5-2010

Toward the development of a model to estimate the readability of credentialing-examination materials

Barbara Anne Badgett
University of Nevada Las Vegas

TOWARD THE DEVELOPMENT OF A MODEL TO ESTIMATE THE READABILITY OF CREDENTIALING-EXAMINATION MATERIALS

by

Barbara A. Badgett

Bachelor of Science
University of Nevada, Las Vegas
2000

Master of Science
University of Nevada, Las Vegas
2003

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy in Educational Psychology
Department of Educational Psychology
College of Education

Graduate College
University of Nevada, Las Vegas
May 2009

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

THE GRADUATE COLLEGE

We recommend the dissertation prepared under our supervision by

Barbara Anne Badgett

entitled

Toward the Development of a Model to Estimate the Readability of Credentialing-Examination Materials

be accepted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Educational Psychology

Alice J. Corkill, Committee Chair

Gregory Schraw, Committee Member

CarolAnne Kardash, Committee Member

Mark Ashcraft, Graduate Faculty Representative

Ronald Smith, Ph. D., Vice President for Research and Graduate Studies
and Dean of the Graduate College

May 2010

ABSTRACT

Toward the Development of a Model to Estimate the Readability of Credentialing-Examination Materials

by

Barbara A. Badgett

Dr. Alice J. Corkill, Examination Committee Chair
Professor of Educational Psychology
University of Nevada, Las Vegas

The purpose of this study was to develop a set of procedures to establish readability, including an equation, that accommodates the multiple-choice item format and occupational-specific language related to credentialing examinations. The procedures and equation should be appropriate for learning materials, examination materials, and occupational materials. To this end, variance in readability estimates accounted for by combinations of semantic and syntactic variables were explored, a method was devised to accommodate occupational-specific vocabulary, and new-model readability formulas were created and calibrated. Existing readability formulas were then recalibrated with the same materials used to calibrate the new-model formulas. The new-model and recalibrated formulas were then applied to sample items extracted from a professional licensing examination and the results were compared.

ACKNOWLEDGEMENTS

I would like to thank Dr. Jack Gerrow Jack D. Gerrow, D. D. S., Marcia A. Boyd, D. D. S., and the National Dental Examination Board of Canada for their continued support. Without the access they provided to curricular and occupational resources as well as examination data, this study would not have been possible. Dr. Boyd's subject matter expertise proved invaluable.

Many thanks also go out to Dr. McCrudden, who directed me to the Buros Center for Testing and introduced me to Dr. Buckendahl. During my internship at Buros, which was granted to me by Dr. Buckendahl, he passed along a research article and said something to the effect of, "Perhaps you can extract an idea for your dissertation from this". Well, Dr. Buckendahl…I was, in fact, able to find a research idea and thank you so very much for your help. Dr. Buckendahl provided further support by connecting me with the National Dental Examination Board of Canada. Without your support, this investigation would not have been conducted. You pointed me in a direction and when I finally developed a research idea, you provided a means for me to collect the necessary information to bring the idea to full fruition. I am at a loss for words to appropriately express my gratitude for your guidance, support, and patience.

I would also like to express my gratitude to my dissertation committee, Gregory Schraw, Ph. D., CarolAnne Kardash, Ph.D., Mark Ashcraft, Ph.D. and Alice J. Corkill, Ph.D., who provided very helpful feedback throughout this process. At the time of the proposal, you offered insight that helped me to better scope the investigation and identified additional steps that were necessary. This resulted in a better-planned study. Thank you for your time and effort.

Dr. Corkill, my dissertation committee chairperson, advisor, friend, and shepherdess deserves very special thanks. You will never know how much it meant to me that you were willing to sacrifice months of weekends to read, re-read, and provide feedback...only to read, re-read, and provide more feedback. Not many advisors are willing to answer panicked phone calls at all hours of the night to calm the nerves of their hysterical advisees. Thank you for all the encouragement….for assuring me repeatedly that I would eventually succeed. It made such a difference and really helped me keep up my spirits and confidence. What, pray tell, will you do with all of your free time now that you are not inundated with my pestering emails, attachments, phone calls, and text messages? I have learned so much from you over the years...not only about educational psychology…but about life. Thank you for being there...time after time.

I would also like to thank Dr. Smith and Dr. Davis for the assistance and support they offered during this process. Dr. Smith, without your macros, I would still be counting words. Dr. Davis, thank you for bouncing around ideas with me early in the process.

Finally, I would like to thank my family and friends. Thank you for understanding that I had to go MIA for a while and why I have not "been there"…for understanding why I have missed so many important occasions. Mom, Daddy J, and Dannette, thank you for helping me sort out those silly word lists. It was very helpful.

My parents and fiancé have offered the most valued support during this very long process. How many times did you talk me off the ledge? How many times did you listen to me weep and express doubt? You always responded with words of encouragement and helped me realize that I could "do it". After each of those dozens of conversations, I was able to pick myself up and push forward. I don't know whether I could have done this

without your love, support, and patience. I don't know whether I could have done this without knowing that you would be proud, regardless of the outcome.

Brian, I especially thank YOU for your patience. Now we can get married!!

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Tests are designed to measure constructs of interest. In order to have confidence that a test score represents the construct of interest tests should be free of unnecessary construct irrelevant variance. One source of construct irrelevant variance is related to the readability of testing materials. Readability refers to the ease with which readers are able to read and comprehend a written text. The values obtained with readability measures reflect the reading difficulty level of a text. Readability of testing materials has received little attention and there is currently no industry-established method for establishing the readability of test items. The following sections include discussions regarding the importance of considering this source of construct irrelevant variance in a particular testing situation: credentialing examinations (i.e., licensing and certification examinations).

The introduction is organized around three main sections: 1) Readability, 2) Readability in Testing, and 3) Readability of Licensure and Certification Examinations. In the first section, readability is defined and a general overview is provided regarding how readability is measured and the variables that are considered. The second section includes a discussion of issues related to applying readability formulas to tests. In the third section, the purposes of licensure and certification examinations and the differences between them are outlined. Issues related to measuring the readability of licensing- and certification-examination items and why their readability levels should be measured are addressed next. Then, a brief discussion is provided regarding the impetus for the current investigation: a model proposed by Plake (1988) that asserts that materials related to

licensure or certification examinations should have the same readability levels as the examinations themselves.

Readability

Readability is a construct related to comprehensibility or the "ease with which a reader can read and understand" a given text (Oakland & Lane, 2004, p.244). The optimal readability level of a text is one that corresponds with, or does not exceed, the reading ability of the reader. When readability levels of texts exceed the reading ability of readers, the readers are likely unable to adequately decipher the author's intended message.

A variety of mathematical equations derived through regression techniques have been developed to assess readability (McLaughlin, 1969). These readability formulas, which typically consist of predictor variables combined with constants, offer a means of quantifying the reading ability that is required for an individual to comfortably read and understand a given text (Felker, 1980; Redish & Selzer, 1985; Stokes, 1978). These readability measures are also used to rank reading materials in terms of difficulty (Fry, 2002).

