level of context needed for closure becomes greater, the difficulty of the test also increases. He interprets these results as an indication that rational deletion techniques are advantageous in that they aid in the development of cloze tests that can validly measure specific language constructs. Building on this work, Farhady and Keramati (1996) would later confirm and reassert Bachman's findings in a study implementing a slightly different rational deletion strategy; rather than focusing on the amount of context as such, their "text-driven method" determined deletion rate rationally based on the number of linguistic and discourse structures of a passage of text. Notably, this study would also corroborate the notion that greater context promotes ease of closure. They correlated a standard fixed-ratio cloze (with every seventh word deleted) with eight different cloze tests whose deletions were chosen based on the linguistic nature of the text itself. The four deletion criteria were: the sentence (above clause level); dependent and independent clauses (at clause level); phrases (below clause level); and around the clause ("cohesive ties") (p. 194).

The rational deletion method chosen in this study by Farhady and Keramati (1996) seemed to reveal a degree of agency in the passage itself; in other words, it is possible that features of the text itself may demand one or another type of deletion. Farhady and Keramati put

forward: “Each text may lend itself to a particular deletion rate which in turn will influence the number of deletions” (p. 199), and assert that trial and error should not be the sole determiner of the type of deletion rate to be employed. They call for a principled rationale for deletion method; in fact, they insist that this justification should be the “fundamental question” (p. 193) operating during cloze procedure design. This study also calls into question the reliability of the fixed-ratio deletion method, suggesting that the lengthier the passage, and thus the more blanks needed for closure, the more reliable (as well as difficult) the test will be. The researchers accept the paradox inherent in this principle; indeed, it follows logically that increased items will translate to increased reliability levels—the counterintuitive element is that this more reliable test will also be more difficult.

A more recent investigation by Kobayashi (2002) “revisits” the question of deletion method indirectly in its focus on scoring methods, an issue that this review will later discuss in some depth. Not only does the examination substantiate previous work advocating the use of a rational deletion technique, but it also reveals much about the effects of *item* characteristics. Kobayashi set out to determine what effect the *type* of deleted word had on cloze performance (with an eye specifically to interrogating cloze correlation with reading proficiency) and examined the effects of five characteristics of deleted items: frequency of the word; its status as content versus function; its part of speech; the number of occurrences of the word in the text; and the available range of alternate answers (due to the employment of the acceptable-answer scoring method). The findings showed that, at least for the Japanese population tested, “relative pronouns, pronouns, and articles proved to be the most difficult types of function words” (p. 577). Since English articles are known to be difficult for Japanese speakers to acquire, this finding suggests an interesting research question: to what degree does the nature of the L1

influence performance on a cloze test in a foreign language? This is a question that recurs in numerous language-test design situations, and will resurface in the second chapter of the present review in discussions of checklist-type vocabulary tests. The extant research on this issue is scant, and the question represents an avenue of investigation outside the scope of the current paper, but Kobayashi's findings suggest that, at least in that context, the learners' L1 represented an important factor in the outcome of cloze performance. Other findings included easier restoration of more frequently occurring words and increased validity as a measure of reading comprehension due to the emphasis on meaning rather than linguistic accuracy. The most important factor impacting performance in the pilot study was items' status as content words or function words. Chapelle and Abraham (1990) make the case that rational cloze tests allow the test designer control over the *types* of deleted words, and thus over the specific language abilities to be measured: "Because items are at the root of cloze performance, it has been suggested that the cloze procedure can be improved by selecting explicitly the words to be deleted, thus creating a rational cloze" (p. 124). Essentially, the use of the rational cloze procedure operates with the assumption that the test designer can deliberately choose items that will address the measurement of specific language traits; this ability to isolate the construct under investigation seems to be the prevailing argument for the use of rational versus fixed-ratio (sometimes known as "pseudo-random") cloze tests.

In their investigation of the "meaning" of cloze scores (as the problem is commonly referred to in the literature) based on an "item difficulty perspective," Abraham and Chapelle (1992) argue that contextual factors directly affect the difficulty, and thus the scores, on cloze tests. They constructed three types of cloze tests: a fixed-ratio cloze with every 11th word deleted following two intact sentences at the beginning (it is customary to allow one or two intact

sentences at the beginning and end), and a rational cloze in which the deleted words were chosen based on their having clearly identifiable contextual clues, demanding the use of information outside the mutilated sentence. The third cloze type was a multiple-choice cloze with the same deleted words as the rational format, with the exception that students were instructed to choose the appropriate word from among several incorrect words (distractors). Each cloze passage contained 35 blanks, or items. The results showed that the fixed-ratio cloze, which included 20 content and 15 function words, was more difficult than the rational cloze which was comprised of only 13 content words and 22 function words. The researchers interpreted the meaning of the fixed-ratio scores as students' ability to retrieve content words from long-term memory or to find them present elsewhere in the passage itself. However, amount of context necessary for restoration, which was specifically investigated for its effects on difficulty, was not shown to be a significant factor; therefore it is logical to surmise that, at least in this study, the fixed ratio cloze could not be seen as a valid measure of students' ability to use context clues. Incidentally, the fact that context was not required in this fixed-ratio situation lends support to the use of fixed-ratio deletion in the measurement of discrete language abilities such as grammar. On the other hand, the rational cloze *was* affected by context levels; Abraham and Chapelle took this result to be an indication of students' ability to utilize context clues during the cloze task. The finding exemplifies the previous assertion that different cloze formats address different language traits and must be chosen with a specific purpose in mind. The multiple-choice cloze test in this study revealed almost nothing, thus, it is suggested that this type of cloze should not be considered a valid measure of the same abilities as the "gap-filling cloze," a distinction in nomenclature that would be reinforced in Alderson's (2000) discussion of reading assessment. Alderson's insistence on the differentiation between the purportedly different "cloze" and "gap-

filling" tests is sensible given the many test formats that fall into the general category of cloze, including the C-Test (a complicated test design that will be explored here in depth) and the cloze elide, which *reverses* the Gestaltist principles of the original cloze by requiring students to choose and eliminate the superfluous words added by the test designer. Unlike the traditional cloze task which demands learners' ability to achieve closure, the cloze elide presents the task of "whittling away" unnecessary words introduced into the original text (for more on cloze elide, see Alderson, 2000).

To reiterate, traditional cloze tests are affected by manipulating variables such as deletion rate, scoring method, and text type. Contributing to the large volume of cloze research is yet another vicissitude in cloze research trends: the C-Test cloze variation. The following section will explore the behavior of this alternative cloze format which is just as contested and perhaps also equally inconclusive as that of the traditional cloze. The C-Test should, nevertheless, be noted for its potential for use in the specific investigation of SL and FL vocabulary, providing its development is supported by a research context allowing for the time and resources necessary for the more labor-intensive process of C-test construction.

The C-test

A variation of the cloze procedure known as the C-Test was developed in an attempt to address many of the widely-discussed problems perceived to be ineradicable in the traditional cloze. The C-Test differs from conventional cloze procedures in that it deletes the second half of every second word, presenting learners the task of restoring not only individual words, but also overall meaning. In their survey of the research on C-Tests, Klein-Braley and Raatz (1984) present a lengthy list of the problems the C-Test was meant to solve. In problematizing the cloze procedure, which they charge is often automatically regarded as a "universal panacea," they set a

background against which C-tests seem to emerge as the answer to all of the alleged flaws associated with the cloze procedure (p. 135). However, what is essential to remember is the need to identify with certainty *what* exactly a language test is intended to measure, and what it is shown to measure, before any assumptions about its validity can be made.

Chief among Klein-Braley and Raatz's (1984) complaints regarding traditional cloze behavior is the fundamental idea that cloze tests fail to accomplish what they were originally designed to do, that is, to produce a random sample of textual elements in order to demonstrate the difficulty or readability of a text. Furthermore, they reiterate the objection often leveled against cloze regarding the uncertainty of what language constructs cloze in fact measures. They remind us that manipulation of any variable in cloze construction (i.e., deletion rate, scoring, etc.) will produce different, and not necessarily valid or reliable, tests. Additionally, and of interest to the present study, the researchers name cloze passage selection as one of the "technical problems [that] cause headaches," (p. 135), referring to the challenge of selecting an appropriately suitable and difficult text for the group of learners involved. And, notably, they point out the following troubling contradiction: native speakers, they claim, seldom achieve perfect scores on cloze tests; at the same time, however, cloze tests in which the acceptable-scoring method is applied rely on native speaker assessment of acceptability, leading to issues with inter-rater reliability.

The C-test was designed to address the problem associated with a possible bias resulting from the content of the cloze passage by presenting learners with multiple text sources, usually five or six. Klein-Braley and Raatz (1984) identify the deletion method employed in the C-test as "the rule of 2," in which "beginning in the second sentence the second half of every second word is deleted until the required number of mutilations is reached. The text then continues to a

'natural break'" (p. 136). Essentially, the researchers argue that the ideal vision of a test of reduced redundancy can be realized in the C-test because the format satisfies the following conditions: the use of several different texts theoretically decreases or altogether eliminates content bias; the test contains 100 or more deletions; native speakers have demonstrated nearly perfect performance; the text deletions provide a highly representative sample; the test must be scored with the exact method; and the test has demonstrated validity and high reliability (p. 136).

Klein-Braley and Raatz (1984) reported that in pilot explorations of C-test behavior, two traditional cloze problems were rectified: it was demonstrable that the tests produced a random sampling of text elements (specifically parts of speech), and that native speakers achieved practically perfect scores. Because the external validation criteria used in early investigations of the C-test included teacher ratings, the researchers were obliged to address the difficulties associated with inter-rater reliability. They dismissed this issue, countering that "their pragmatic validity in the context of the school system… is a fact of life" (p. 136). This appeal to the practical reality of language testing is incongruent with the researchers' later admonitions that the C-tests "have, or should have, no place in the ongoing teaching process" (p. 144) due to their required levels of difficulty. Because the C-test is a norm-referenced test, that is, that target groups should be expected to achieve scores of only 50%, they argue that the difficulty level should increase incrementally across text passages, culminating in a "very difficult" final text. It would appear from their discussion that the most promising characteristic of the C-test is its ability to discriminate among subjects, and that the resultant rankings agree strongly with teacher judgments.

Klein Braley and Raatz's (1984) review concludes by summarily stating the "the C-Test does everything that the cloze test promised" (p. 145). However, although they report

satisfaction with C-test performance in all of the research with which they have been involved, they nonetheless warn against the automatic adoption of the test as reliable and valid, recalling similar warnings regarding the traditional cloze. Later work with C-tests would scrutinize many of the first findings and offer a slightly more complicated interpretation of its merits. Chapelle and Abraham (1990), although admitting that the C-test improves the cloze procedure psychometrically, suggest that the ubiquitous problem of identifying the elements of language that it tests is no clearer than it is in the traditional cloze procedure. At the time of their study, the researchers assert that "C-test research ha[d] failed to clarify evidence for the specific language traits that it measures" (p. 126).

Although it was first proposed as a measure of global competence, Klein-Braley (1985) found that semantic and textual proficiency seemed to subordinate to grammatical competence. The study aimed to compare the effects of various types of cloze tests, and in doing so, reported a finding that appears counterintuitive to Klein-Braley and Raatz's (1984) previous interpretation of the C-test as a measure of grammatical rather than semantic competence: "The C-test, *correlating most strongly with the vocabulary test*, produced, on average, the highest correlations with the language tests. Why did this apparently more grammatically-based test correlate so well with written text-based tests—even better than the fixed-ratio cloze?" (p. 140, my emphasis). Indeed, high correlations were certainly obtained with the vocabulary test (r = .862). However, because the aim of this study was to investigate the possibility of systematically comparing the relationships between test types rather than to explore specifically how C-tests behave when correlated with vocabulary measures, it is necessary to discuss a later study which does address the question regarding whether C-tests are valid for the purposes of measuring FL vocabulary.

Chapelle and Abraham (1990) discuss the possibilities with regard to "how researchers can bring to bear essentials of measurement theory on L2 research by weighing validity justifications pertaining to the use of the C-test method for vocabulary assessment in L2 research" (p. 157). Interrogating the C-test in consideration of its possible merits and drawbacks for L2 vocabulary research specifically, they rationalize the isolation of vocabulary as its own construct, and discuss its measurability by means of the C-test. Two justifications have been proposed in defense of the C-test as a vocabulary measure, namely, that C-tests are contextualized, and that they have obtained correlation with other language tests (Singleton and Little 1991, p. 67). These questions, along with the health of the test's construct validity, were investigated in relation to the C-test and vocabulary research of the time, and yielded hopeful results.

Chapelle (1994) operated with the new, unidimensional definition of validity proposed by Messick (1989), where validity is "the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores" (p. 13). Chapelle's inquiry into the validity of the C-test in the context of vocabulary assessment was informed by an important assumption regarding the nature of vocabulary ability, in that she defined it as "a capacity for language use in context" (p. 163). This perspective, according to Chapelle, derives from communicative language theory, in which an "interactionalist" model imagines ability as both linguistic knowledge and the capacity for contextualized use. Although an interactionalist view would reject the notion of quantifying learners' vocabulary in an absolute sense, in say, the manner of Nation and Meara, it has been previously suggested that vocabulary *size* can in fact play a contributing role in the scores obtained in a C-test (Klein-Braley, 1985). In the end, however, the analysis could not offer any

firm claims regarding the C-test as valid for vocabulary research; indeed, Chapelle (1994) concedes that she "found motivation both for and against use of the C-test in L2 vocabulary research" (p. 183). Perhaps the only conclusive directive to emerge from this analysis was that, in order for the C-test to be used validly as a vocabulary measure, rational deletion of content words must be elected over the typical fixed-ratio deletion of the second half of every other word. This only serves to confirm the superiority of the rational deletion technique that has been demonstrated with regards to cloze testing in general.

Singleton and Little's (1991) argument for the C-test as indicative of the condition of a learner's interlanguage (the intermediate state of an individual's language ability during the acquisition process) complicates our understanding of the global/local debate with regard to cloze testing and the relative merits of the C-test over the traditional cloze. Furthermore, Singleton and Little (1991) have contended that in a C-test the "command of L2 syntax can be engaged only to a very limited extent and its most lexical dimension…The principal challenge set by the [C-test] is very obviously lexical in nature" (p. 5). If this is the case, it would seem that the C-test, rather than operating as a global indicator, might be put to better use as a vocabulary assessment. Overall, C-tests have performed as promising measures of both "lower-level" (localized) ability and "higher level" (global) competence. However, Read (2000) concurs with the general consensus that, although empirical evidence seems to support Singleton and Little's (1991) claim that the C-test is a useful test of vocabulary knowledge, the research has yet to conclusively define the role of vocabulary in C-test performance (p. 113).

C-tests seem to merit further investigation with regard to their usefulness as vocabulary measures. Unfortunately, for the purposes of the present thesis, the C-test is impractical due to the complexity of its construction, i.e., the selection of at least five passages of text and an

analysis of their relative difficulty levels and suitability for the target population. The issue of text selection is complicated on its own in terms of selecting even one passage for construction of a traditional cloze test, indicating that C-test construction requires more time and expertise than the present study is able to offer. Furthermore, the time necessary for the administration of a test composed of five passages certainly exceeded the time allotted for the administration of the present study.

Scoring the cloze test

Scoring method, like deletion rate, has also come under scrutiny for its presumed influence on cloze performance and outcome. The question of cloze scoring relates to the problem of deciding which responses are to be accepted as correct for the purposes of quantifying data. Research in cloze procedure scoring has investigated a wide range of different scoring techniques: that in which only the deleted word occurring in the original text is acceptable (exact answer method); that in which any word that accomplishes the same meaning is correct (semantically acceptable-word answer); that in which any word in the same form class is acceptable; and the method in which any grammatically acceptable word is correct (see especially Alderson 1979 for more on the latter two). Alternatively, the method known as *clozentropy* would later be developed; this technique "logarithmically weights" satisfactory responses based on the frequency of their occurrence in a pretest administered to native speakers (Brown, 1993). The complexity of this scoring procedure was clarified by Alderson (1979) who describes clozentropy as a technique "which gives weighted credit for responses which are the same as responses given by a criterion group (usually native speakers of the language)" (p. 219). Despite some appealing aspects of this scoring procedure, it is perhaps logistically the most difficult to execute, and therefore will not be explored in this review in any depth. Furthermore,

the multiple-choice method (distinguished as closed-ended rather than open-ended), having been shown in many applications to produce the easiest tests, will not be considered at length. Rather, we will focus on a discussion of the two most widely-investigated scoring methods, as described in the following sections.

The thrust of the debate regarding scoring methods seems to be the competition between the *exact-word* and *acceptable-word* methods. Early manipulation of scoring methods in cloze tests of native English speakers' reading proficiency seemed to point to exact scoring (in which items are counted as correct only when the response is the same word deleted from the original text) as the optimum method, due to its seeming simplicity and reliability (see Oller, Atai, and Irvine, 1974). Oller and Conrad (1971) suggested that other scoring methods yielded approximately similar, though not superior, performance discrimination. However, in a later study, Oller (1971) would alter his position with regard to the issue of scoring; this investigation would serve to demonstrate superior differentiation between ESL speakers of varying proficiency when "any contextually acceptable word is counted as correct" (Oller and Conrad 1971, p. 191). In a different discussion, Oller (1973**)** suggests that tests scored by the acceptable-word method should undergo scoring by native speakers of the target language, or at least by highly proficient L2 speakers. Kobayashi (2002) arrived at a compilation of acceptable answers by consulting the opinions of six highly-literate native speakers of English.

In perhaps the first study to examine the relative outcomes of the four major scoring methods (exact answer, semantically acceptable answer, clozentropy, and multiple-choice) without the use of an external criterion measure, Brown (1980) hypothesized, among other things, a lack of difference in levels of reliability, validity, and difficulty index. However, the results in fact demonstrated clear differences between all four scoring methods in terms of

reliability and difficulty. The only variable that appeared unaffected by scoring method was validity. An analysis of the relative reliability levels across all four scoring methods demonstrated the any-semantically-acceptable method as having the highest reliability coefficient. In terms of difficulty, the exact scoring method was predictably the most difficult, with only 30% of subjects correctly closing 100% of the blanks. Brown interprets this result as indicative that exact scoring produces a test that is too difficult to yield high levels of reliability. Conversely, the level of test difficulty when the multiple-choice scoring technique is employed is suggested to be underdemanding; Brown reports a very high percentage (64%) of subjects responding correctly to the average item (see e.g. Oller, 1972, discussed below, for more on scoring and cloze difficulty). These and other results led Brown (1980) to assert that the any-acceptable response method of scoring is superior overall. His position regarding this method's superiority, particularly over the exact method, is expressed with conviction in his pronouncement that: "There is something inherently repugnant about counting an answer wrong, which is actually correct, simply because the author of the original passage did not choose to use that word; yet, precisely that often happens when the EX [exact] method is used" (p. 316). Oller raised the issue of fairness with respect to the exact-answer method, supporting his argument with the idea that guessing the exact word deleted from a passage is "not necessarily a language skill in the ordinary sense of the term" (p. 152). Although Brown qualifies his previous categorical statement by adding that the purpose and conditions of the test should ultimately be considered when choosing a scoring method, his use of the word "repugnant" strikes this author as representative of rather an extreme position. It could possibly be argued that the exact scoring method may be useful for studies designed to measure subjects' knowledge of words deliberately

chosen for one or another theoretical justification; in such instances, cloze procedures should probably consider correct only exact responses.

Oller (1972) set out to specifically investigate levels of textual difficulty as variables when various scoring methods were employed. He categorized the study's scoring methods as: M1, in which only exact restorations are counted; M2, allowing for both exact and "any other contextually acceptable responses," and M3-M5, each of which weighted and differentiated different types of responses (i.e., restoration of the original word; responses not the original word but entirely acceptable; responses violating discourse constraints; responses violating local constraints; and entirely incorrect or unanswered items). Three cloze tests were constructed to reflect three levels of difficulty: beginning, intermediate and advanced. Subjects also took the ESL Placement Examination from the University of California at Los Angeles which served as the criterion measure; this test consisted of vocabulary, grammar, reading, and dictation. Results revealed that M2 yielded significantly superior correlation coefficients than M1, with negligible correlation obtaining in M3-M5. For the two more difficult cloze forms, M2-M5 outperformed M1 in all cases but two. Oller concluded that the acceptable-word method used in this study was clearly superior to an exact-word method in terms of item discrimination and validity, regardless of test difficulty. Unlike Alderson's later findings, (1979), Oller (1972) found that correlations of cloze with a synonym-matching vocabulary test from an ESL proficiency measure were low when controlled for the integrative skills of grammar, dictation and reading. Oller suggests, therefore, that cloze tests, because they correlate best with tests that demand higher-order language skills, measure overall integrated proficiency rather than discrete-point skills.

Alderson (1979), in attempting to methodologically examine the effects of different scoring techniques and their respective influence on the relationship of the cloze tests and

various EFL measures, administered cloze tests and observed the effects of the following five scoring methods: exact word only; any semantically acceptable word; any word with an identical form class as the deleted word; any word with the same grammatical function; and any word that was grammatically correct, irrespective of class, function, or semantic value (p. 221). Alderson concluded that the cloze format employing the SEMAC (any semantically acceptable) method of scoring yielded the best correlations with external proficiency measures (notably dictation), and was specifically superior to the exact method, preempting and substantiating Brown's later (1980) findings. On the other hand, while Brown found no significant variance in validity across different scoring methods, Alderson's results pointed to the semantically-acceptable scoring method as producing the most valid test, at least in the domain of EFL testing. Of additional interest to the present study is the fact that Alderson identified vocabulary as one of the language skills most closely related to the cloze procedure, along with grammar and reading comprehension.

In an effort to investigate the interaction between isolated item characteristics and ESL cloze performance (particularly reading comprehension), Kobayashi (2002) employed the following three scoring methods: exact word; both semantically and syntactically acceptable word; and the semantically acceptable but syntactically unacceptable word. Among the item characteristics analyzed in this study were a word's status as either content or function; the part of speech; the frequency of the word; the number of times the word occurred in the text; the subjects' "knowledge base," or level of proficiency; and the number of alternate answers deemed acceptable by a panel of educated native Japanese speakers. Both of the "acceptable" word scoring methods yielded higher scores overall across all item characteristics. Results obtained through the use of the semantically (only) acceptable word yielded only slightly higher scores

than the semantically-and-syntactically acceptable words, but these higher scores occurred consistently, which has implications for differentiation between *types* of acceptable-word scoring methods. Additionally, Kobayashi found higher, although not remarkably higher, levels of reliability when either of the acceptable-word scoring methods was employed. Of note, according to Kobayashi, was the observed difference in reliability in relation to an item's role as content or function word with regard to the employed scoring method. It is reported that the exact-scoring method contributed to lower reliability estimates in the case of content word blanks, and that, conversely, reliability jumped "dramatically" when acceptable-word scoring was utilized. What is interesting is that "the values were even higher than those of function words in many of the texts" (p. 575). These findings carry implications regarding the effects of scoring methods with regard to a word's syntactic role, and Kobayashi calls for further research on this perceived pattern. Lastly, in addition to fairly high overall correlations with the proficiency measure, the study revealed slightly higher correlations when the acceptable-word scoring method was implemented; correlations were higher yet when the scoring allowed only for a semantically-acceptable response. This finding reinforces the recurrent idea that cloze procedures may measure similar constructs as proficiency tests. Individual scoring methods correlated very highly among themselves. Of the 24 correlations, and at a high .88, only one correlation dipped below .90. Although Kobayashi warns that these results should be interpreted cautiously due to some of the study's limitations (including a small number of test items), she also acknowledges the "complex relationship between cloze item characteristics and scoring methods" (p. 582). What this study does show unequivocally is that cloze procedure is affected by the choice of scoring method, and some may interpret this particular study as evidence

suggesting the superiority of the acceptable-word method (whether semantically-syntactically or semantically-only).

Performance discrimination in native and nonnative speakers

This review has made brief reference to the problem associated with native speaker performance on cloze tests. As has been mentioned, several studies to date (i.e., Klein-Braley and Raatz, 1984) have shown that native speakers seldom achieve perfect scores on cloze tests, and offer the C-test as a remedy for this particular issue of reliability. What follows is a discussion of research that speaks to the differences between native and nonnative speaker cloze performance, and why, if the cloze test can be claimed to measure overall language proficiency, it must be able to sensitively discriminate between the performance of native and nonnative speakers.

In their early attempt to ascertain the extent of the cloze procedure's ability to discriminate between varying levels of ESL proficiency, Oller and Conrad (1971) also examined their native-speaking control group along a continuum of beginning, intermediate, and advanced ESL students. They found that "Differentiation of levels of proficiency among the ESL groups seems adequate, but ENL [native English speaking] freshmen are not significantly distinct from advanced ESL students though they are significantly inferior to ENL graduate students" (p. 183). The researchers' interpretation of this study's results points to the interesting possibility that ESL students might be disadvantaged in comparison with native speakers for the very fact that they *are* studying English in a second-language context: they suggest that more advanced ESL students would benefit from more content-based instruction (like the curriculum to which native speakers are exposed), which would eliminate what Oller and Conrad refer to as the emphasis on unnecessary "habits or intuitions" that are characteristic of designated ESL classes (p. 191).

Indeed, in their study, some advanced ESL learners produced better scores than ENL college freshmen in an exact-scoring procedure. Alderson (1980) would later interpret these results as evidence that the cloze "discriminated falsely among native speakers, [while] failing to discriminate where it should—between native and nonnative speakers" (p. 62).

Lado (1986) contested an aspect of Oller and Conrad's (1971) above findings regarding the cloze test's potential as a viable tool for placement purposes, and therefore its ability to reveal individual differences. To this end, he replicated the study in order to support Carroll et al.'s (1959) hypothesis that cloze tests are "inadequate" for matters of "diagnosis and placement" (p. 131). Lado's findings would indeed uphold Carroll et al.'s (1959) hypothesis in an item-analysis procedure. Moreover, in an assertion regarding cloze as a valid test of higher-order, discourse-level skills, Lado states:

> From a psycholinguistic performance point of view, it is not surprising that the cloze does not encourage high level thinking since it requires that the subject use the context to search for specific words missing in the text… That there should be significant correlations between cloze and other types of language tests has no more significance than the fact that cloze has a substantial language component even though the process of using it is backwards from normal communicative use. (p. 136)

Lado's observation here that the use of context is necessarily part of the cloze task goes without saying. However, it is not as clear whether the component of "searching" the text is present, or even if it indicates local ability. Lado (1986) goes on to reference a "mundane" technical flaw in Oller and Conrad's (1971) original study, namely three errors in the counting off of deletions. Because of previous evidence that has suggested the important influence that deletion starting points have on the outcome of cloze tests, Lado insinuates that the three deletion frequency errors could have influenced the results of the study. It is beyond the scope of the present paper to engage with Lado (1986) and Oller and Conrad (1971) in a discussion of cloze

procedure and individual differences. For interesting discussions on individual learner differences and proficiency-level discrimination among nonnative speakers, see respectively Stansfield and Hansen (1983) on the individual variables of field-dependence/independence, and Yamashita's (2003) report on a think-aloud cloze that differentiated well between skilled and less skilled EFL students, thereby supporting the argument for cloze as global proficiency measure.

Alderson (1980) examined the differences between native and nonnative cloze performance with respect to the local-global proficiency debate (what Alderson dichotomizes as lower-order/higher-order skills). He scrutinized the assumption that proficiency tests (such as cloze is suggested to be) measure some factor or another that is inherent to all native speakers, and this work warrants a thorough examination for what it suggests about cloze as an indicator of overall language proficiency. Immediately striking is Alderson's tightening of the definition of "cloze tests." He proposes that only the pseudo-random cloze procedure should be designated as a proper cloze; other *rational* procedures, in which deletion rate is determined on the basis of a theoretical hypothesis relating to specific text or language characteristics, he suggests be referred to as "gap-filling tests" (pp. 59-60). This clear distinction is a response to a general consensus that different deletion rates produce different tests, measuring different constructs.

The study begins by stating that linguistic proficiency has been commonly understood as a language element that only native speakers possess, and that standardized proficiency tests such as the TOEFL are often designed to measure just that: the extent to which the examinee's language processing and use approximates that of a native speaker. It follows therefore, that tests of proficiency should reveal the differences between native and nonnative speakers and between nonnative speakers of different proficiency levels (both in L1 and L2), but should not

discriminate significantly between different native speakers. To restate, for a cloze procedure to be said to be a valid measure of *overall proficiency*, it must, by definition, produce tests on which "native speakers will perform uniformly well," and by which "nonnative speakers will be clearly distinguished from native speakers" (p. 60). Three different cloze tests were produced from three texts of pre-established levels of difficulty: low, medium and high. Each test experienced use with four deletion rates: every 6^{th}, 8^{th}, 10^{th}, and 12^{th} word deletion rates were imposed. As discussed in the scoring section of the present paper, Alderson employed 5 scoring techniques on these three tests, and it was expected *a priori* that native speakers would perform perfectly in tests where any grammatically-acceptable response was permitted, and in tests where any acceptable form class word (with accompanying grammatical functionality) was accepted (p. 63). At an item level, "the effect of the changes in deletion frequency was similar for native and nonnative speakers" (p. 64), whereas at the task level, changes in deletion rate were found to produce differentiation effects that were irregular at best.