Readability formula results are reported as numerical indices. The indices from several readability formulas are reported in terms of grade level (e.g., Dale-Chall, 1948, 1995; FOG, 1952; FORCAST, 1973; Fry, 1965; Harris-Jacobson, 1974; SMOG, 1969; Spache, 1953). Results from other formulas represent difficulty levels on a scale (e.g. Flesch, 1948 & Lexile, 1987).

Scholars have investigated the predictive power of syntactic and semantic variables for estimating readability (DuBay, 2004; Fry, 2002; Klare, 1963; Oakland & Lane, 2004;

Sharrocks-Taylor & Hargreaves, 1999; Sydes & Hartley, 1997). Syntactic variables most often addressed include: 1) average sentence length (as measured by the number of letters, syllables, or words); 2) number of personal sentences (e.g., quotes, questions, commands, requests, or other sentences directed at the reader); 3) number of personal references; 4) number of sentences per passage; and 5) number of prepositional phrases. Semantic variables most commonly investigated include: 1) average word length (as measured by letters and syllables); 2) number or percentage of difficult words (difficult words are identified by determining whether they are included in familiar word lists such as *The Dale-Chall list of 3,000 familiar words*, 1943; or *The Living Word Vocabulary*, Dale & O'Rourke, 1976, 1981); 3) number of personal pronouns; 4) number of elemental words (i.e., words that are essential to the meaning of the sentence); 5) number of monosyllabic words; 6) number of words with three or more syllables; 7) number of words including affixes; 8) number of personal words; 9) percentage of concrete words; 10) percentage of abstract words; 11) percentage of polysyllabic words; and 12) percentage of simple localisms. Of these syntactic and semantic predictor variables, sentence length, word length, and the percentage of difficult words (vocabulary) have shown to be the most powerful in estimating readability (Stenner & Burdick, 1997). Below are two of the more popular and widely used readability formulas:

Flesch-Kincaid Grade level (US Navy, 1976) = .39 (*wl*) + 11.8 (*sl*) – 15.59

(Where *wl* = word length and *sl* = sentence length)

Dale-Chall Cloze (Chall, 1995) = 64 – (.95) (X_1) – (.69) (X_2)

(Where X_1= number of unfamiliar words and X_2=average sentence length.)

Although readability formulas are useful for determining text difficulty, not all texts lend themselves well to the formulas because the formulas generally require several 100-word passages for proper implementation (Allan, McGhee, & van Krieken, 2005; DuBay, 2004; Hewitt & Homan, 2004; Homan, Hewitt, & Linder, 1994; Klare, 1984; Oakland & Lane, 2004). Readability formulas do not yield valid results for materials such as multiple-choice test items or documents with long word lists (Allan, McGhee, & van Krieken, 2005; Hewitt & Homan, 1991, 2004; Homan, Hewitt, & Linder, 1994).

Popham (1981) was one of the first researchers to address the need for a readability measure useful for estimating the readability of individual sentences. He developed the *Basic Skills Word List* to assign words to grade levels for a set of basic skills tests. The criteria he used to devise the word list were as follows: 1) word frequency in published reading texts, 2) word frequency in general reading material, and 3) readers' familiarity with particular words (according to Dale and O'Rourke's Living Word Vocabulary, 1976). Although Popham did not develop a readability formula, his was one of the first concerted efforts to address the readability of individual sentences and test items (Hewitt & Homan, 1991).

Homan and Hewitt (2004) as well as Homan et al. (1994) also worked to develop a method for estimating the readability of individual sentences and phrases. The authors created and validated the Homan and Hewitt readability formula for single sentences that occur in multiple-choice tests at 2nd- through 5th-grade levels. Hewitt and Homan (2004) further investigated the use of the Homan and Hewitt readability formula and the relationship between item difficulty and readability with their examination of social studies items from a major standardized test.

The Homan-Hewitt formula includes three predictor variables: 1) number of difficult words (WUNF), 2) word length (WLON), and 3) sentence complexity (WNUM). Difficult words are identified as those not included in *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981). Word length is established by counting words per sentence that include more than six letters. Sentence complexity is determined by establishing the average number of words per Hunt's T-Unit. Hunt's T-Unit is a measure of syntactic complexity that considers the number of clauses per sentence. The resulting formula is:

$$Y = 1.76 + (.15 \text{ X WNUM}) + (.69 \text{ X WUNF}) – (.51 \text{ X WLON}).$$

Although Homan et al. (1994) published validation results for their readability formula designed for use with multiple-choice test items; it has not been adopted for use with standardized tests. No researchers, other than the developers, have published or presented studies using the formula (databases queried include ERIC Ebsco, Eric First Search, and Pychinfo). Test manuals seldom include estimates of item readability or information regarding the methods used to design and develop items (Homan et al., 1994). It appears that the Homan-Hewitt formula is the only formula that has been specifically designed for use with single-sentence, multiple-choice questions.

The Homan-Hewitt formula was designed for and validated with materials appropriate for elementary school-age children. Therefore, it would not be considered appropriate for use with adult-level reading material. Nevertheless, the variables and methods that Homan and Hewitt (1994, 2004) used to develop the formulas might offer valuable information for the development of a formula suitable for multiple choice items written for other populations.

Readability in Testing

Although the Homan-Hewitt formula, according to validation study results (Hewitt & Homan, 2004; Homan et al., 1994), may offer useful information about the readability of multiple-choice items, readability is not typically formally addressed in the development of high-stakes tests. More traditional readability measurement approaches are not appropriate for use with test items. Test items are typically constructed to be concise. Multiple-choice items, for instance, include stems that are usually between one and three sentences long with response options that are shorter. The length of test items inhibits accurate estimations of readability because readability formulas generally require several 100-word samples for reliable evaluation.

It is not useful to simply combine test items into a single continuous prose segment in order to meet the length requirement of readability formulas for two reasons. First, prose subjected to readability formulas should be continuous and test items are distinct pieces of text. Second, if items were combined to create quasi-continuous prose of appropriate length and a traditional readability formula were applied, it would be impossible to determine the readability levels of individual items. Instead, the readability index obtained would offer an overall estimate of the entire instrument (Homan et al., 1994). This would make it inappropriate to use the results to identify the readability levels of specific items.

Failure to consider the readability of test items can pose a critical problem in high-stakes, standardized testing. Specifically, without the assessment of the readability of test items, the test developer risks creating items that do not properly correspond to the reading abilities of examinees for whom the test is intended. If the readability level of a

test item is beyond the reading ability of an examinee, the item is not likely to solely measure the construct of interest; instead, it likely also measures examinee reading ability. In other words, a test item with a particularly high readability level will require that a candidate have reading comprehension skills that enable him/her to effortlessly decipher the intended message. If the candidate does not have reading comprehension skills that correspond to the readability level of the test item, the item measures dual constructs: the construct of interest and reading comprehension. Unless the construct of interest is, in fact, reading ability, incongruence between readability and reading ability introduces a critical, irrelevant confound in the measurement of the construct of interest. This, then, becomes an additional source of measurement error (Cronbach, 1980; Plake, 1988). For example, if a mathematical word problem includes text at an inappropriate readability level for examinees, it no longer simply measures their ability to solve word problems; it also measures examinee reading ability. Therefore, examinees who have the ability to correctly solve a variety of word problems, but have poor reading comprehension skills, may fail to select the correct response because they are unable to understand the details of the text. This would result in different test performance outcomes for examinees with similar mathematical skill levels but with different reading ability levels. The higher reading ability examinees would have an advantage over examinees with lower reading ability due to a construct-irrelevant skill, which would negatively affect the validity of the results (Plake, 1988).

Readability of Licensure and Certification Examinations

Credentialing examinations used for licensure or certification generally serve "gate-keeping" purposes (Plake, 1988, p.543). Passing scores are required for examinees to be

allowed to perform particular jobs or tasks. These examinations are essential in order to maintain public safety. Appropriate correspondence between the readability of test items and the reading ability of examinees is, therefore, especially important for licensure or certification, high-stakes examinations. Examinees should have the reading comprehension skills necessary to effectively read and decipher texts used during instruction and job practice. It follows, then, that the readability levels of instructional materials, credentialing examination items, and job related materials should be congruent.

Licensure and certification examinations are used to license and certify, respectively, people to practice particular professions. Both types of credentialing examinations are designed to ensure that prospective practitioners possess the appropriate knowledge, skills, and abilities to practice their professions. The principal purposes of these measures are to maintain public safety and provide service patrons some confidence in the capabilities of practitioners (Downing, 2006).

Certification and licensure examinations are different in that licensure is generally granted by the state, whereas a professional organization or board generally grants certification. In addition, licensure is typically mandatory; certification can be mandatory or voluntary (Downing, 2006). Permission to legally practice professions or occupations such as medicine, dentistry, and cosmetology require licensing. Certification is generally required for an individual to practice a specialty within the field in which he/she is licensed (Downing, 2006). A clear distinction between the uses of the two types of examinations can be illustrated with an example from the medical field. Dermatologists must take a licensure examination to become licensed to practice dermatology in their state. They may then take additional courses or attend seminars to learn how to use the

newest laser skin-treatment devise. After such a mini-course they might take a certification examination and upon passing would be certified to use the laser in their practice.

Credentialing examinations, like other high-stakes tests, are often largely comprised of multiple-choice items. Unfortunately, the format of multiple-choice test items prevents them from being well suited for the use of readability formulas. Readability estimations of credentialing examination items are further impeded by discipline-specific technical language (Allan, McGhee, & van Krieken, 2005). For example, imagine that the Homan-Hewitt formula were applied to items from a licensure examination designed for registered nurses. Words such *tracheoesophageal* would artificially inflate readability estimations. This is because readability formulas, including the Homan and Hewitt readability formula (1994), are specifically designed to be sensitive to semantic variables such as word length and vocabulary. *Tracheoesophageal* is a lengthy, polysyllabic word and certainly not included in *The Living Word Vocabulary* list of common words (Dale & O'Rourke, 1981). The especially high readability estimates would be appropriate if the test were taken by examinees without medical backgrounds, but the test is designed for examinees with extensive medical knowledge. Any person who takes a licensure examination to become a nurse is, or should be, familiar with such terms. Therefore, valid measures of readability should not be affected by such domain-specific vocabulary.

Although, to date, there are no external criteria available to identify the level at which certification and licensure examinations should be written, Plake (1988) asserts that readability checks should be included in the validation process of those examinations. This is because construct-irrelevant variance due to inappropriate levels of reading

difficulty poses a potential threat to the validity of credentialing examination results. When items are written at readability levels above which candidates are able to comprehend, the language has the potential to hinder candidate performance based on constructs irrelevant to what the examination is designed to measure. Credentialing examinations, aside from technical language, should have difficulty levels low enough to ensure that anyone qualified to do the job in question is able to read and understand the items.

According to Plake's *Model for evaluating the readability level of a licensure/certification examination for a trade profession* (1988), readability of credentialing examinations in a trade profession should correspond to materials that are necessary for job performance. This is in accordance with *Standard 9.8* of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999, p. 99), which reads, "In employment and credentialing testing, the proficiency level required in the language of the test should not exceed that appropriate to the relevant occupation or profession." Plake also contended that the readability level of curriculum or learning materials used in necessary educational or training programs should correspond to the readability of the respective credentialing examination. This notion is supported by Downing (2006), who asserts that to offer acceptable validity evidence, the content of a credentialing examination should be determined with attention to curricular documents, teaching syllabi, instructional materials and content, and textbook content—as well as other relevant sources.

Plake (1988) holds that learning, testing, and occupational materials should have equal readability levels. Unmatched levels of readability among materials could open the

door for candidate appeal. She asserts that incongruence can occur in one of two ways. First, students might be assessed with language that is more difficult to read and understand than the materials with which they were taught. Second, language used in a test might be at a higher level than is required by the occupation or profession. In light of Plake's model, both cases involve the introduction of avoidable construct-irrelevant variance. If the creators of certification/licensure examinations do not adequately address issues of examination readability, the validity of the results may, and perhaps should, be questioned.

In summary, readability essentially reflects the difficulty level of a given text and the reading ability level required to comprehend that text. Various formulas have been developed to quantify readability of continuous prose according to semantic and syntactic variables. To date, high-stakes test development does not involve formal measures of test item readability, most likely because no well-established formula appropriate for use with individual multiple-choice items is available.

Readability estimates for licensure or certification examination items are necessary to establish that student learning materials, examination materials, and occupational materials are of equivalent readability levels. Before the readability levels of credentialing examination items can be considered, however, a process designed to accommodate the multiple-choice item format and occupational-specific language must be developed. Until a method is created that is capable of accommodating credentialing examination format and content, the relationship between learning materials, examination items, and occupational material readability levels is a moot point. The first step in

investigating these relationships, therefore, is to design a process for measuring the readability of credentialing examination items.

The goal of this investigation was to develop a set of procedures to establish readability, including an equation, that accommodates the multiple-choice item format and occupational-specific language related to credentialing examinations. The procedures and equation should be appropriate for learning materials, examination materials, and occupational materials. If successful, the new-model would offer a means for investigating and comparing readability levels of credentialing-related learning, examination, and occupational materials.

Establishing equivalence in readability levels across the materials would offer credentialing programs additional evidence that respective examinee exam scores are valid representations of the constructs of interest. Specifically, equivalence in readability levels across the materials would suggest that unnecessary measurement error introduced via construct-irrelevance variance due to inappropriate readability levels of the examination items would not likely be a matter of concern. In contrast, determining that the readability levels of examination items are greater than the readability levels of either the learning or occupational materials would potentially inform a credentialing program's future item-development efforts.

CHAPTER 2

LITERATURE REVIEW

The concept of readability and approaches to measuring it has received substantial attention throughout the 20th century. The following sections include discussions of readability. The first section includes a definition of readability and descriptions of its more popular uses. The methods used over the years to calibrate readability measures are described in the second section. The third section includes a history of readability research and formula development conducted by reading researchers as well as a description of a readability measure devised by measurement scholars. The last section includes an explanation of the need for a readability formula suitable for use with test items.