Besides reiterating the supposition that varying the scoring method will vary the cloze test that is derived, Alderson's findings with respect to the effects of scoring show surprisingly little native-nonnative performance discrimination. Of particular importance is the observation that native speakers, while not only failing to attain perfect scores (even when the test becomes easier through more lenient scoring), struggled more with *semantically*-loaded items than with grammatical restorations. Although some minor differences were found between native and nonnative speaker cloze performance, Alderson did not interpret these as robust evidence that superior native-nonnative speaker discrimination is a behavior of the cloze procedure. He does, however, state that the cloze *task* is essentially the same for both native and nonnative speakers. Alderson's findings led him to suggest that perhaps using native-speaker norms as external

validation represents flawed methodology: “attempts to use native speakers as criteria for nonnative speakers, as in clozentropy or much criterion-referenced testing, are misguided. If native speakers vary in their ability to do cloze tests, then the value of using them as criteria is doubtful” (75). Furthermore, he calls into question the practice of validating proficiency tests with nonnative speakers on the grounds that native speakers will perform perfectly, since the results of his study indeed demonstrated otherwise. Based on the few studies examined here, it would seem that the cloze procedure has demonstrated an unsatisfactory capacity to discriminate between native and nonnative speakers, thereby violating the definition of how proficiency tests should behave. This simply represents further complicating evidence with regard to the highly contested, often inconsistent behavior of cloze tests.

Cloze procedure and reading comprehension

An aspect of cloze procedure as yet not specifically discussed in this review is that of its capacity as an assessment tool for reading in a second language. The applicability of cloze procedure to the field of reading assessment is demonstrated in much of the literature reviewed in the current paper up to this point. The argument that cloze procedure may be an “ideal” method for assessing reading (Alderson, 2000, p. 207) can be supported by the fact that the original purpose of the cloze technique (Taylor, 1953), was to assess the difficulty or “readability” of journalistic texts. We have already established that one of the most apparent characteristics of the cloze procedure is the ambiguity of *what* it is measuring; and the discrete versus integrative model is once again prominent in discussions of cloze as a reading assessment.

Reading assessment, like any other language skill, can be tested using either the *discrete-point* or the *integrative* approach (as defined in Chapter 1). This choice is dependent, obviously, on the goals of the test itself; that is, is the test designed to elicit evidence regarding learners’

global textual understanding, or is it meant to reveal elements related to singular aspects of language? Alderson (2000) remarks on the debate regarding the soundness of discrete-point tests of reading, in which some have argued that it should be tested using a "global, unitary approach" (p. 207). Interestingly, it is the very ambiguity of what precisely cloze can be said to measure that lends support to both the global and the local arguments. Alderson (2002) suggests that some have used the difficulty of ascertaining what exactly is assessed in a cloze procedure as a justification for considering the cloze as "ideal" for global reading assessment. On the other hand, he notes further that skepticism has also been expressed on those same grounds, in that the ambiguity surrounding the construct assessed in cloze tests prevents our ability to categorically assert that cloze measures global ("unitary") skills (p. 207). If cloze is designed to measure overall language proficiency, and reading is conceived as a unitary construct, the possibility exists that cloze procedure may be a viable format for reading tests. The reality, however, is that identifying the construct actually measured in any given cloze test is a difficult enterprise.

Reading comprehension and cloze: An attempt to differentiate task and test difficulty

Difficulty levels have been shown to affect the validity and reliability of tests designed to measure reading skill. Since reading comprehension is necessarily a component of the cloze task, this study pursued the question of how the difficulty of the cloze *task* (i.e., filling in discourse gaps) and the text itself might be isolated and differentiated. To that end, it became necessary to devise a way in which reading comprehension might be measured with respect to the cloze passage. The following section discusses the construction of reading comprehension questions, with a view toward implementing such questions as a textual comprehension measure alongside a cloze passage.

Alderson (2000) emphasizes the distinction between *item* difficulty and *text* difficulty, both of which affect a test's overall difficulty, and notes, "It is clearly possible to ask easy questions of difficult texts, and difficult questions of easy texts" (p. 86). In particular, scores on reading tests may be misinterpreted due to confusion between the difficulty of the test items and the overall readability level of the passage itself. Alderson suggests that it is ultimately very difficult to fully distinguish between item and text effects on difficulty, since the two are necessarily interactive (p. 86). Among the factors that influence difficulty levels of reading tests is the language of the questions. It is generally agreed that questions should be no harder than the passage itself (Nuttal, 1982; Alderson, 2000), in order to avoid adding unnecessary difficulty not related to the purpose of the task. The second-language testing context contributes the additional concern of whether to word the items in learners' L1 or the target language. Could it be that wording test items in the first language will eliminate any irrelevant task effect possible when students are asked to formulate and write words and sentences in the target language *in addition* to comprehending the text passage? For homogenous groups such as those in the present study, this decision is made easier by a common L1, meaning that test questions can be composed in one language. Shohamy (1984), in a study of the relative difficulty of reading questions presented in the L1 and the L2, found that both multiple-choice and open-ended questions were easier in the L1, particularly for beginners. Notably, item analyses showed that difficulty was particularly affected by L2 vocabulary, especially in the multiple-choice format.[5] In addition to the alleviation of anxiety (especially in lower-proficiency students), the practice of wording reading questions in the L1 may be more task-appropriate since learners are likely to

[5] For a comprehensive review of the role of vocabulary in reading comprehension, see Zhang and Annual (2008).

initially access their answers in the L1 (Alderson, 2000). Another compelling argument in the debate regarding the language of questions is Nuttal's (1982) contention regarding the language of learner response: "the inability to express themselves in the FL needlessly limits the kinds of response students give, and the quality of the response too" (p. 131). Furthermore, in an answer to the argument that reading the questions *is* part of the reading task, Nuttal maintains that because the ultimate task in a reading test is text comprehension, overly-difficult questions that "distract" from students' demonstrating comprehension should be avoided. Moreover, the clues present in the questions about the text itself may be a desirable effect of presenting questions in the L1. Like Alderson (2000), Nuttal (1982) concluded that test questions, whether written in L1 or L2, should be composed of the clearest language possible.

Nuttal proposed 5 types of reading questions:

> Type 1: Questions of literal comprehension
> Type 2: Questions involving reorganization or reinterpretation
> Type 3: Questions of inference
> Type 4: Questions of evaluation
> Type 5: Questions of personal response (p. 132)

The first type of question is one in which the answer is plainly accessible in the text, and may often be answerable in the text's own words; it demands literal comprehension. Nuttal contends that ability to cope with this type of question is a prerequisite for handling the increasingly complicated questions represented by Types 2-4. Question type 2 involves accessing the literal information of Type 1 that appears in different parts of the text, and reassembling it to form an interpretation. Nuttal considers the value of these types of questions to lie in their demand that "the student consider the text as a whole rather than thinking of each sentence on its own; or in making him assimilate fully the information he obtains" (p. 132). Question 3, in asking the student to draw inferences about what may be implied in the text,

requires the student to grapple more with intellectual rather than linguistic difficulties; Nuttal refers to these types of questions as requiring the comprehension of the text's "joint implications," that is, conclusions that can only be drawn by understanding aspects of the text and applying them to one another. Evaluative questions like those represented by Nuttal's Question 4 are the most "sophisticated" type of question, and require the student to at once comprehend, respond to, and analyze the text. These sorts of questions involve, for example, the student's assessment of the writer's intentions and or/credibility, and are appropriate only for advanced students. Finally, questions in category 5 are those which ask the reader for his own personal reaction to the text, including responses such as "I'm convinced"; "I'm not interested"; "I'm moved"; and "I'm horrified" (p. 133). The research on reading comprehension informs the present study in that the cloze task certainly demands reading ability, which should be accounted for in the administration of a cloze test.

Conclusion: Important Evidence from Three Decades of Cloze Research

This chapter, though by no means exhaustive, has highlighted much of the major research efforts undertaken to explore the behavior of the cloze procedure for designing tests of language ability. Indeed, it would be nearly impossible in a study such as this thesis to address every cloze study published—such is the work of a doctoral dissertation. What the current paper *has* done, however, is shed light on the major debates embedded in cloze design. To briefly recapitulate: Taylor's (1953) cloze procedure, while initially designed to establish the readability (difficulty) levels of journalistic texts, was later adopted by first- and second-language acquisition theorists. The fundamental controversy that would arise from the use of cloze in L2 applications was, and perhaps still is, the question of what language abilities are actually being measured by the cloze test. Oller's extensive work with cloze throughout the 1970's, particularly in light of his theory

of the "general language proficiency factor" (also known as "expectancy grammar") has produced a convincing argument for cloze as measuring overall, integrated language proficiency. An especially attractive aspect of Oller's argument is that the theory of expectancy grammar erodes the commonly-held notion of a receptive-productive skill dichotomy, a dualistic approach to language acquisition that, while having been historically taken for granted, is perhaps an erroneously reductive model of SLA/FLA. Thus, given the evidence reviewed in this paper, it is the conclusion of this researcher that the cloze procedure can, when designed properly, serve as an index of students' general proficiency level in a given L2 or FL.

Beyond the primary controversy of global versus local proficiency, other lingering questions remain, namely the issue of scoring the cloze test. As shown in this review, it would seem that a scoring system allowing for any semantically-acceptable answer will produce an appropriately difficult test with superior levels of reliability. However, if Kobayashi's (2002) study is any indication, such a scoring method demands the input of a panel of native or near-native speakers of the target language. Thus, research projects that lack access to substantial numbers of native speakers are well-advised to exercise caution when employing the semantically-acceptable scoring method. Such was the case in the current study: the project's advisor, while having some knowledge of Spanish, and the secondary researcher, having good working knowledge, can by no means be said to have near-native command of the Spanish language. This fact represents a justification for the use of the exact-word-only scoring method in quantifying the data derived on the cloze test designed for this study.

The cloze procedure has been manipulated over time in such a way as to create several different permutations of Taylor's original vision: the multiple-choice cloze, the cloze elide, and the C-test all represent efforts toward improving what some have perceived as design flaws in

the original cloze format. It is the conclusion of this review that, given the high correlations that have been found between the C-test and vocabulary tests, the C-test is likely the superior measure to employ in an investigation of FL vocabulary. However, given the time and resource limitations of the current study, the C-test proved impractical to design and administer. Given the fact that designing a C-test requires the selection and mutilation of at least five texts, in order of increasing difficulty level, it would seem that this format would be suitable under different circumstances; were I to continue this line of inquiry for a doctoral dissertation, the C-test would likely be the superior option.

Regardless of the limitations of the current study, the cloze test used in this investigation was designed in a principled manner, with all prior research informing each decision. Because this is a preliminary study, the results obtained will say less about students' Spanish proficiency or vocabulary, and more about the implications of the test's design. This is a desirable outcome, in that it will reveal aspects of the study that could be improved in future replications.

CHAPTER 3: THE YES/NO VOCABULARY TEST

Introduction

The various arguments for the importance of developing measurements of vocabulary size are by now well-established in second language acquisition literature. Size is an important dimension of vocabulary assessment, and according to Meara (1992) "can be used as a rough guide to other language skills too" (p. 5), suggesting that learners' vocabulary size may have a role to play in the description of overall language proficiency. In a study examining the relationship between lexical competence and language proficiency, Zareva, Schwanenflugel and Nikolova (2005) asked 64 subjects to self-rate their "familiarity" with 73 lexical items and also to provide word associations in order to verify the words as known. They concluded generally that quantity, as well as quality, of lexical competence correlated positively with proficiency levels, suggesting that research into the relationship between vocabulary size and overall language proficiency is a worthwhile undertaking.

Very early efforts in estimating the size of vocabulary in first-language contexts using a "checklist" format included Sims (1929) and Tilley (1936) (cited in Meara and Buxton, 1987). These first attempts involved the construction of a list of words on which students were instructed to mark the words whose meaning they knew, creating a lexical decision task. Because the nature of the checklist task is part self-assessment and part conventional test, it tended not to behave well as an accurate estimate of vocabulary size in its early vicissitudes; that is, the test's "self-reporting" component tended to yield scores based on learners' overestimation of their own vocabulary size. In response to this problem, Anderson and Freebody (1983) incorporated the use of pseudowords—words that do not exist in English but obey its morphosyntactic and orthographic rules—as a built-in compensatory measure purported to

mitigate learner overestimation. The terminology used in describing items that resemble words but do not exist in the target language has undergone an evolution of its own. Meara and Buxton's (1987) Yes/No Vocabulary Test used the term "imaginary words" to refer to the fictional items; Read (1997) would use the term "non-word," and Beeckmans et al. (2001) would decide on the term "pseudoword," rationalizing: "We prefer the term 'pseudowords' to 'non' words (Read 1997a) or 'imaginary words'…since these words obey the phonotactic and morphological rules for word formation in the given language. Therefore the term 'pseudo' appears the most appropriate" (p. 236).[6] Nomenclature aside, Anderson and Freebody argued that the inclusion of pseudowords had the additional benefit of ensuring task relevance (and validity), which had been called into question in multiple-choice situations; later, Eychmans et al. (2007) would re-interrogate the nature of the Yes/No task (hereafter denoted Y/N) and its potentially biasing effects. Anderson and Freebody's inclusion of pseudowords would allow future research to apply theories of *stimulus detection*, thereby deriving formulas to mathematically correct for overestimation.