What is Readability? Definitions and Popular Uses

In this section, the concept of readability and readability formulas is introduced. First, readability and readability formulas are defined and examples are offered of readability scholars' definitions of each. Second, an explanation is offered regarding the manner in which the results of readability formulas are reported. Third, the reading levels targeted by formulas are discussed. Finally, some of the specific uses for which readability formulas have been developed are outlined.

Readability Defined

Readability is a construct related to the comprehensibility of a given text. Definitions of readability vary slightly among scholars; but the gist of the definitions is the same. Readability generally refers to the reading difficulty level of a text. It is affected and determined by the elements that influence a reader's comprehension (Dale & Chall,

1949). Readability formulas are mathematical equations designed to predict and quantify the reading ability required for a reader to understand a text (Felker, 1980; Stokes, 1978). The results enable the ranking of reading materials in order of difficulty (Fry, 2002).

Reporting Readability Formula Results

Readability formula results are reported as numerical indices. Some readability formula results are reported in terms of grade levels (e.g., Dale-Chall, 1948, 1995; FOG, 1952; FORCAST, 1973; Fry, 1965; Harris-Jacobson, 1974; SMOG, 1969; Spache, 1953). Results from other formulas are represented as difficulty levels on a scale. For example, results from the Flesch Reading Ease formula (1948) are reported on a scale from 0 to 100 with 100 representing the lowest level of reading difficulty. Results from the Lexile Framework are reported on a scale from 0 to 2,000, where higher Lexile values reflect higher levels of reading difficulty.

Readability Formula Targets

Different readability formulas were designed to estimate the readability of written materials for audiences at particular ability levels. For instance, the FOG formula (Gunning, 1952), FORCAST formula (Caylor, Stitch, Fox, & Ford, 1973), and the Army's Automated Readability Index (ARI; Smith & Senter, 1967) were developed specifically for use with adult-level materials. The Dale-Chall formula (1948) and the Flesch Reading Ease formula (1948) were developed to identify appropriate levels of difficulty for readers from 4th-grade to adult. The Fry Readability Graph (1968) was initially designed for primary and secondary school materials but through extrapolation was later extended to include preprimary levels. The Spache (1953) and Harris-Jacobson (1974) formulas were designed specifically for use with materials at preprimary levels.

Intended Uses

Readability formulas are often designed for specific uses. The formulas are used to determine and help select reading materials of appropriate difficulty levels for students (e.g., Spache, 1953; Harris-Jacobson, 1974; Fry, 1968). They have also been developed to determine the readability of technical and training materials intended for adult readership. For instance, the Boeing Company contracted Jablonski to devise a readability formula to determine the readability of their maintenance manuals (Klare, 1974-1975). After an extensive study of the reading demands of military occupational specialties, Caylor and Stitch (1973) developed the FORCAST formula for use with U.S. Army materials. Readability is also a concern for materials meant for the general adult population. DuBay (2004) reports that readability formulas have been cited in research related to: political literature, corporate annual reports, customer service manuals, drivers' manuals, dental health information, palliative-care information, research consent forms, informed consent forms, online health information, lead-poison brochures, online privacy notices, environmental health information, and mental health information.

Readability estimation is valuable to help ensure that readers are provided textual materials that correspond to their reading abilities. Without consideration of such alignment between text levels and reading ability, readers may not be able to comfortably read and understand the intended message of a given text. Therefore, congruence between reading materials and reader ability should be considered.

Calibration Methods for Readability Formula Development

Existing readability formulas were calibrated with the use of the McCall-Crabbs Standard Test Lesson in Reading (McCall & Crabbs, 1925, 1950, 1961) and the cloze

technique. The McCall-Crabbs criterion was used in the earlier years of readability research and has since been largely replaced by the cloze technique. These calibration methods are discussed in the following section along with an explanation as to why the cloze technique is now the calibration method of choice.

McCall-Crabbs

The following subsection includes as discussion of the first popular means by which readability formulas were calibrated, multiple-choice scores on the McCall-Crabbs *Standard Test Lessons in Reading* (McCall & Crabbs, 1925, 1950, 1961). First, the methods used to norm the passages are described. Second, the most popular formulas developed using these criteria are presented. Then an overview of how the McCall-Crabbs Standard Test Lessons in Reading were used as a criterion for formula development is offered. Finally, the shortcomings of the McCall-Crabbs Standard Test Lessons in Reading for use as a criterion in readability formula development are addressed.

Norming passages.

Readability formulas are often developed using text passages of known difficulty. McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs, 1925, 1950, 1961) offers grade-level scores against which numerous early readability formulas were measured. McCall and Crabbs originally designed their test lessons in 1925 and renormed them in 1950 and 1961 (DuBay, 2004; Felker, 1980; Klare, 1984).

The initial 1925 grade-level assignments were created with the multiple-choice test results of 2,000 New York City school children on 376 text passages (Felker, 1980; Stevens, 1980). The test lessons were administered to grades three through six. Each text

passage was approximately 150 words and was followed by eight or ten multiple-choice questions. Grade-level equivalents for the passages were derived by the number of correct responses from students in a particular grade. For instance, two correct answers for a passage might result in that passage being given the grade level of 3.2 (second month of grade three); six correct responses might be equivalent to 6.4 (Felker, 1980). These normed text passages with grade-level assignments have been widely used to calibrate readability formulas.

Until about 1960, most readability formulas were developed using the McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs, 1925, 1950, 1961) as the criterion (Klare, 1984). Among those readability formulas are the Lorge formula (1939), Lorge formula revised (Tretiak, 1969), Dale-Chall formula (1948), Flesch reading ease formula (1948), Flesch reading ease formula revised (Powers, Sumner, & Kearl, 1958), Farr-Jenkins-Patterson formula (1951), Danielson-Bryan formula (1963), FOG formula(Gunning, 1952) and SMOG grading formula (1969; DuBay: 2004; Klare, 1974-1975; Olsen, 1986). McCall and Crabbs renormed the passages with new groups of children in 1950 and 1961 because of concern that the Standard Test Lessons in Reading results had become outdated and less useful (DuBay, 2004; Klare, 1974-1975, 1984; McCall & Crabbs, 1950, 1961). Several readability formulas that were originally calibrated based on the 1925 version were recalibrated based on the new criteria (e.g. Dale-Chall formula,1995; Farr-Jenkins-Patterson formula, 1958; Flesch reading ease formula, 1958; and Lorge formula revised, 1969).

Formula development using McCall-Crabbs as a criterion.

Readability formula developers who used the McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs, 1925, 1950, 1961) as their criterion constructed the formulas so that they predicted the average grade level of students who correctly answered a set percentage of multiple-choice questions for a passage. The set percentage of correct responses for the average grade levels varies by formula. The percentage-correct criterion for grade-difficulty-level assignments based on the McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs) for each of the formulas is: 50% with the Lorge formula (1939), Lorge formula revised (Tretiak, 1969), Dale-Chall formula (1948), and Flesch reading ease formula (1948; Powers, Sumner, & Kearl, 1958); 75% with the Farr-Jenkins-Patterson formula (1951); 90% with the FOG formula (Gunning, 1952); and 100% with the SMOG grading formula (1969).