The Y/N technique as a response to limitations of multiple-choice vocabulary assessment

Not until Meara and Buxton (1987) was the checklist test used in a second-language context. Their proposed Yes/No Vocabulary Test would, over the next twenty years, be replicated and refined, eventually becoming the format of placement tests such as the Eurocentres Vocabulary Size Test (Meara and Jones, 1990). Meara and Buxton first imagined the technique as a solution to some of the problems associated with multiple-choice vocabulary testing, complaining of at least two major difficulties inherent in any of the various MC formats. They would examine the Y/N test as a possible "alternative" and perhaps more wieldy instrument that could improve on some of the undesirable behavior and effects of the MC. The

[6] This review will use "pseudoword" and "non-word" interchangeably.

first problem the Y/N format was hoped to alleviate was the multiple-choice test's questionable reliability, which can present in any of three ways. First, a student asked to choose the correct definition of a given word from a list of four possible definitions will likely need to know the meaning of not only the correct item, but also the words that make up the definitions of three incorrect (distracting) items. That is, without understanding the distractors, the student is probably unlikely to supply a principled response. Additionally, the student might also encounter problems with the syntax of the question itself if it is written in the target language. Finally, it is certainly possible for a student to know one meaning of a word, but not necessarily other, more nuanced meanings; that is, a student may be able to recognize the word, and even use it, but fail to supply the answer required of that particular item. This represents a failure of the test to account for the alternate or nuanced semantic properties possible in some words.

Thus, the reliability of the MC format is unstable due to some confusion regarding what a student needs to know about a word, what words the student must know that surround the target word, and to what extent a student must be able to cope with the syntax of a vocabulary question. A second problematic characteristic of the MC vocabulary test is its inability to efficiently test overall vocabulary *size*, the dimension that researchers such as Meara, Nation, and Laufer (e.g. Nation and Laufer 1995) are interested in. Despite the fact that the MC test is effective for the testing of individual words, it cannot be used as a reliable measure of the range of a learner's lexicon. This limitation can be observed when learner vocabulary size increases, resulting in the need for a high sampling rate, which in turn demands increasingly more test items, which become impractical to administer. High sampling rates enable researchers to see a more accurate and detailed picture of a learner's vocabulary size due to thorough coverage.

> With [a vocabulary of] 1000 words and 25 items, the sampling rate is one [word] in 40. This figure is not too bad, but as the size of the actual vocabulary increases, the sampling rate gets progressively worse. For a vocabulary of 2,000 words, a 25-item test samples only one word in 80; for 4,000 words, the sampling rate falls to one word in 160; while for a fairly advanced student with a vocabulary in the 10,000 range, a 25-item test samples only one word in 400. (Meara and Buxton, 1987, p. 144)

These claims suggest that even *if* the MC format is satisfactorily reliable for students in entry-level language courses, the problem of inadequate sampling is only aggravated by higher vocabulary sizes, probably rendering MC tests *insufficiently* reliable for advanced student populations.

All of this is to say that MC vocabulary tests are difficult to construct well. According to Nation (1990), a well-crafted multiple choice vocabulary test will have well-chosen distractors; even so, so-called "good" MC tests still require careful and costly pre-testing and analysis. Despite their reputation as easily scored, and the fact that tests such as the TOEFL make use of the technique, MC tests at one point became the impetus for the development of an alternative and seemingly "simpler" test design method thought to be tenable in a checklist-style format.

Preliminary investigations of the Y/N test in L2 contexts

In spite of the optimism with which Meara and Buxton (1987) discussed their Y/N test's preliminary results, retrospective discussions of the Y/N vocabulary test's first use have resulted in the acknowledgement of several serious problems inherent to the test that must be examined empirically before it can be regarded and used as a valid and reliable vocabulary size estimator. The central problem is the issue of scoring methods, which in turn hinges directly on the introduction of pseudowords. Subsequent replications of the format have focused in large part

on examining the available scoring procedures and proposing new ones, and Beeckmans et al. (2001) assert that addressing the scoring problems inherent in the Y/N test is the prerequisite for dealing with any additional issues. Indeed, the question of how to score the Y/N test is the center around which all other discussions of the test revolve, and any other problematic elements of the test seem in fact to be residual fallout from scoring difficulties. This is to say that, in effect, all problems seemingly unrelated to scoring are both satellite to and bound up in the engulfing issue of scoring. What follows is a review of the available research[7] addressing the problem of Y/N scoring and other interrelated issues involved in the construction and use of Y/N vocabulary tests, as well as an exploration of how various studies have attempted to improve upon the test's design.

To begin, we will look closely at the first proposal of the Yes/ No technique as a way to produce tests of second-language vocabulary knowledge. Unsatisfied with the aforementioned drawbacks in multiple-choice vocabulary assessments such as the vocabulary component in the Cambridge First Certificate (CFC) examination, Meara and Buxton (1987) launched a preliminary investigation of the Y/N technique for constructing vocabulary tests. Based on observations of Anderson and Freebody's (e.g., 1983) work with the Y/N technique in testing the reading comprehension of native speakers, Meara and Buxton identified several advantages to the format, including ease of construction and the ability to test large amounts of items in a short period of time. In assessing vocabulary size, testing large numbers of items translates into the ability to gain a far more representative sample of "known" words, allowing for a more precise measure of overall vocabulary breadth.

[7] Several studies which might have added insight to this review are published in Dutch and French and therefore can be neither accessed nor read.

Meara and Buxton set out to test the performance of the Y/N format by conducting an experiment which correlated scores on a 100-item Y/N vocabulary test with a 25-item multiple-choice vocabulary test approximating one that might appear on the CFC examination. Their Y/N test consisted of 60 real French words and 40 "imaginary" words supposedly derived using morphological/orthographical patterns allowable in French. Students were given the directions: "Tick the words you know the meaning of e.g. milk" (p. 154). Beeckmans et. al (2001, p. 239) would later take issue with these directions, arguing that they rendered the test's format ambiguous; that is, if the Y/N test cannot truly be a *forced-decision* task, it will be impossible to distinguish between un-ticked words representing those unknown to the reader and omitted responses (unanswered test items). For the history of how Y/N instructions have evolved, see Beeckmans et al. (2001, footnote on p. 240).

The Y/N technique, despite its seeming simplicity, functions with internal complexity. In order to cope with the obvious problem that learners may tend to either over- or under-estimate their knowledge of a word depending on individual response style (leading to a *response bias*), Meara and Buxton followed Anderson and Freebody (1983) in the incorporation of pseudowords. Since the Y/N test consists of both real and "imaginary" or (pseudo) words, four types of learner response are possible. The inclusion of pseudowords provides a check on the learner's accuracy of self-assessment, or "truthfulness" (p. 145). Meara and Buxton adopt the following notation system for the four possible types of responses: RY indicates the correct acceptance of a real word; IY means an incorrect acceptance of a pseudoword; RN is no knowledge of a real word; and IN is correct rejection of a pseudoword (p. 145).

In order to correct for the occurrence of IY responses and adjust scores downward accordingly, Meara and Buxton (1987) replicated Anderson and Freebody's (1983) application

of *signal detection theory* (Green and Swetts, 1966), or SDT, which allows for necessary corrections to test scores according to the rate of "false alarms" (acceptances of pseudowords). The researchers illustrate the process in the following example:

> Suppose, for example, that the [learner] claimed to know 50% of the real words, but also claimed to know 20% of the imaginary words. The IY score would indicate that the [learner's] score of 50% on the RY cell needed to be adjusted downwards. Signal detection theory allows us to do this in a principled, non-arbitrary way. (Meara and Buxton, 1987, p. 146)

Later replications of this seminal study would continue to apply theories of signal detection in order to mathematically correct scores for learner overestimation of vocabulary knowledge. Eventually, several SDT-based advanced models would be developed in order to improve upon Meara and Buxton's (1987) first use, which employed Anderson and Freebody's (1983) original formula derived and adapted directly from SDT: words known = real words marked – nonwords marked / 1 – nonwords marked (Nation 1990).

Signal detection theory expresses this formula in terms of "hits" and "false alarms" as follows: $P(k) = P(h) – P(fa) / 1-P(fa)$, where $P(h)$ is the probability of a "hit," or correctly acknowledged real word expressed as a proportion of real words recognized; $P(fa)$ is the probability of a "false alarm," or claiming knowledge of a pseudoword; and $P(k)$ is the likelihood of acknowledging a real target. For practical purposes, $P(k)$ can be thought of as how many of the target words the learner can be said to know.

The results of Meara and Buxton's preliminary use of the Y/N technique in second-language vocabulary testing yielded satisfactory correlation between the Y/N format and the MC format (r=.703), but this was notably lower than Anderson and Freebody's (1983) robust .84 derived from native speakers. The researchers argue that the two important differences between Anderson and Freebody and Meara and Buxton are that the latter were able to construct both

tests using the same items, and had access to a homogeneous subject group. In the 1987 work, of 100 subjects, the largest single group (*n*=18) was a homogenous group of French speakers and produced a .829 correlation between the Y/N and the MC, approaching Anderson and Freebody's finding of .84. Meara and Buxton here accurately predict that a recurring difficulty in later replications would be the influence of L1 transfer and pseudoword cognates, and call for more definite research on how to construct pseudowords. Some further consideration of pseudowords would occur, and it would later be suggested that Romance languages, particularly French (Meara, 1992), present specific problems in Y/N test construction due to their being highly cognate with English; however, comparatively little specific empirical evidence exists to guide the construction of non-existing words—what is written is largely hypothetical. Meara and Buxton conclude that, in addition to correlating well with a multiple choice vocabulary test, the Y/N test also later discriminated much more accurately between candidates of the Cambridge First Certificate in English examination, accurately predicting the results of all but five of the 26 candidates from the original pool of 100 who sat the CFC. The researchers optimistically propose the possibility of using the Y/N vocabulary test to assess a far greater number of words than is feasible in a multiple choice test, which in turn will significantly enlarge the sample of learners' vocabulary that can be examined. Meara and Buxton (1987) predicted that with the Y/N vocabulary test, "several hundred items can be tested effectively in the space of a few minutes" (p. 150). What this meant for the study of vocabulary assessment is that the Y/N test might make possible a means by which to quickly quantify a learner's vocabulary when used in tandem with a "formal sampling mechanism." Meara and Buxton's (1987) preliminary work with this format in an L2 situation concludes with an optimistic view of the potential for the use of the test in second-language vocabulary research.

Extrinsic motivation and response bias

Response bias in the Y/N format provides an effective lead-in to a discussion of the details of the format's psychometric problems. Huibregtse, Admiraal and Meara (2002) describe the phenomenon of "response style" as that in which the learner, when faced with a yes or no decision of which she is unsure, will unsystematically tend toward one or the other response (see also Nunnally & Bernstein, 1994). Learner response bias, or a tendency to "lean" toward either a "Yes" or "No" response, can be characterized as a test-taking strategy, as explained by Eychmans et al. (2007). The study of response behavior has its roots in Signal Detection Theory. Hoshino (1991) conducted a study of word-frequency in recognition memory in which subjects demonstrated a bias in favor of the positive response to high-frequency words as opposed to low-frequency words, citing McCormack and Swenson (1972) as having introduced the investigation of the effects of word-frequency on response bias in the Y/N test format. Because the observable phenomenon associated with response bias is a high-false alarm rate (Eychmans et al. 2007), these rates were examined among two lists of words: one of high-frequency and one of low frequency-words. Later studies (McNicol and Ryder 1971, cited in Hoshino 1991) would go on to provide empirical evidence for the presence of a learner bias toward positive responses to high-frequency words in recognition memory. It is interesting to consider how evidence such as this might inform the construction of a Y/N vocabulary test, since vocabulary recognition is to some degree bound up in memory (Ortega, 2009).

Twenty years after Meara and Buxton's introduction of the checklist test to second-language assessment, Eychmans, Van de Velde, van Hout, and Boers (2007) proposed the necessity of examining the interaction between the Y/N task and learner strategies in order to isolate possible response bias. They argued for the presence of evidence pointing to "a complex

interplay between the multitude of characteristics that determine the learner (cultural, psychological, sociological, etc.) and which constitute his attitude to the task" (p. 61). Thus, individual subject differences influence the style of test response. Eychmans et al. argue that the effects of attitudinal differences within individual learners work against the assumption that learners can (if they choose) present an absolutely accurate picture of their vocabulary knowledge, that is, that they can monitor and refrain from the use of such test-taking strategies as guessing and overestimating the extent of their knowledge. Furthermore, the use of pseudowords, designed as a way to make learner strategies "explicit" and to "quantify and compensate for them" (p. 59), while it has been assumed to be a sufficient means of controlling for guessed responses, deserves a closer examination in terms of the effect it may have on the test's reliability. Eychmans et al. state their purpose as being "to question how and why test-takers use strategies like guesswork, how varied they are in doing this, and to ask whether compensating for guesswork in the scores such tests present is really possible" (p. 59). Their discussion centers around the idea that *extrinsic motivation*, or forces outside of the domain of the immediate situation, i.e., the vocabulary test, may bias the learner in favor of one type of response strategy or another. Extrinsic motivation is particularly influential in a high-stakes testing situation, in which a learner's best interests are served by utilizing strategies to improve performance. As stated previously, the Y/N technique for producing vocabulary tests is unusual in that it could actually be functioning as a self-assessment task disguised as an objective vocabulary checklist. The nature of the self-assessment task is that it depends heavily on a *decision criterion.* What this means is that no matter the quality of a learner's knowledge of a word, the nature of the Y/N task demands either a Yes response or a No response—the task does not allow for ambivalence. In middle-ground cases in which a learner is not entirely certain or

confident (e.g., Yule, Yanz and Tsuda, 1985) in her knowledge of a particular word, her tendency toward either Yes or No could be the result of the nature of her response *style*, that is, overestimation, underestimation, or omitted response.

Eychmans et al. (2007) make use of an economic model noted by Shohamy (2001) as analogous to extrinsic motivation in the context of language testing, and this model is particularly useful in illustrating how external factors conspire to affect learner attitudes and ultimately their test performance. Shohomy argues that a test's power to cause a behavior change in the learner can be compared to theories of economics that emphasize the strategies producers and consumers employ in order to maximize gain (e.g., Bourdieu 1999; cited in Shohamy 2001). In this view, the prospect of various types of capital (financial, cultural, and personal) serves as the motivating objective for individual behavioral choices. Bourdieu's economic model of extrinsic motivation operates as follows:

> …various situations, which may not be governed by strictly economic logic, may nonetheless conform to a logic that is economic in a broader sense because the individuals concerned are trying to gain some kind of capital (e.g. cultural or symbolic capital), or the increase of some kind of symbolic 'profit' (e.g. honour or prestige.) (Eychmans et al., p. 2007)

The applicability of this model to language testing (particularly in high-stakes testing situations) is demonstrated by the fact that test performance to some extent may influence educational and career opportunities, and increase or decrease recognition from teachers and peers. Eychmans et al. (2007) argue that this sort of interference is not beyond the control of the researcher; what they do question, however, is if and how the Y/N task may in fact encourage or cause such biased response (e.g., overestimation of vocabulary size for the sake of inflating the representation of one's ability).