Shortcomings of the McCall-Crabbs criterion.

According to Klare (1974-1975, p. 66) the use of McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs, 1925, 1950, 1961) was well-suited for readability formula calibration, "These lessons have been convenient statistically because there are a large number of reading passages, covering a wide range of difficulty, resting upon extensive testing, and providing detailed grading scores." When Dale and Chall developed their first readability formula they touted the McCall-Crabbs Standard Test Lessons in Reading as the best criteria available but also acknowledge that it has "serious deficiencies" (Dale & Chall, 1948, p. 15).

Critics have more specifically addressed the deficiencies to which Dale and Chall allude. McCall and Crabbs never published a guide or outline of how to use their

Standard Test Lessons in Reading for readability formula calibration because the instrument was not designed for such use (Stevens, 1980). Stevens corresponded with McCall about the use of the McCall-Crabbs Standard Test Lessons in Reading (McCall & Crabbs, 1925, 1950, 1961) as a criterion for readability formula development and reported that McCall stated:

> When, last year, I learned for the first time the number of readability formulas resting on my G [grade level] scores, I was vastly surprised….Probably all the formulas were defensible during the rude early days of scientific education. The formulas builders never approached me as you have done. (p. 414)

According to Stevens—based upon her correspondence with McCall—the authors never intended the test lessons to be used for formula development or extensive testing. Instead, these lessons were meant for use as a practice exercise in reading (Stevens, 1980). Crabbs and McCall (1925) offer a more specific description of their intended use of the standard test lessons for students (p.1-3):

1) Teach them how to comprehend rapidly all kinds of materials
2) Help them enjoy their reading lessons
3) Make it easier for them to learn their other lessons
4) Test and teach them at the same time
5) Test them with a standard test
6) Automatically indicate their proper grade classification in reading
7) Teach them how to read carefully and accurately
8) Teach them how to read for the main points, to judge the relative importance of the various ideas presented, to follow the sequence or thread of thought, to

reorganize material in order to answer questions that cut across this thread of thought

9) Teach them how to read as rapidly as they can understand what they read and to regulate their speed according to the purpose for which the reading is being done
10) Teach them how to skim
11) Enable them to score their own or each other's tests
12) Motivate and improve their oral expression
13) Provide them with opportunity for the practice of leadership
14) Help prevent the dull pupils from becoming discouraged and the bright pupils from loafing
15) Make it possible for them to appreciate more difficult literature, and literature of a wider range
16) Increase their joy in literature by reserving the appreciation period primarily for appreciation
17) Save their time.

McCall's response to Stevens' inquiry and the description of intended uses of the standard test lessons described by Crabbs and McCall brings into question the validity and reliability of the passages for formula development. The use of these passages for readability formula development has also been criticized on the simple basis of their design.

The McCall-Crabbs *Standard Test Lessons in Reading* (McCall & Crabbs, 1925, 1950, 1961) consists of four booklets (A-D) each comprised of approximately 70 graded

passages each. Each passage is followed by a set of multiple-choice questions. The grade-levels were assigned according to the number of correct responses by pupils of known reading achievement levels (Stevens, 1980). According to McCall and Crabbs' intended design, the books (A-D) are ordered according to difficulty (A is least difficult or contains the easiest reading passages) as are the passages within each book. This is a necessary characteristic if the passages are to be used for readability formula criterion. Olsen (1986) tested this assumption with six readability formulas that were designed with the McCall-Crabbs *Standard Test Lessons in Reading* (McCall & Crabbs, 1925, 1950, 1961) as the criterion: Flesch formula, Dale-Chall, FOG Index, SMOG index, Spache Index, and Wheeler-Smith. These formulas were applied to the first and last third of passages from each book. If the books and the passages within them were arranged according to difficulty, the results of formulas based upon them should have consistently indicated such. That was not the case. None of the formulas resulted in consistent within- or between-book progressions from least to most difficult. In addition, there were vast differences among some of the formula results for the reading selections (within and between books).

If formulas that were designed based on the McCall-Crabbs *Standard Test Lessons in Reading* (McCall & Crabbs, 1925, 1950, 1961) do not yield results consistent with the test lessons original design, there may be reason for considerable concern about the validity and reliability of the McCall-Crabbs *Standard Test Lessons in Reading* for use as a calibration instrument. In fact, it is no longer widely used as a criterion for readability formulas but not because of the issues mentioned here. It was replaced by a newer, more convenient method, Taylor's (1953) cloze technique (Klare, 1984).

The Cloze Technique

The following subsection includes an introduction of the *cloze technique*: a method for calibrating readability formulas that largely replaced the multiple-choice method discussed previously. First, its original development and validation are discussed. Then, research and advances in the use of the method as a means of calibration for readability formulas are described. Finally, a list of formulas that have been calibrated or recalibrated based on the cloze technique is provided.

In 1953, Taylor developed the cloze procedure for measuring the readability of text. The name "cloze" is a derivation of "closure", which is a term used in Gestalt psychology to refer to people's tendency to complete familiar patterns. Eventually this method largely replaced the use of the McCall-Crabbs *Standard Test Lessons in Reading* (McCall & Crabbs, 1925, 1950, 1961) for the calibration of readability formulas.

Considering how the method works, the name reflects it well. The cloze procedure involves deleting words from a text passage using a random-number system or by counting out every n^{th} (usually 5^{th}) word. A blank of standard length is placed in the position of the deleted words. Participants are then presented with the modified text passages and asked to fill in the blanks using the surrounding contextual clues. Cloze totals for each passage are derived by simply counting the number of blank spaces that are filled with the correct words. Synonyms are not counted as correct and misspellings are not counted as errors. Passages for which participants receive high scores are considered more readable and passages for which they receive low scores are deemed less readable. The cloze procedure differs from sentence-completion tests in that development of sentence-completion tests involves the deletion of pre-evaluated words so

that a person's knowledge of specific information can be assessed. The "cloze procedure deals with [a] contextually interrelated series of blanks, not isolated ones" (Taylor, 1953 p. 417). In addition, cloze does not deal with meaning; instead, its sampling procedure is gauged toward identification of language patterns.

Taylor (1953) was adamant that the cloze method is not a readability formula. It does not involve counting language elements that are thought to correlate with ease of comprehension. Although, he did claim that the procedure "measure[s] whatever effects elements actually may have on readability" (p.417).

Taylor (1953) conducted two experiments to test the cloze procedure as a measure of readability. For experiment one, Taylor used 24 juniors and seniors enrolled in journalism courses at the University of Illinois. He compared participant cloze scores for passages from Flesch's *How to Test Readability* (1951) to results from the Flesch formula (1948) and the Dale-Chall formulas (1948). The cloze procedure resulted in the same rankings of the passages as the readability formulas. In addition, analysis of variance results showed that cloze scores for each passage were significantly different from one another. Taylor concluded that the cloze procedure was measuring the same constructs as the readability formulas and showed sufficient power of discrimination.

Taylor (1953) conducted a second experiment as a follow-up to the first experiment. In the second experiment, the "cloze procedure was 'pitted' against those standard formulas" (Taylor, 1953, p. 415) with 72 subjects from the same population as the first experiment. Taylor added the following passages, which were thought to be difficult for the readability formulas to appropriately gauge, to the second experiment: Caldwell's *Georgia Boy*; Stein's *Geography and Plays*; Joyce's *Finnegan's Wake*; Swift's *Gulliver's*

Travels; and Dickens' *Bleak House*. Taylor believed that the Flesch and Dale-Chall formulas would inaccurately rank the passages taken from these texts. In a pilot study, six subjects were used to establish cloze predictions (median cloze scores). Scores from the second experiment for the cloze procedure, Flesch formula values, and Dale-Chall formula values were compared to these predicted cloze scores in terms of readability rankings. The cloze test rankings agreed perfectly with the predicted cloze test rankings obtained in the pilot study. The standard formulas agreed with one another relatively well with a rank correlation of .70 ($p < .05$). The results, however, did not significantly correlate with the predicted or experimental cloze test scores. In addition, analysis of variance between the experimental cloze scores showed that they were significantly different from one another. Taylor interpreted these results to substantiate those from the first experiment. In addition, he touted, "previous cloze results were more successful than those of the two standard formulas in predicting the ranks of future results for the population used" (p. 427). Although he wrote this as if it were quite an accomplishment, it seems fairly obvious that cloze procedure results would be expected to agree better with other cloze procedure results than those of other readability results. On the other hand, he points out that the cloze scores for prediction and those from the second experiment were derived from independent populations. With his two experiments, Taylor clearly illustrated that the cloze procedure is at least as accurate as the standard formulas in identifying or ranking the readability of text.

The cloze procedure devised by Taylor (1953) offered a viable means of gauging the readability of texts. This was later substantiated by Coleman (1965) who was the first to use the cloze technique instead of multiple-choice tests to develop a readability formula.

He devised four formulas that yielded multiple correlations of .86, .89, .90, and .91 with cloze criterion scores (DuBay, 2004).

A set of 36, 150-word passages calibrated for complexity by Miller and Coleman (1967) ended the need for participants in the development of readability formulas calibrated with the cloze technique. They enlisted 479 college students to complete cloze tests on the 36 passages, which ranged in difficulty from first-grade to difficult technical material. The majority of the prose passages were taken from McCall and Crabbs *Standard Test Lessons in Reading* (1925, 1950, 1961) and the *Handbook of Experimental Psychology* (Stevens, 1958).

Miller and Coleman constructed and administered three types of cloze tests for the 36, 150-word passages: Cloze Test I (CT I), Cloze Test II (CT II), and Cloze Test III (CT III). They constructed five versions of CT I. For each of the tests they deleted every fifth word. For the first version of CT I, they started with the first word, for the second version they started with the second word, and so forth. Each version of CT I included 30 deletions. The authors created 150 versions of CT II. Each version had only a single word deleted. For CT III, Miller and Coleman deleted every word in the passage and required participants to guess each word. After participant attempted to guess the word, the correct word was revealed to them and they moved on to the next word. With this approach, the participants were exposed only to words preceding the blank for which they were guessing.

Twenty participants completed the five versions of CT I (four participants per version for each passage). There were a total of 600 responses for CT I. Miller and Coleman (1967) had 450 participants complete CT II (three participants per version for

each passage). This resulted in 16,200 responses or 450 guesses for each passage. The participants who took CT III worked over several days and completed the test for all 36 passages. This resulted in 1,350 responses for each passage, for a total of 48,600 participant responses.

Miller and Coleman (1967) transformed the scores from each cloze test into percentage correct values. The mean percentage scores and standard deviations for each test averaged across the 36 passages were as follows: CT I: M = 54.6, SD = 14.5; CT II: M = 63.8, SD = 11.0; CT III: M = 33.7, SD = 7.6. They found that the three types of cloze tests resulted in similar rankings of the passages. The correlations between the results of the methods were as follows: CT I and CT II: r = .95; CT I and CT III: r = .87; CT II and CT III: r = .87.

Miller and Coleman (1967) contended that the high degree of agreement among the three cloze test methods was evidence of stability. Miller and Coleman's 36-passage readability scale, and the cloze technique in general, were later validated by Coleman and Miller (1968) and Aquino (1969). Subsequently, the passages became widely used for readability formula development (Klare, 1984).

Bormuth (1967; 1968; 1969) did extensive research concerning the viability of cloze techniques for readability formula calibration. He offered a frame of reference for the interpretation of cloze scores by establishing cloze scores comparable to multiple-choice scores (1967, 1968, 1969). Bormuth used the multiple-choice standards put forth by Thorndike (1916): 75% correct on multiple-choice tests indicates that the tested passage is suitable for supervised (classroom) instruction; 90% correct score indicates that the passage is suitable for independent reading. These percentages had long been the

conventional guidelines used by educators and textbook authors but they were not based on scientific study (Bormuth, 1968; Dubay, 2004; Klare, 1966; Taylor, 1953). In fact, these criteria can be traced back to Thorndike (1917) who derived them from teacher opinions, who, in turn, adopted them based on oral tradition.

Bormuth investigated how cloze and multiple-choice scores corresponded. His aim was to establish a frame of reference for interpreting cloze scores according to Thorndike's (1916) multiple-choice test score guidelines. He conducted two studies to develop criterion scores for cloze tests that correspond with the criteria traditionally employed with multiple-choice comprehension tests (i.e., 75 and 90%). These studies were described in his 1969 work but were published individually in 1967 and 1968.

In the first study aimed at establishing a comparable criterion score, Bormuth (1967) administered 50-item cloze and 31-item multiple-choice tests over the same nine passages to 100 4^{th}- and 5^{th}-grade students. Through inspection of scatter plots and computing correlations between the scores from the different tests, he determined that the scores (cloze and multiple-choice) were linearly related (r = .946). Bormuth pooled the multiple-choice and cloze scores of the 4^{th}- and 5^{th}-grade students to create one set of multiple-choice scores and one set of cloze scores. Through regression analysis of the two sets of scores, Bormuth established that 38% correct cloze score corresponded to 75% correct multiple-choice score. When the multiple-choice scores were corrected for guessing, a 43% cloze score corresponded to a 67% multiple-choice score. A cloze score of 50% corresponded to 90% for multiple-choice (87% when corrected for guessing).

In his 1968 study, Bormuth's objective was to establish cloze criterion scores comparable to 75 and 90% completion test scores obtained in an oral reading test. He

used the four forms of the *Gray Oral Reading Paragraphs*, each of which consisted of 13 paragraphs with unique difficulty levels. Bormuth's participants were 120 4th-, 5th-, and 6th-grade students (40 per grade level). Participants completed cloze readability tests over two of the paragraphs at each level of difficulty and then completed oral comprehension tests of the other two paragraphs immediately after orally reading those paragraphs.