To illustrate further the idea of task type influencing test validity, Eychmans et al. reference several uses of the Y/N Vocabulary Test with French-speaking learners of Dutch (Beeckmans et al., 2001; Eychmans, 2004). They report findings which repeatedly pointed to an inherent response bias in the Y/N task; that is, the majority of the subjects they tested exhibited some sort of predisposition, tendency, or preference for either a Yes or No response. The study was able to show that this bias was unrelated to (independent of) the state of the subjects' vocabulary knowledge. The researchers attribute this underlying bias to the type of task required in a Y/N test, and allege that the task in the Y/N test is actually one of self-assessment—an object of examination in its own right—rather than demonstration of vocabulary knowledge. In order to clarify further what a response bias actually "looks" like, it is useful to contrast the Y/N task with that of the True/False (T/F) task. The two are often confused, but in fact, according to Eychmans et al., they differ in that the response bias present in the nature of the T/F task is actually relevant to the competence that the test seeks to measure, unlike what has been observed in MC tests. The example used in their discussion is the learner who tends toward using only simple grammatical structures, therefore producing a frequent response of "false" on a T/F grammar test composed of mostly complicated grammatical structures. Again, this type of response bias can be seen as necessarily part of the task, and indeed, part of the learner's ability that we wish to measure in the T/F situation. A response bias in this situation does not undermine the test's validity, since the task is relevant to the construct under examination. In contrast, Eychmans et al. argue that the task effect produced by the Y/N test is unrelated to vocabulary knowledge, and indicates extra-linguistic motivating factors—a blow to the concurrent validity of the Y/N model.

Several attempts to re-confirm the validity of the Y/N test began with efforts at lessening or eradicating the observed response bias. Eychmans (2004) concluded that the "false alarm rate" is the "surface phenomenon through which a response bias is revealed" (cited in Eychmans et al., 2007, p. 64); therefore, it followed that the false alarm rate should be the target of reduction. They hypothesized that response bias could be accounted for and resolved by issuing clearer instructions on the test; this plan was based on the idea that ambiguous or imprecise *task description* may account in part for the manifestation of response bias. While the clarifying of instructions was accompanied by a reduction in the rate of false alarms, it did not serve to improve the test's concurrent validity when compared to a translation test of the claimed words from the Y/N checklist. These results indicate two things: improving the directions is not an automatic fix for the response bias inherent in the Y/N task, and also, a translation test of claimed words is not a reliable confirmation measure for the Y/N task, as evidenced by the fact that learners "made a very poor job" of translating the words they had marked as known (p. 64). Eychmans et al.'s major experimental effort at reducing response bias, however, also involved a different examination of the test's characteristics, as will be seen below in the discussion of the Computer-Adaptive format of the Y/N Vocabulary Test.

Pseudowords and the role of the first language

In his review of the Eurocentres Vocabulary Size Test, Read (2000) reiterates the need for additional work to describe the influence of learners' linguistic backgrounds and their "reactions to non-words." He raises a concern regarding pseudowords which exist as words in the learners' L1, and asserts that such words should not be used. An example of such an instance is Meara and Buxton's (1987, p. 148) example of the English pseudoword *observment*. Because this non-word resembles a morphosyntactically/orthographically possible word in Romance

languages but not Germanic languages, a French speaker might have a greater chance of claiming it as a known word than a German speaker who would likely correctly reject it. Again, the cognate effect in the literature has shown French as particularly problematic in this respect because of the "close relationship" between the French and English vocabularies (p. 129). Arguing against the need for pseudowords, Shillaw (1996) applied the Rasch Model in an item-analysis of a series of conventional Y/N tests. The Rasch analysis showed that tests containing pseudowords were no more reliable than tests containing only real words; the real-word tests also were also themselves highly reliable. Mochida and Harrington (2006) would later support the argument that pseudowords are superfluous.

Nation's (2001) discussion of the Y/N format references Meara, Lightbrown and Halter (1997), who observed the cognate effect and discovered that tests in which cognates comprised 50% of the words tended to significantly elevate scores as compared to a version with no cognates. However, Nation (2001) argues that their study failed to report an important question: did the other 50% of the words also resemble cognates? Nation insists that, in order for the test to behave as desired, the pseudowords constructed for the test must resemble the test's real words, "so that the only way of distinguishing real words from non-words is through familiarity with the real words" (p. 348). The argument that cognates represent a major variable in the behavior of Y/N vocabulary tests is well-documented (van Heuven et al., 1998). Mochida and Harrington's (2006) study is an example in which the cognate effect was *not* an operant, since the subjects' first languages were all non-alphabetic: Chinese, Cantonese, Vietnamese, Japanese, Korean, and Thai (p. 94). Interestingly, although this study did not systematically isolate, control for, or examine the absence of cognates, Mochida and Harrington offer the idea that the lower false alarm rates shown in their results may be due to the very fact that cognate effect was

not an issue. Extending this argument might also lead to support for the hypothesis that the Y/N test *is* sensitive to L1-transfer, and furthermore, that Romance languages are indeed more susceptible to cross-linguistic interaction.

Much in line with their overall argument regarding the role of extrinsic motivation in the Y/N task, Eychmans et al. (2007) investigated cognate effect in an item-analysis in order to confirm the idea that it might be implicated in response bias, despite previous evidence that the presence of cognates does not result in overestimation (e.g. Meara, Lightbown and Halter, 1994; Eychmans 2004, cited in Eychmans et al. 2007). In fact, the presence or absence of cognates was not shown to interact significantly with the false alarm rate, an indication that cognates do not contribute significantly to response bias. Eychmans et al. (2007) interpreted their failure to observe a cognate effect as evidence for the view that "the response bias functions independently of lexical skills or linguistic factors" (p. 75). These findings add further support to their argument that the Y/N task is in fact operating extralinguistically to some degree.

The first pseudowords were constructed following lenient and ill-defined guidelines. Anderson and Freebody's (1983) method consisted of changing one or two letters in real words, and "by forming unconventional base-plus-affix combinations (e.g., *observement*, *adjustion*)" (Nation 1990). Beeckmans et al. (2001) assert that pseudowords should be of the nature to allow instructors who are both native and nonnative speakers of the target language to construct Y/N tests, and offer two alternate procedures which, they argue, can obtain the same resulting non-word as the traditional technique. The first involves changing the affix of a real word (their Dutch example is to change *prettig* to *pretachtig)* (p. 245). It must be noted that not all words lend themselves to this type of manipulation; in this study, this change was administered to as many real words as was possible, creating 22 pseudowords. The remainder of the words in this

study were subject to a different type of mutilation: "the substitution of one or two graphemes with respect to the phonotactic and morphological rules for word formation in Dutch…Example: *timmerman* is turned into *tommerman*' (p. 245).

Mochida and Harrington (2006), in their study of a Y/N listening test, would take the view that the inclusion of pseudowords in the Y/N test is not only problematic, but unnecessary. The pseudowords in this study were derived through a one- or two-letter alteration process of real words which preserved English phonotactics, as in Anderson and Freebody's (1983) first use. Mochida and Harrington's pseudowords were chosen based on corresponding frequency to words on the Vocabulary Levels Test (Nation) in order to establish symmetrical variation in length between real and non-real words. They note: "Pseudowords at the lower frequency levels were on average slightly longer than those at the higher levels" (p. 82). Additionally, Mochida and Harrington incorporate Ziegler and Perry's (1998, cited in Mochida and Harrington, 2006) concept of pseudoword neighborhoods in order to explain the high false alarm rates in their study. Neighboring words, according to this view, are similar in either orthography or phonology (or both). According to Ziegler and Perry (1998), the factors of item similarity and "strength of representation" can be used to estimate the probability that a learner might confuse a word and a non-word (in Mochida and Harrington, 2006). The argument put forth by Mochida and Harrington with regard to the effect of pseudowords on false alarm rate is:

> One important moderating factor in false alarm performance is the number of neighbors a pseudoword item shares…Pseudowords at the various levels [of the VLT] may differ in their similarity to words the individual knows, and this pool of potentially confusable words may help explain the [high false alarm rate]. (p. 93)

The important point to glean from these findings is that the number of distracting pseudoword neighbors has an effect on the rate of false alarms in a given YN score. The idea of

neighborhood size (the number of words closely resembling and often confused with a given word) is particularly significant in situations involving lower-proficiency language learners. Mochida and Harrington postulate that since lower-level learners are probably just beginning to acquire recognition familiarity of an increasing amount of words, they are burdened with a large number of potentially-confusable items. That is, beginners are likely to recognize common word suffixes and affixes that recur in many of the words they encounter, but not evaluate their use. In Spanish, examples of this could be *–mente* and *–ación,* which are common adverbial and nominal suffixes respectively. A beginning Spanish learner might overextend these word endings and accept non-words such as *mentiramente*, attaching a familiar verb *(mentir)* to a familiar ending (*--mente)*. Furthermore, although it is unlikely that a beginner will associate the non-existant *mentiramente* with the low-frequency real *metidamente,* this issue probably becomes a concern as vocabulary breadth increases. Mochida and Harrington's (2006) comments regarding how learner proficiency interacts with confusable neighbors implies a relationship between the two, namely that "lower proficiency individuals should have more false alarms at the higher frequency levels, with the false alarm rate progressively decreasing as a function of increasing word difficulty" (p. 93). They call for systematic research to investigate this possible "proficiency effect," and propose a possible alternative for the use of pseudowords altogether. They argue that replacing the pseudowords that might be used in a more conventional Y/N test with the same number of real low-frequency words would preserve the balance of positive and negative responses, thus eliminating the difficulties associated with the neighborhood effect. Although it is possible for the cognate effect to operate in this situation, remaining a concern, Mochida and Harrington offer the unique perspective that the elimination of pseudowords would enhance the "face validity" of the test for practical classroom use, and

expose learners to contact and practice with a greater number of real target words. These conclusions support those of previous work in which the Y/N test did not perform well with lower proficiency subjects (Beeckmans et al., 2001; Meara, 1996).

Scoring: The prerequisite to validating the Y/N vocabulary test

Until now, this review has explored the ancillary difficulties of the Y/N vocabulary test format that arise from the central question of scoring. Beeckmans et al. (2001) assert that "the problem of establishing an adequate scoring method is more than just one relevant issue among others: it constitutes a prerequisite in order to be able to address many…problems" (p. 241). The authors explore the problem of scoring from the perspective of a *continuous-discrete model*, the focal point of which is the element of response bias, and problematize previous working definitions of response bias that failed to account for the possible ways in which it manifests. Indeed, the "central issue" for Beeckmans et al. is "controlling the response bias for the Y/N format in order to be able to eliminate it from the raw data," since the very practice of adjusting raw scores downward with the use of correction formula carries with it questions of reliability (p. 259).

In general, the continuous-discrete model of scoring can be understood as describing how various scoring methods handle the element of response bias. The discrete method, resting on the basis of the "threshold theory" which assumes an "all-or-nothing" nature of subjects' responses (p. 260), does not contend with response bias in a direct sense. On the other hand, the continuous approach (taken from SDT), "posits a continuum ranging from 'being sure of the presence of a signal/word' to 'being sure of the absence of the signal/pseudoword'. The middle of the continuum corresponds to maximal doubt" (p. 260). Thus, while the discrete model fails to take into account variance in learner response due to doubt, guessing, strategy, and omitted

response (all types of bias), the continuous model employs a technique by which such variables are acknowledged and corrected for. For the purposes of this review, it is less important to engage in a full description of the complexities within various measurement theory models justifying the resulting mathematical formulas than to provide an overview of the types of scoring methods that have been explored to date. Therefore, what follows is a chronological account of the evolution of Y/N scoring formulae and their observable differences in behavior.

According to Beeckmans et al. (2001), the simplest way to obtain overall test results would be to simply interpret the correct response rate, that is, to look at the proportion of Yes to No responses to yield a raw score. However, the introduction of pseudowords saw the emergence of a need for more complicated scoring techniques, for which theories from the study of signal detection were borrowed. Because SDT (Green and Swets, 1966) is concerned with describing decision behavior in tasks of "detection," the field of language testing adopted it in order to sort out the complex procedure of factoring in the presence of false alarms.[8] The devastating argument in favor of more advanced scoring models than the "number of correct responses" technique (Huibregtse, Admiraal & Meara, 2002) is the idea that the older scoring technique is incapable of correcting for individual response style (p. 230).

The discrete approach to Y/N scoring falls on the lowest end of the "correction for guessing" models, in that it imposes the least amount of correction to raw scores. More aptly named the "correction for blind guessing" model, it assumes that two possibilities exist for each test item (Yes or No), and that learners will either know the correct response, rendering the probability of a correct response as 1, or guess completely at random (blindly), with the probability of a correct response decreasing to $1 / k$ (where k equals the number of response alternatives). Furthermore, it assumes that if the subject knows a word, the response given will

[8] For a comprehensive description of SDT as applied to language testing, see Beeckmans et al (2001).

be correct; if a subject does not know the answer, she will either fail to respond or guess blindly. This technique was applied in the earliest uses of the Y/N vocabulary test format in native-speaker contexts (Sims 1929; Tilley 1936), consistently leading to overestimation of learners' vocabulary size in response to inflated scores. Obviously, the issue of learner bias is not accounted for in this scenario.

Green and Swets' (1966) SDT formula seemed to be a solution for the shortcomings of the blind-guessing model. Their original formula was expressed as $P(h)=P^*(h) + P(f)$ [1-$P^*(h)$] where $P(h)$=the observed hit rate; $P^*(h)$=the true hit rate; $P(f)$=the false alarm rate. Anderson and Freebody (1983) reformulated this calculation in order to apply it to the needs of the Y/N test, resulting in the following formula: $P^*(h) = P(h) – P(f) / 1 – P(f)$. The SDT-based method is the advanced model of correction, and was expressed by Meara (1992) as Δm. Meara first proposed Δm in his introduction of the EFL Vocabulary Tests. In order to make these formulas meaningful, it is useful to observe how they have behaved in context. Huibregtse, Admiraal, and Meara (2002) compared the effects of three scoring techniques; this work centered around the three crucial factors present in any Y/N scoring situation: 1) "there are two types of correct and two types of incorrect answers"; 2)"participants have the possibility of (sophisticated) guessing"; and 3) "participants show different response styles" (p. 230). The terminological distinction made by Huibregtse et al. between "blind guessing" and "sophisticated guessing" has important implications, in that it reinforces the argument derived from SDT that subjects do not guess at random in a decision task such as the Y/N test.