To establish comparable cloze criterion scores, Bormuth (1968) identified the most difficult levels upon which a participant was able to earn comprehension scores of 75% and 90%. The participant's two cloze scores at the corresponding difficulty levels were averaged. The results were similar to those of Bormuth's (1967) multiple-choice study: a 43.69% cloze score corresponded to a 75% completion test score and a 57.16% cloze score corresponded to a 90% completion test score (corrected for guessing).

Bormuth (1969) conducted a pilot project to demonstrate that it was possible to establish a rationally based criterion for minimum cloze performance that would correspond to a passage of suitable difficulty level. This was the first study of its kind in that it was the first attempt to establish empirically based criterion scores of any sort and deserves a thorough explanation. Therefore, it will be discussed in greater detail than Bormuth's other cloze studies.

Bormuth (1969) used 260 participants who were formed into matched reading ability pairs based on scores from a 52-item cloze readability test. The participants were of varying ability levels: 25 pairs from grade 3; 23 pairs from grade 5; 15 pairs from grade 7; 28 pairs from grade 11; 24 pairs from junior college; and 15 pairs from graduate school. Two passages, A and B, were extracted from the same source as the 52-item cloze

test and then multiple-choice comprehension and cloze readability tests were constructed from each.

To determine the difficulty of a passage for each pair of participants, one member completed a cloze readability test over that passage. Then, to establish the extent of information gain from reading the passage, the second member of each pair was given a multiple-choice test over the passage without reading it. At a one-week delay, the second member read the passage and immediately completed the same multiple-choice test.

Bormuth (1969) established the amount of information gain by subtracting the second member's first score from his/her second score on the multiple-choice test, both of which were corrected for guessing. The researcher then plotted the information gain scores for each pair against their cloze difficulty scores and regressed them using stepwise polynomial regression analysis to ascertain the relationship between cloze difficulty and information gain for the passage. This was done separately for each passage.

For both passages, the first three powers of information scores accounted for significant amounts of variance: passage A multiple correlation = .69 and passage B multiple correlation = .62. The polynomial curves for each passage were compared and were not significantly different from one another. Therefore, Bormuth (1969) combined the data sets for each passage into a single data set to which an eighth degree polynomial regression fit. The use of a higher degree polynomial allowed Bormuth to see the holistic nature of the relationship as well as the error fluctuations in the data. This revealed that pairs who could correctly answer less than 25% of cloze items gained little information from the text. Pairs that who were able to correctly answer more than 25% of the cloze

items showed a sharp increase in information gain from the passage. The gain continued to increase until cloze scores reached 35 to 40%.

Bormuth (1969) did not observe a ceiling effect. Only twelve of the 260 participants scored better than 90% on the second multiple-choice test. He attributes the leveling off of information gain to prior knowledge. Specifically, the first and second multiple-choice scores were significantly correlated (r = .42). When passages were particularly easy for participants, they earned high scores on the second administration but also performed fairly well on the first administration because of prior knowledge of the topic. This resulted in the appearance that they had gained less information from reading the passage than had students for whom the passage proved more difficult.

Bormuth (1969) interpreted his findings to indicate that it was possible to create a rationally based criterion for judging appropriate difficulties of reading materials for students at particular ability levels. He specifically emphasized that two passages were employed and showed very similar curves. Bormuth construed this to imply that a fixed relationship existed between cloze readability and information gain. Based on his preliminary data, he estimated a cloze criterion score of 35%. He qualified this estimation with attention to a limitation: he did not account for the influence of passage difficulty on student affect. Bormuth explained, “It is desirable, of course, to provide students with materials from which they can gain information, but it is even more desirable to provide them with materials which they will study without any more duress than is ordinarily involved in instruction” (p. 50). Specifically, his concern was that when students are required to study materials that are too difficult for them, they may become frustrated or

inattentive. He, therefore, clarified that the 35% criterion indicated the most difficult materials from which a student was likely to benefit.

In addition, Bormuth (1969) held that the 35% criteria should be considered with some apprehension because it is possible that the criterion may vary according to student reading ability, passage difficulty, individual student differences, or any sort of interaction of these variables. He contended that adopting a single criterion might be an over simplification of a complex matter. Bormuth made this supposition based on the work of Coleman and Miller (1968) and Kammann (1966). Coleman and Miller varied their passage difficulties and found some evidence that information gain may decrease at the extreme poles of passage difficulty. Kammann found that passage difficulty and student temperament affected student ratings of their interest in a passage.

Bormuth (1969) also admitted some methodological or material-related limitations for the 35% criterion. He used two passages and held that the number of passages and the methods used to select the passages were not sufficient to generalize the results to all passages. In addition, the methods used to create the multiple-choice tests were not sufficient to account for potential systematic bias. The results, however, might have been different if different writers had created the test. Finally, Bormuth acknowledged that offering the same multiple-choice test twice might have biased the results of the second administration. Nevertheless, Bormuth (1969) explained that good reasons remained for using it. It was the only rationally based criterion available at the time. The same limitations of the 35% criterion exist for the more traditionally accepted 45 and 55% criteria. Therefore, for the final section in his investigation (the calculation of several readability formulas), Bormuth employed 35%, 45%, and 55% as his criteria.

A modification of the cloze procedure, the limited-cloze procedure, was developed and validated by Cunningham and Cunningham (1978). Their primary rationale for this modification was that classroom teachers either refused to use the technique at all or failed to properly follow cloze procedure scoring guidelines. Specifically, classroom teachers tended to be too lenient in their scoring by counting synonyms of deleted words as correct responses. In a limited-cloze procedure the deleted words are placed above the passage in random order. The students are told that the words should be used to fill in the deleted words in the passage. This alleviates any concern about the use of synonyms because the correct words are provided. Cunningham and Cunningham established the validity and reliability of the limited-cloze procedure with 163, 7th-grade students (study I) and 203 5th-grade students (study II).

The cloze technique has been used as a criterion for the development of several readability formulas. In addition, it has been used to recalibrate several existing formulas that were previously calibrated based on the McCall-Crabbs *Standard Test Lessons in Reading* (McCall & Crabbs, 1925, 1950, 1961). Readability formulas devised with the cloze technique include: Coleman formulas (1965); Bormuth Mean Cloze formula (1969); a modification of the Bormuth formula: Degrees of Reading Power (College Entrance Examination Board, 1981); Coleman-Liau formula (1975); FORCAST formula (Caylor, Sticht, Fox, & Ford, 1973); Hull Formula for Technical Writing (1979); William, Siegel, Burkett, and Groff formula (1977); and Hull formula (1979; Dubay, 2004). Popular formulas that were recalibrated using the cloze technique include the Flesch-Kincaid formula (Kincaid, Fishburne, Rogers, & Chissom, 1975) and the New Dale-Chall Readability Formula (Chall & Dale, 1995).

Some readability formula developers who turned to the cloze technique for calibration purposes have used previously calibrated passages (e.g., Chall & Dale, 1995; Coleman & Liau, 1975). Others have used their own passages and participants with the cloze technique (e.g., Caylor & Stitch, 1973; McLaughlin, 1969). In either case, the cloze technique is simpler, less costly, and introduces less measurement error than creating multiple-choice tests for passages to be administered to participants. In addition, the cloze technique likely offers a more accurate means of calibration than using the McCall-Crabbs passages (McCall & Crabbs, 1925, 1950, 1961), which was not devised for readability formula calibration.