Because so-called "correction for guessing" models of scoring are in fact based on a "blind guessing model" (it assumes only two possible responses), Huibregtse et al. argue that,

among other problems with this formula, there exist some "peculiarities of the equation," namely:

> In the event that the observed hit rate equals 1.0, the corrected hit rate will also equal 1.0, regardless of the false alarm proportion. This means that participants who show correct responses on all items…would obtain the same score as participants who give 'yes' responses to all real words as well as to most of the pseudowords…In the event that the false alarm rate equals 1.0, the equation cannot be solved. (p. 232)

This difficulty represents one of several ongoing mathematical problems with regard to scoring method, and is currently undergoing continued investigation. In their comparison of the scoring measures *h-f* (proportion of hits minus the false alarm rate), "correction for guessing" (cfg), and Meara's (1992) Δ*m,* Huibregtse et al. (2002) found none of these techniques to satisfactorily account for the conditions of sophisticated guessing and response style, and proposed a new scoring formula: the I_{SDT} method. This formula calculates an index (expressed as a value between 0 and 1), representing a subject's score. The results obtained using this scoring formula are interpreted as assuming a "probabilistic response based on graded vocabulary knowledge" (Mochida and Harrington, 2006). Huibregtse et al. (2002) found their I_{SDT} formula to perform better as a predictor of performance on the Vocabulary Levels Test, with the mean hit rate serving as the best mean predictor of VLT performance relative to the three other possible responses. Mochida and Harrington's (2006) later testing of I_{SDT} against other scoring formula (*h-f*, cfg, and Δ*m*) would show little difference in scores using all formulae (Mochida and Harrington, 2006); in fact, the researchers concede that all formulae will tend to either over- or underestimate scores to some degree. Meara (2010), acknowledging some "incorrect mathematical assumptions" inherent in his 1992 Δ*m* scoring model, admits that the model's treatment of false alarms resulted in excessively harsh scoring effects when the total number of hits is relatively low. Furthermore, he contends that past efforts in addressing the

scoring problems of the Y/N test (i.e., Beeckmans et al., 2001; Huibregtse et al., 2002; and Mochida and Harrington, 2006) were not altogether satisfactory either. At the time of his 2010 publication, Meara stated that he was currently working on a Bayesian model, which he hoped would finally provide a definitive solution to some of the problems discussed here. Until such a model is shown to consistently eliminate scoring problems, it is necessary to choose one of the advanced SDT scoring formulae. To that end, the current study will make use of Huibregtse, Admiraal and Meara's (2002) I_{SDT} formula, given that: it has performed relatively well as a predictor of performance on the Vocabulary Levels Test; it is the most recent adaptation of the SDT models; and it has been shown to behave with relative consistency.

Practical applications of the Y/N vocabulary test

The Y/N Vocabulary Test was later developed into the computerized checklist test called the Eurocentres Vocabulary Size test (Meara and Jones, 1990), which estimates vocabulary size up to 10,000 words. Known as the VOC, the computerized version uses the same methodology as the EFL Vocabulary Tests (Meara 1992), which test large amounts of words from five frequency bands in a traditional paper-and-pencil format. The VOC was developed for the Eurocentres language schools in the U.K. which, because of short courses and a high rate of student turnover, were in need of a fast and efficient placement test. This need arose out of the observed shortcomings of Eurocentres' traditional placement test battery, the Joint Entrance Test (JET), which was composed of listening comprehension, grammar and reading, and an oral interview—an impracticality given the volume of placement testing that needed to be accomplished. The computer-adaptive Y/N test was proposed to answer the need for an alternative to the JET. Indeed, it required only 10-15 minutes to complete and was scored instantly by the computer. Meara and Jones (1988) base their proposition that a Y/N vocabulary

test could function as an indicator of *overall language ability* for the purposes of a placement test on Anderson and Freebody's (1981) suggestion that vocabulary knowledge is strongly related to all practical language skills, and that, at least in speech, a broader vocabulary will predict superior performance than a more limited vocabulary in various measures of ability.

Although based on the same principles as the original Y/N test, the Eurocentres Vocabulary Size Test is remarkable in its excellent sampling rate (one word in 20) and its sensitivity; learners are tested in five frequency bands sequentially but cannot advance to the next frequency band without scoring highly on the previous section. The computer presents sets of 10 words and 10 pseudowords until the testee establishes a passing rate; once the student passes a section, she is allowed to move on to the next frequency band. A small-scale study of the Eurocentres Vocabulary Size Test as a placement indicator (VOC) yielded a "surprisingly" high overall correlation with the JET (r= .644 in the Cambridge section and r = .717 in London, respectively). They found that native French speakers scored consistently lower than other homogenous groups, showing that although homogenous groups tend to demonstrate higher correlations than mixed groups, this is not a firm rule. It is interesting to recall here Meara and Buxton's (1987) correlation of .829 between the original Y/N prototype and a multiple-choice vocabulary test in their French-speaking group. Meara and Jones (1988) would interpret their low French-speaker correlations as the possibility of an inherent systematic bias against particular languages in the VOC, but they state that a bias in the JET is equally possible. In terms of the VOC's quality as a placement indicator, 18 students (out of n = 250) would eventually be re-placed into more appropriate courses (four in Cambridge and 14 in London). Meara and Jones suggest that, because the test's format assumes that word recognition mirrors word knowledge, the test probably slightly overestimates vocabulary size, and that this could be

remedied by appropriately adjusting the raw VOC score. Meara and Jones concluded overall that the test functioned well, and again called for further work on the construction of pseudowords and the effects of varying L1 transfer. The Eurocentres Vocabulary Test would later be incorporated into the European Dialang project, where the Y/N format is now used for diagnostic purposes.

Eychmans et al. (2007) discussed the DIALANG Y/N Vocabulary Test as an example of a test whose format may play a role in the presence of response bias. The DIALANG test, although computerized, mimics the conventional paper-pencil format by presenting items on the screen as a list, like in the original Y/N test. Other conventional features of the DIALANG include the absence of a time limit, the students' ability to change or omit responses, and their ability to choose the order in which they respond to items. These features pose the threat of response bias by enabling learners to develop strategies, which, in a valid and reliable test, must be controlled for. The study proposed an alternate computer format in which control is exerted over these elements by programming the following conditions:

1. They suggested that the computer could show items sequentially and in an isolated manner rather than in a list which can be viewed in its entirety. This tactic has several advantages, including the fact that learners would not be afforded a view of the whole test; they might not remember which and how many words they have rejected and accepted; and the nature of the forced-decision task means that item omission is impossible.
2. In order to prevent sequence effects (i.e., poorer performance on the end of the test due to boredom or fatigue), the computer could randomize item sequence among subjects.
3. A time limit has the benefit of making the test more "uniform" among all learners. I suggest that a time limit also serves the goal of preventing students' lingering over individual items, accessing implicit knowledge, and strategizing.

4. The directions (the task description) reappear at the top of each screen "in order to remind the test-takers of the exact nature of the decision they are expected to make," which the researchers argue might lead to consistency of response behavior. (p. 66)

Based on the hypotheses above, the study predicted that this increase in researcher control would mitigate the presence and effects of response bias. However, a high false alarm rate (over 20%), coupled with systematic negative correlation between an uncontrolled computerized Y/N test, pointed to an even higher level of response bias than yielded by the uncontrolled format alone. Eychmans et al. (2007) concluded that, at least for the particular format adopted in this study, the computer-adaptive format was not successful in reducing response bias, and may have even served to bias responses toward "Yes," as was revealed in an item analysis. Despite these findings, the computer-adaptive test nonetheless makes possible the efficient testing of large quantities of items which, in tests of lexical size, may trump arguments that the computerized Y/N test could introduce bias.

CHAPTER 4: THE STUDY

The purpose of the present study was to construct and correlate a cloze test and a Y/N vocabulary test in a preliminary effort to investigate the relationship between vocabulary and overall language ability.

Subjects

The subjects tested in this study were 26 student volunteers, all enrolled in one of two university Spanish 1 courses during the Fall semester of 2010. All subjects were at least 18 years of age, were informed of the purpose of the study, and participated willingly. Group A (n=14) sat for the cloze and comprehension questions only; Group B (n=12) took the cloze and the Y/N tests.

Materials

The cloze passage

The cloze test was based on a passage of text entitled "Las recreaciones y los deportes" from the out-of-print graded reader *Muchas Facetas de México* (Burnett, 1973). The passage describes popular recreational activities in Mexico, including the *fiesta* and athletics. In particular, the passage contains a description of the music, food, dancing, and religious components of the *fiesta*, with athletics comprising about a third of the whole passage. The text was altered based on several syntactic changes suggested by a native Spanish speaker and professor of Spanish; these changes did not alter the meaning of the text and were undertaken in order to update the prose to one more likely to be familiar to the subjects. The passage, in its final intact form, contained a total of 316 words. As is customary in the cloze procedure, two sentences were left intact at the beginning of the passage during deletion in order to supply context, and one intact sentence remained at the end. Using Abraham and Chapelle's (1992)

study as a precedent, an every 11-th word deletion rate was imposed; this rate was adhered to except when the 11th word fell on a function word, in which case, the next linearly-encountered content word was deleted. Abraham and Chapelle (1992), examining the effects of different deletion rates, found that the amount of context necessary for restoration (that is, the amount of intact text between blanks), did not significantly affect the test's ease or difficulty; this finding was contrary to their hypothesis that greater context promotes ease of closure. Indeed, the results of the current study appear to align with this theory, in that our 11th-word deletion rate seemed to produce a very difficult test. In all, twenty-two words were deleted and transformed into blanks of uniform length. Each blank represented a deleted content word, with the exception of numbers 5, 7, 11, and 16, which happened to be function words. These deletions were preserved for a principled reason: number 5 (*como*, or "like" / "just as,") was preserved as a deletion due to its semantic importance to the passage's overall meaning. Number 7 (*menos*) is a component of the idiomatic phrase *por lo menos*, meaning "at least," and we were interested in students' knowledge of this multi-word lexical item. Finally, both numbers 11 and 16 were *también*, meaning "also," an important discourse connector whose meaning preserved the continuity of the text. Abraham and Chapelle's (1992) fixed-ratio cloze, which contained twenty content and fifteen function words, was more difficult than their (rational) cloze consisting of 13 content and 22 function words. That the cloze test in the current study contained a majority of content words is perhaps another factor contributing to the test's apparent difficulty.

A split-half reliability estimate was calculated for the cloze test using Chronbach's Alpha and the statistics software program "R." The reliability coefficient (0.71) fell short of what Lado (1961) claimed as a good reliability coefficient for vocabulary (and structure and reading), which he located within the range of .90 to .99. However, since a minimum of fifty deletions has

generally been recommended for sufficient reliability (Brown 1993), the twenty-two deletions in our cloze test may have been insufficient for optimum reliability. That said, the time constraint under which the current study was administered (i.e., the inability to test the recommended fifty items in the 25 minutes allotted) should be acknowledged insofar as it restricted the length of the test. Additionally, Brown (1993) reported highly inconsistent levels of reliability in cloze tests generally, ranging as widely as .31 to .96. Thus, it is difficult to infer anything conclusive from the reliability level of the cloze test used in the current study.

Comprehension questions

As has been mentioned, it was determined early in the test-construction process that the cloze passage was extremely difficult to close, regardless of the chosen method of scoring. Cloze difficulty was first observed in the preliminary results of a small pilot experiment ($N = 4$) using a cloze test based on information from a Mexican tourist website.[9] This choice of text was rooted in an attempt to create a cloze test of authentic, "real-world" language. The very low scores on this cloze version lead to consultation with two educated native Spanish speakers, neither of whom were able to close more than two deletions. These preliminary results would lead to the hypothesis that, aside from the possibility that the words necessary for closure were unknown to the pilot subjects, perhaps the cloze task itself was unfamiliar and difficult for beginning-level SFL students and native speakers alike; this theory is also supported by Alderson (1980). Additionally, it was hypothesized that the readability of the authentic text chosen for the cloze was excessively difficult for the beginning-level population; in response to this observation, a simplified text from a graded reader was chosen for construction of the final version of the cloze test.

[9] http://www.visitmexico.com

To ascertain the difference between task effect and vocabulary deficit, eight comprehension questions were constructed in order to provide a "check" on the cloze results. In other words, if subjects obtained very low scores on the cloze passage but successfully answered the comprehension questions, it can be assumed that the task of supplying words to close the blanks was too difficult for the students' proficiency level, although the passage itself was appropriately readable and its vocabulary perhaps known. The comprehension questions were devised based on Nuttal's (1982) spectrum of reading comprehension questions, discussed in Chapter 2. Of Nuttal's five question types, four are represented in the eight comprehension questions administered in the present study. Students were instructed to write their answers to these questions in English. The following are the eight comprehension questions written to assess subjects' overall passage comprehension irrespective of their ability to close the blanks:

1. What is the favorite recreational activity in Mexico? (TYPE 1)
2. What are some activities that the United States and Mexico have in common? (TYPE 2)
3. How do Mexican *fiestas* differ from similar activities in the United States? (TYPE 3)
4. Describe the religious aspects of some *fiestas*. (TYPE 1)
5. Name three foods sold during a *fiesta*. (TYPE 1)
6. What do the dancers act out at celebrations? (TYPE 1)
7. A remarkable feature of Mexico City is what? (TYPE 3)
8. What is your overall opinion of this passage? (TYPE 5)

The Y/N test

The words used in the Y/N checklist test were taken from Davies' (2006) *Frequency Dictionary of Spanish: Core vocabulary for learners*. Twenty-five words were chosen from both the 1-1000 and 1000-2000 word-frequency bands (every 40th sequential word). Fifty words were

then chosen from the 2000-3766 word-frequency band, for a total of 100 words. These words were randomly sequenced into a list.

As mentioned in Chapter 3, Mochida and Harrington's (2006) study would lead them to contend that pseudowords are both overly problematic and unnecessary. However, their study consisted of a Y/N test of listening, which may affect the relevance of that claim to the current study. Anderson and Freebody's (1983) introduction of the pseudoword was a major contribution to investigations of Y/N test scoring, and allowed for the cross-disciplinary influence of Signal Detection theory (Green & Swets, 1966) and mathematically advanced scoring models. Furthermore, because the interests of the current study lay largely in replicating, in Spanish, Meara and Buxton's (1987) preliminary investigation, pseudowords were included in the Y/N test.

Of the 100 real words, 40 words were randomly chosen for transformation to pseudowords, yielding the conventional real-to-pseudoword ratio of 60:40. Pseudoword construction was based on the guidelines provided in the literature; that is, all pseudowords obeyed the morphosyntactic/orthographic rules governing Spanish words, and no pseudoword existed as a real word in Spanish. For the most part, the pseudowords were created by changing one or more consonants and/or vowels in the real word. When possible, the number of syllables in the real words was preserved in the pseudoword distortion, with some exceptions. When the number of syllables *was* altered, the alteration typically added one syllable. Two real words with tildes (ñ) and one with an acute accent (á) were transformed into pseudowords bearing the same orthographic marks. The resultant pseudowords conformed to Mochida and Harrington's (2006, p. 82) finding that lower-frequency real words that are transformed tend to yield "slightly longer" pseudowords than higher-frequency words. The intention during pseudoword

construction was to create a fictional word that could plausibly mimic the frequency of the real word from which it was derived. In the pilot study, all pseudowords were acknowledged by two native speakers as orthographically/morphologically possible in Spanish; one of the native speakers also provided one false alarm response in an informal trial of the test.

The cognate variable in Y/N vocabulary tests has been noted by Meara, Lightbrown and Halter (1997), who found that tests in which 50% of the FL words were cognate in the L1 tended to yield significantly elevated scores as compared to a version with no cognates. Cognates in the current study's Y/N test were relatively few, comprising only 20% of the total number of items (20 real cognates out of 100 items; no pseudowords were cognate). Mochida and Harrington (2006) offered the alternate view that the lower false alarm rates shown in their results may be due to the very fact that cognate effect was not an issue; however, it is worthwhile to question whether the small number of cognates in our study may have contributed to the relatively low false alarm average of .022. Below are the words appearing on the Y/N test component. Boldface words are the pseudowords derived from and replacing the real word to the left.

extraer
suyo
tal
indudablemente→ **insubreciente**
misa
sentimental
sucesivamente→ **pelebremente**
grano→ **lingro**
circo
aguardar
consumo
presidir→ **relipicar**
lago
primo
contraer
novela
patio
animo
temprano
categoría
recto→ **gimplio**
cerrado
obligación
educar
prudente→ **boquelante**
huevo
examinar→ **meslogar**
existencia→ **liquetencia**
rana
estimulo→ **esperago**
excusa
bañar→ **pisteñar**
liso→ **lorto**
precaución
entero
cansancio
enfermero
núcleo
señor→ **peleñiro**
mezclar
efectivo
recinto
voz
critico

barrer→ **olistar**
cuando→ **puento**
frió→ **frisclo**
sorprender
religión
repentino
favorable→ **relostable**
seguridad
el→ **ul**
cifra
coincidir
señal→ **menrol**
descargar
soltar
plenamente→ **dueblemente**
abuela→ **doquela**
espectáculo→ **empegánculo**
honesto
comerciante→ **quemoliante**
parar
extenso
desorden
usar→ **reldar**
lazo→ **desco**
mal→ **nol**
brindar→ **pentar**
población
castellano
hombre
próximo
vuelta
raza→ **rezlo**
permanentemente
multiplicar
incluir
inquieto→ **implantico**
despedir
malestar→ **molebrar**
alguien→ **elguien**
solo
abuelo
hierba→ **fierlo**
texto
maduro→ **rapunto**
hoja
profesional→ **comensional**
caracterizar→ **relaterizar**
diferencia
exigente→ **empligente**
social
escuchar→ **oscuejar**
índice
demonio
aflicción→ **esfleccion**
ocurrir
frenar→ **plenar**

Administration and Scoring

Fourteen subjects (Group A) were allotted 25 minutes to complete Form A of the test, which contained the cloze passage and the eight comprehension questions. Two days later, twelve different subjects (Group B) were given Form B of the test, containing the cloze passage and the Y/N checklist. Group B was also permitted 25 minutes to complete the test. Both groups were given brief oral instructions prior to administration, and were asked to read the test directions carefully before beginning the test.

Cloze scoring

The twenty-two-item cloze passage was scored using the exact scoring method, primarily on the basis of Oller's (1973) argument that an any-semantically-acceptable scoring method must be done by a native speaker. Item 1 represents one exception to the exact scoring method

employed: following the collection of data and during the scoring process, it was discovered that an error had been made in the mutilation of the first non-intact sentence of the passage which affected the first item, effectively creating a repetition of the first blank. It is necessary to recall here Lado's (1986) observation of a similar "mundane" technical flaw in Oller and Conrad's (1971) cloze investigation, namely, an error in deletion rate; Lado suggests that such errors may, to some extent, influence the study's results. In an effort to compensate for this error, scoring for item 1 was adjusted so that any word that was semantically possible given the context surrounding the deletion was accepted for credit.

Comprehension question scoring

Each of the eight comprehension questions was assigned two points, resulting in sixteen possible points. Questions 1-7 were objective questions and full credit was given if the subject provided a fully correct answer. If an answer failed to provide 50% or more of the correct answer (i.e., instances in which the question demanded three discrete components and only one was provided), one point was deducted. An example of this is Question 5, which asks the student for three foods sold during a *fiesta.* If a student provided only two correct foods, or two correct foods and one incorrect food, one point was awarded. Similarly, answers which were technically correct (that is, found in the text) but failed to answer the question were also awarded half credit. Answers that contained a fully correct response and provided additional information of questionable accuracy were awarded full credit. For example: Subject 2, in response to Question 5 ("How do Mexican *fiestas* differ from similar activities in the United States?") answered "They celebrate people, public figures and religion." Although this answer contains the necessary component of religion, and although *fiestas* do often honor patron saints, the passage does not mention the celebration of living public figures in the traditional sense. This

student was awarded the full two points. Question 8 was a subjective question that asked participants for their personal opinion of the passage itself, and all responses were awarded full credit. Omitted responses to Question 8 and/or any other question resulted in zero points awarded for the item.

Y/N scoring

The Y/N vocabulary test was scored using Huibregtse, Admiraal, and Meara's (2002) I_{SDT} formula: $I_{SDT} = 1 - 4h\,(1 - f) - 2(h - f)\,(1 + h - f) / \; 1 - 4h\,(1 - f) - (h - f)\,(1 + h - f)$, where h=hit rate and f= false alarm rate. A tally of hits, misses, correct rejections, and false alarms was taken of each test and converted to a decimal proportion for use in the above formula. Scores obtained from the formula represent indices of students' relative vocabulary size and are represented by values between 0 and 1; scores closer to 1 are considered superior to those closer to 0.

Results

Cloze and comprehension question results.

The table below summarizes the descriptive statistics obtained for the cloze, the comprehension questions, the four Y/N response types (miss, hit, correct rejection (CR) and false alarm (FA)), and the Y/N scores represented as I_{SDT}. The cloze scores (n=26) showed a relatively normal distribution with a mean score of 2.54 and standard deviation (SD) of 2.55. Scores ranged from 0 to 10, with a median score of 2. The notably low cloze score average was immediately evident in the few individual pilot cases and persisted in the final administration, despite the simplification of the cloze passage. Scores on the comprehension questions (n=14) ranged from 5 to 15, with a mean of 9.86 and SD of 3.30. The mean score is difficult to interpret due to the small number of observations (n=14), and the fact that individuals' cloze and

comprehension scores were not specifically analyzed at an item level. The possible relationship between the cloze test and the comprehension questions, though not systematically tested for, presents a desideratum for further research.

Y/N results

Because the formula used in scoring the checklist test depends first on a simple count of the four possible responses (hit, miss, correct rejection, and false alarm), it is worthwhile to mention the statistical occurrences of each response type. The average number of "misses" was 41 (SD 5.61), with a range of 31-51. "Hits" ranged from 10-30 and averaged 19.92 (SD 5.65). The mean number of "correct rejections" was 36.92, with a range of 32-40 and a SD 2.07.

Finally, "false alarm" responses averaged 2.17 (SD 1.95), with 0 being the smallest number of occurrences and 7 being the highest. The final I_{SDT} scores were calculated by transforming simple counts of the four response types into proportions. The average Y/N index (of n=12) was .371, with a median of .367, range of .328-.429, and SD of .037. Surface observation would suggest that false alarm behavior obeyed the principle that "among subjects with the same hit rate, those with higher false-alarm rates end up with a lower test score" (Beeckmans et al., 2001, p. 238). When interpreting the results of the Y/N test, it should also be recalled that Meara (1996), Beeckmans et al., (2001), and Mochida and Harrington (2006) all showed that lower proficiency subjects tended to perform relatively worse on the Y/N test than intermediate and higher proficiency subjects, a principle that was likely at work in the current study.

Table of Descriptive Statistics

Statistic	Cloze	Comp Qs	*Y/N Miss*	*Y/N Hit*	*Y/N CR*	*Y/N FA*	Hit rate	FA rate	I_{SDT}
Average	2.54	9.86	41	19.92	36.92	2.17	0.19	0.02	0.37
Median	2	10	43	18	37	2	0.18	0.02	0.36
Min	0	5	31	10	32	0	0.10	0.00	0.32
Max	10	15	51	30	40	7	0.30	0.07	0.42
SD	2.55	3.30	5.61	5.65	2.07	1.95	0.05	0.01	0.03

Regression

Weak positive correlations presented for both the cloze and Y/N I_{SDT} scores (r=.36), and the cloze and comprehension scores (r=.28). Despite a positive correlation between I_{SDT} and cloze, no statistical significance was obtained at either the p<.05 or p<.10 levels, since the t statistic (1.23) is less than the critical values of 2.228 and 1.812 respectively. Based on a regression of cloze scores as a function of the I_{SDT} index, the estimated coefficient was β=19.8. This result indicates that an increase from 0 to 1 in I_{SDT} index is associated with an increase of 19.8 in the cloze score; an increase of .05 is associated with an increase of 0.99. A regression analysis of cloze scores as a function of scores on the comprehension questions reveals that, since the t statistic (1.01) is less than the critical values 2.179 and 1.782, r=2.80 is not significant at the p<.05 or p<.10 levels. The estimated coefficient, β=.25, indicates that an increase of 1 in the comprehension question score is associated with an increase of 0.25 in the cloze score, and that an increase of 3 in the comprehension score (based on an SD of approximately 3) is associated with an increase of approximately 0.75 in the cloze score.

Discussion and Conclusions

Broadly articulated, this study began with an interest in a possible relationship between SFL vocabulary and overall SFL proficiency. To that end, an examination of past work in the development of various language tests was undertaken in order to inform the construction of two language assessment measures, specifically, the cloze procedure (claimed to measure proficiency), and the Y/N vocabulary test, a seemingly efficient way to measure vocabulary size. The literature on cloze procedure vigorously interrogates the question of what cloze procedure measures, and it is the conclusion of this paper that cloze has, after decades of research, been shown to correlate satisfactorily with recognized measures of overall language proficiency. Thus, the cloze procedure was chosen as the method by which to construct the test of overall ability used in this study. The construct of vocabulary is multidimensional; however, what was of interest in this particular context was the breadth, or size, of a learners' FL vocabulary with respect to general ability. Because the results of this study were not statistically significant, it is impossible to draw principled inferences regarding our central hypothesis from the data obtained. However, what these results do warrant is continued inquiry into how vocabulary and proficiency might interact given a superior test design. To that end, it is necessary to consider aspects of the study's characteristics that may have contributed to less than optimum behavior, and the modifications/improvements that could be made if the investigation were ever to be replicated in the future.

Despite Brown's (1993) "natural cloze" experiment, which he suggested may point to the cloze test as being resilient to changes in difficulty level, the pilot and results of the current study seem to have been influenced strongly by the difficulty of the cloze test, possibly stemming from low readability, high task difficulty, or a combination thereof. There are at least two theoretical

explanations for the generally very low cloze test scores in this study: a) the text's readability level was excessively difficult for the beginning-level population tested, that is, the test demanded ability above students' present level; or b) the nature of the test task itself was unfamiliar and distracting. If task effect was indeed an operant variable, the validity of the cloze test can be assumed to have been influenced; that is, perhaps the test was actually assessing the unfamiliar skill of negotiating text interrupted by blanks, which disrupted the process of accessing words in working memory. It is also possible that the text chosen for use in the cloze procedure exceeded the readability level at which the subjects were equipped to perform; this factor, combined with the relatively small number of items, may have also affected the validity of the test.

The question of precisely which characteristic(s) of the cloze test design created difficulty for students is an empirical one; a possible research project to this end could be the investigation of textual difficulty levels alongside task difficulty. For further replications, it would be essential to implement a formal device with which to measure the readability of a given text (for example, the Fernandez-Huerta readability estimate), and to have more complete knowledge of the particular population's current ability levels. The administration of comprehension questions relating to the cloze passage was meant to provide some indication of the difference between poor performance because of task effect and that due to non-comprehension. A surface observation reveals that, without exception, each subject's comprehension question score at least matched (one subject) or exceeded (thirteen subjects) her cloze score (often by at least eight points), indicating that gap closure may not have performed well as an indicator of passage comprehension. Although this informal observation requires additional empirical investigation in order to be meaningful, it is at least a small indication that task difficulty, and not passage

readability, may have negatively affected subjects' cloze performance. Regardless, the cloze test proved problematic from its initial construction and pilot administration to the low scores obtained in the final study's results. To recapitulate, because of the possible effects of text and task difficulty on the cloze procedure, it would seem essential that a systematic method of ascertaining readability should be implemented, as well as a means by which readability and task effect may be differentiated by way of an item-level estimate of correlation between the cloze and an outside comprehension check.

Next, this study was administered under conditions which rendered impractical the construction and administration of a computer-adaptive format of the Y/N checklist test. As has been discussed, the computer adaptive Y/N test may offer some distinct advantages over the traditional paper-and-pencil format; the most important of these advantages is the format's ability to test an enormous quantity of items, with a superior sampling rate (and thus superior coverage) than the traditional format permits. Therefore, it can be argued that this study likely did not test a sufficient number of words to provide the best overall picture of subjects' vocabulary size. It is a conclusion of this paper that, despite the drawbacks that some have associated with the computer-adaptive Y/N format (i.e., the problem with task relevance), the ability to efficiently test large numbers of words in a short amount of time—as is possible in a computer-adaptive situation—is a highly desirable circumstance.

A final problem affecting this study was the difficulty of recruiting a satisfactorily large sample of volunteers, both for the pilot and for the official study administration. Though tentative conclusions might be drawn from the sample population in this study, the undersized subject pool (n = 4 in the pilot and n = 26 in the official administration) likely depressed the statistical significance of this study's results. If replicated under different circumstances in

which more time, a greater sample size, and a computer-adaptive test format were practical, this study would likely yield results from which firmer conclusions could be drawn.

Possibilities for Future Research

A survey of the literature of cloze research reveals that the C-test, though itself disputed, is a worthwhile candidate for further investigation; of particular interest is its tendency to correlate well with vocabulary measures. Because the labor and time necessary for the construction of a valid C-test exceeded what this study could supply, it is interesting to consider investing further work in a larger-scale study using the C-test as an alternative to the traditional cloze. As it stands, the efficacy of the traditional cloze test in measuring overall language proficiency remains an active empirical question.

The Y/N vocabulary test is itself an instrument worthy of extensive further study in that its unusual blend of components (the self-rating task aspect, the construction of pseudowords, and the continuing development of mathematically-advanced scoring formulae) represents a sizable space of investigative opportunity. Indeed, the Y/N test is both elegant and sensitive when operating under optimum conditions. The present study's examination of past work with the Y/N test has also brought to light an unexpected (and complicating) finding regarding the Y/N vocabulary test. Mochida and Harrington (2006) validated the Y/N test as a placement test and as a measure of overall English proficiency in their examination alongside an Australian school placement battery. Not only did actual placement decisions correlate fairly strongly with the accuracy of the Y/N vocabulary test (r= .6), the Y/N also showed relatively high and consistent correlations with the listening and grammar measures of the placement battery, at .7 and .8 respectively. Mochida and Harrington interpreted these results as indicating that vocabulary recognition ability underlies each of the test battery's components. These findings,

coupled with the time and cost efficiency of the computerized Y/N test suggest a remarkable prospect: that an objective test of vocabulary ability might be capable of effectively discriminating between learners for placement purposes more efficiently than a traditional four-skills battery. Therefore, perhaps the Y/N test is more usefully studied as a measure of overall language proficiency than as an estimator of vocabulary size. Further work in this area would be of significant benefit to the field of language acquisition research, and of tremendous insight for those particularly interested in the importance of SL/FL vocabulary.

Additionally, and because size is merely one aspect of the nuanced construct of "vocabulary," further studies of a similar nature as that of Zareva, Schwanenflugel and Nikolova (2005), in which proficiency is investigated alongside the dimensions of word "familiarity" (which resembles some aspects of the Y/N task) and word association, could likely lead to a better understanding of what can be said regarding the nature of the interaction between vocabulary and language ability. Because the lexicon forms the basis by which ideas are conveyed, it would seem intuitive that increasing a learner's vocabulary size would expand her opportunities for practicing the expression of a broader range of ideas in the target language. Further investigation of vocabulary assessments such as the Y/N test may also inform our pedagogical practices with regard to the teaching of vocabulary, that is, how to develop the most effective methods by which to help learners encounter, incorporate, and retain lexical items in the target language. Thus, the continued study of vocabulary and its assessment remains tremendously valuable for our overall understanding of language acquisition.

REFERENCES

Abraham, R.G. & Chapelle, C.A. (1992). The meaning of cloze test scores: an item difficulty perspective. *Modern Language Journal 76*, 468-479.

Alderson, C.J. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly 13*, 219-227.

Alderson, C.J. (1980). Native and nonnative speaker performance on cloze tests. *Language Learning 30*, 59-76.

Alderson, C.J. (2000). *Assessing reading*. Cambridge, U.K.: Cambridge University Press.

Anderson, R.C. & Freebody, P. (1983). Reading comprehension and the assessment and acquisition of word knowledge. In B.A. Hutson, (Ed.), *Advances in reading/language research.* Greenwich: JAI Press.

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

Bachman, L.F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *TESOL Quarterly 19* (3), 535-556.

Bachman, L.F. (2005). Building and supporting a case for test use. *Language Assessment 2,* 1-34.

Bachman, L.F. & Palmer, A.S. (2000). *Language Testing in Practice*. Oxford, U.K.: Oxford University Press.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing 27*, 101-118.

Beeckmans, R., Eyckmans, J., Janssens, V., Dufranne, M., & van de Velde, H. (2001). Examining the Yes/No vocabulary test: some methodological issues in theory and practice. *Language Testing 18* (3), 235-274.

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

Brown, J.D. (1993). What are the characteristics of *natural* cloze tests? *Language Testing 10*, 93-115.

Brown, J.D. & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly 32*, 653-675.

Brown, G., Malimkjaer, K., & Williams, J. (Eds.). (1996). *Performance and competence in second language acquisition.* Cambridge: Cambridge University Press.

Burnett, J. (1973). *Muchas facetas de México.* Skokie, IL: National Textbook Company.

Carter, R.A. (1998). *Vocabulary: applied linguistic perspectives.* London: Routlage.

Carroll, J.B., Carton, A.S. & Wilds, C. (1959). *An investigation of 'cloze' items in the measurement of achievement in foreign languages.* College Entrance Examination Board Research and Development Reports. Laboratory for Research Instruction, Graduate School of Education, Harvard University.

Carroll, J.B. (1972). Defining language comprehension: some speculations. In J.B. Carroll & R.O. Preedle (Eds.), *Language comprehension and the acquisition of knowledge.* (pp. 1-26). New York: Wiley.

Chapelle, C.A. (1994). Are C-tests valid measures for L2 vocabulary research? *Second Language Research 10*, 157-187.

Chapelle, C.A. & R.G. Abraham. (1990). Cloze method: what difference does it make? *Language Testing 7*, 121-146.

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

Clark, D.F. & Nation, I.S.P. (1980). Guessing the meanings of words from context: strategy and Techniques. *System 8* (3): 211-20.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Davies, M., & Face, T.L. (2006). Vocabulary coverage in Spanish textbooks: how representative is it? In N. Sagarra & A.J. Toribio (Eds.), *Selected Proceedings of the 9th Hispanic Linguistics Symposium* (pp. 132-143). Somerville, MA: Cascadilla Proceedings Project.

Davies, M. (2005). *A frequency dictionary of Spanish: core vocabulary for learners.* London: Routledge.

Davies, M. (2005). Vocabulary range and text coverage: insights from the forthcoming Routledge Frequency Dictionary of Spanish. In D. Eddington (Ed.), *Selected Proceedings of the 7th Hispanic Linguistics Symposium.* Somerville, MA: Cascadilla Proceedings Project, 106-115.

Eyckmans, J. (2004). *Measuring receptive vocabulary size: reliability and validity of the Yes/No Vocabulary Test for French-speaking learners of Dutch* (Doctoral dissertation, Radbound University Nijmegen, 2004). Retrieved from http://repository.ubn.ru.nl/.

European Dialang Project. http://www.dialang.org

Eyckmans, J., Van de Velde, H., van Hout, R., & Boers, F. (2007). Learners' response behavior in Yes/No vocabulary tests. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.),

Modeling and assessing vocabulary knowledge. (pp. 59-76). Cambridge, U.K.: Cambridge University Press.

Farhady, H., & Keramati, M.N. (1996). A text-driven method for the deletion procedure in cloze passages. *Language Testing 13*, 191-207.

Foley, J. (1983). More questions on assumptions about cloze testing. *RELC Journal 14*, 57-69.

Green, D.M., & Swets, J.A. (1996). *Signal detection theory and psychophysics.* New York: Wiley.

Harrington, M., & Carey, M. (2009). The on-line Yes/No test as a placement tool. *System 37*, 614-626.

Hazenberg, S., & Hulstijn, J.H. (1996). Defining a minimal receptive second-language vocabulary for non-native university students: an empirical investigation. *Applied Linguistics 17*, 145-163.

Hoshino, Y. (1991). A bias in favor of the positive response to high-frequency words in recognition memory. *Memory and Cognition 19* (6), 607-616.

Hughes, A. (2003). *Testing for language teachers.* Cambridge, U.K.: Cambridge University Press.

Huibregtse, I., Admiraal, W., & Meara, P. (2002). Scores on a yes-no vocabulary test: correction for guessing and response style. *Language Testing 19* (3), 227-245.

Irvine, P., Atai, P., & Oller, J. (1974). Cloze, dictation, and the Test of English as a Foreign Language. *Language Learning 24* (2), 245-252.

Klein-Braley, C. (1985). A cloze-up on the C-Test. *Language Testing 2,* 76-104.

Klein-Braley C., & Raatz, U. (1984). A survey of research on the C-Test. *Language Testing 1* (2), 134-136.

Kobayashi, M. (2002). Cloze tests revisited: exploring item characteristics with special attention to scoring methods. *Modern Language Journal 86* (4), 571-586.

Koda, K. (1996). L2 word recognition research: a critical review. *The Modern Language Journal 80* (5), 450-460.

Lado, R. (1961). *Language testing.* London: Longman.

Lado (1986). Analysis of native speaker performance on a cloze test. *Language Testing 3* (2), 130-146.

Laufer, B., Elder, C., Hill, K., & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing 21* (2), 202-226.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics 16* (3), 307-322.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing 16*, 33-51.

Lee, S.H. (2008). Beyond reading and proficiency assessment: the rational cloze as stimulus for integrated reading, writing, and vocabulary instruction in ESL. *System 36*, 642-660.

Markham, P.L. (1985). The rational deletion cloze and global comprehension in German. *Language Learning 35* (3), 423-430.

Meara, P. (1990). Some notes on the Eurocentres vocabulary tests. In J. Tomola (Ed.), *Foreign language comprehension and production.* (pp. 103-113). Turku: AFinLa Yearbook.

Meara, P. (1992). EFL Vocabulary Tests. Swansea University.

Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer, &

J. Williams (Eds.), *Performance and competence in second language acquisition*. (pp. 35-53). Cambridge, U.K.: Cambridge University Press.

Meara, P. (2005). Designing vocabulary tests for English, Spanish and other languages. In C.S. Butler, M. Gómez-González, & S.M. Doval Suárez (Eds.), *The dynamics of language use.* (pp. 271-286). Philadelphia: John Benjamins.

Meara, P. (2008). Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Learning Abstracts 13*, 221-246.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing 4*, 142-154.

Meara, P., & Jones G. (1988). Vocabulary size as a placement indicator. In P. Grunwell (Ed.), *Applied linguistics in society.* Papers from the annual meeting of the British Association for Applied Linguistics, 3. Nottingham, U.K.

Meara, P., & Jones, G. (1990). Eurocentres vocabulary size test 10KA. Zurich: Eurocentres.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Mochida, A., & Harrington, M. (2006). The Yes/No test as a measure of receptive vocabulary knowledge. *Language Testing 23*, 73-98.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* New York: Newbury House.

Nation, I.S.P. (1993). Using dictionaries to estimate vocabulary size: essential, but rarely followed, procedures. *Language Testing 10*, 27-40.

Nation, I.S.P. (2001). *Learning vocabulary in another language.* Cambridge, U.K.: Cambridge University Press.

Nattinger, J.R., & DeCarrico, J.S. (1992). *Lexical phrases and language teaching.* Oxford, U.K.:
Oxford University Press.

Nuttal, C. (1982). *Teaching reading skills in a foreign language*. London: Heinemann.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory*. New York: McGraw-Hill.

Oller, J.W. (1972). Scoring methods and difficulty levels for cloze tests of proficiency in English as a second language. *Modern Language Journal 56* (3), 151-158.

Oller, J.W. (1973). Cloze tests of second language proficiency and what they measure. *Language Learning 23*, 105-118.

Oller, J.W. (1975). Cloze, discourse and approximations to English. In M.K. Burt and H.C. Dulay (Eds.), *New directions in second language learning, teaching, and bilingual education.* (pp. 345-355). Washington, DC: Teachers of English to Speakers of Other Languages, Inc.

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

Oller, J.W., & Conrad, C.A. (1971). The cloze technique and ESL proficiency. *Language Learning 21*, 183-194.

Oller, J.W., & Inal, N. (1971). A cloze test of English prepositions. *TESOL Quarterly 5* (4), 315-326.

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.

Pawly, A., & Syder, F.H. (1983). Two puzzles for linguistic theory: nativelike selection and nativelike fluency. In J.C. Richards and R.W. Schmidt (Eds.), *Language and Communication.* London: Longman.

Richards, J.C. (1976). The role of vocabulary teaching. *TESOL Quarterly 10*, 77-89.

Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal 10*, 12-18.

Read, J. (2000). *Assessing vocabulary.* Cambridge, U.K.: Cambridge University Press.

Read, J., & Chapelle, C.A. (2001). A framework for second language vocabulary assessment. *Language Testing 18*, 1-32.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford, U.K.: Oxford University Press.

Shillaw, J. (1996). The application of Rasch modeling to Yes/No vocabulary tests. Retrieved from http://www.lognostics.co.uk/vlibrary/shillaw1996.doc.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing 1* (2), 147-170.

Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing 18* (4), 373-391.

Singleton, D., & Little, D. (1991). The second-language lexicon: some evidence from university-level learners of French and German. *Second Language Research 7*, 61-81.

Stansfield, C. (1980). The cloze procedure as a progress test. *Hispania 63* (4), 715-718.

Stansfield, C., & Hansen, J. (1983). Field dependence-independence as a variable in second language cloze test performance. *TESOL Quarterly 17*, 29-38.

Stubbs, T., & Tucker, G. (1974). The cloze test as a measure of English proficiency. *Modern Language Journal 58* (5-6), 239-241.

Taylor, W.L. (1953). Cloze procedure: a new tool for measuring readability. *Journalism Quarterly 30*, 415-433.

Thorndike, E.L. (1924). The vocabularies of school pupils. In J.C. Bell, (Ed.), *Contributions to education*. (pp. 69-76). New York: World Book Co.

Van Heuven, J.B.W., Dijkstra, T., & Grainger, J. (1998). Orthographic neighborhood effects in bilingual word recognition. *Journal of Memory and Language 39*, 458-483.

Wesche, M., & Paribakht, T.S. (1996). Assessing second language vocabulary knowledge: depth versus breadth. *Canadian Modern Language Review 53*, 13-40.

Yule, G., Yanz, J.L., & Tsuda, A. (1985). Investigating aspects of the language learner's confidence: an application of the theory of signal detection. *Language Learning 35* (3), 473-488.

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

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: variable sensitivity. *Studies in Second Language Acquisition 27*, 567-595.

Zhang, L. J., & Annual, S.B. (2008). The role of vocabulary in reading comprehension: the case of secondary school students learning English in Singapore. *RELC Journal 39*, 51-76.