Measuring Readability

In this section, the readability research and formula development that occurred throughout the 20th century is discussed. The discussion begins with attention to the precursors to formal readability measurement. Readability research projects are then discussed in chronological order because, in most cases, they largely build upon one another. Deviations from chronological order occur in cases where readability formulas were revised over the years. In these instances, the original formula is presented followed by the revised versions. In the review of these research and readability formula development projects, discussions are offered regarding the information provided by the authors related to calibration methodologies, materials, and validation studies.

In the Beginning

Attempts to measure readability began as early as 900 C.E. when Talmudists counted words and individual ideas of the Torah scrolls. This was done to clarify unusual

meanings and to devise appropriate divisions of the Torah into approximately equal comprehension units for weekly readings (Lorge, 1944b).

According to Lorge (1944b), word counts were further employed by scholars throughout the centuries in an effort to identify lists of words that people of specific populations should know. For example, Kaeding (1898), a German scholar, was one of the first to use word counts to establish basic vocabulary. His count was based on nearly eleven million words and was done to determine word frequency for a shorthand system. In 1902, Reverent J. Knowles created a 350-word basic vocabulary list for the blind that was comprised primarily of passages from the Bible. Eldridge created a much larger list in 1911. He created a six thousand common English word list from issues of the Buffalo Newspaper (Lorge, 1944b).

Formal readability research began in the 1920s and stemmed from two main sources: studies of vocabulary control and studies of readability measurement (Chall, 1988). Studies of vocabulary control concentrated on vocabularies that would be most suitable for learning to read and were particularly focused on the frequency and difficulty of "new words" in textbooks. Readability measurement began with attention to the difficulty of content area textbooks. In the early years of readability measurement studies, scholars created procedures and instruments to discriminate between easier and more difficult texts and to rank them in terms of difficulty.

An important contribution by Thorndike (1921), *A Teacher's world book*, paved the way for objective measures of readability (Lorge, 1944b). Over ten years, Thorndike compiled a list of 10,000 words that was the first comprehensive listing of English words by frequency of use (DuBay, 2004). This provided an objective measure of word

difficulty (Chall, 1984) and laid the ground for most future readability research (DuBay, 2004). A decade later, Thorndike (1932) extended his work with the publication of *A Teacher's world book of 20,000 words*. Then, in 1944, Thorndike and Lorge (1944) added another ten thousand words in their publication, *A Teacher's world book of 30,000 words.* A variety of vocabulary word lists were subsequently created by several readability scholars (e.g., Dale, 1943; Leary, 1938; Spache, 1953).

Contemporary Readability Measures

Lively and Pressley (1923) used Thorndike's (1921) word list to create the first readability formula (Chall, 1988; DuBay, 2004; Klare, 1984; Hewitt & Homan, 1991). Their work was stimulated by junior high science teachers' concern that their textbooks were overly laden with technical jargon. Teachers complained that they spent the majority of their time explaining vocabulary, rather than teaching content (DuBay, 2004). Lively and Pressley examined three methods of measuring readability. For each 1,000 word passage they counted: 1) the number of different words; and 2) the number of words not included in Thorndike's (1921) 10,000-word list. After obtaining the word count totals, they identified the median for all passages sampled. They determined that the median index was the best indicator of readability level, where higher median index values indicated easier reading materials and lower values indicated more difficult reading materials.

Gray and Leary.

In 1935, Gray and Leary published a monumental study of readability that examined more style elements and relationships between them than any readability research that has been published since (DuBay, 2004). Gray and Leary's focus was to determine what

makes books readable for adults with low levels of reading ability. The researchers began their investigation by surveying 100 experts and 100 library patrons about what makes a book readable. They divided the 289 answers into four categories: content, style, format, and organization. The researchers then cut the exhaustive list to 44 style variables they believed they could reliably count.

Gray and Leary (1935) administered several reading comprehension tests to thousands of adults and found that of the 44 factors, 20 showed a significant relationship to the ability to answer comprehension questions. Through multiple regression, they identified five style factors that accounted for the greatest variance in reading difficulty: 1) the number of different difficult words, 2) the percentage of different words, 3) the average sentence length in terms of words, 4) the number of prepositional phrases, and 5) the number of personal pronouns. These five variables had a correlation of .65 with reading difficulty. The first four variables were positively related to reading difficulty and the last variable (number of personal pronouns) was negatively related to reading difficulty. Gray and Leary's use of style variables and multiple regression became the most common method of investigation for readability in further research.

Lorge.

Lorge (1939) created a readability formula that he later revised (1948) to correct an error made in the first version (Klare, 1974-1975). In his 1939 study, Lorge examined predictors employed by readability scholars. He examined the five factors used by Gray and Leary (1935) as well as weighted vocabulary scores based on Thorndike's (1932) 20,000-word list. Lorge (1939) also explored four factors used by Morris and Holversen (1938): 1) number of elemental words, 2) percentage of simple localisms, 3) percentage

of concrete word-labels, and 4) percentage of abstract words. Through multiple regression, Lorge identified three predictors that correlated .77 with his criterion: average sentence length in words (X_1), number of prepositional phrases per 100 words (X_2), and number of uncommon words (according to Dale's list of 769 words; X_3). This three-factor prediction equation was combined with a constant to offer a grade-level estimate.

In 1948, Lorge revised the formula by slightly altering the constant because he found that he had made a mistake in the constant used in 1939 (Lorge, 1948a). Lorge's (1939) formula was as follows: grade placement = .07 X_1+ .1301 X_2+ .1073 X_3+ 1.6126. His revised readability formula took the following form: grade placement = .06 X_1+ .10 X_2+ .10 X_3+ 1.99.

Flesch.

Flesch (1943, 1948) was the next scholar to make a significant contribution to readability research with his attention to adult-level reading material. He published his first readability formula in 1943 and included three language elements: 1) average sentence length in words, 2) number of affixes, and 3) number of references to people. This formula was widely used and applied to newspaper publications, bulletins and leaflets for farmers, adult education materials, and children's books. In 1948, Flesch reevaluated the formula based on an important shortcoming: Flesch's (1943) formula was partly based on Lorge's (1939) erroneous calculations and it sometimes yielded inconsistent results. For example, the formula showed that *Reader's Digest* was more readable than *The New Yorker* magazine (Flesch, 1948). Flesch took issue with this because he contended that most educated readers found the *Reader's Digest* boring and *The New Yorker* magazine much more readable. In addition, practical applications of the

formula led to misinterpretations because the element that was easier to estimate (sentence length) was overemphasized and the element that was more difficult to estimate (number of affixes) was underestimated. Furthermore, practitioners had difficulty using the scoring system. Flesch (1948), therefore, modified his formula.

Flesch's (1948) reanalysis involved four factors: 1) average sentence length in words, 2) average word length in syllables, 3) average percentage of personal words, and 4) average percentage of personal sentences (e.g., quotes, questions, commands, requests, and other sentences directed to the reader). Analyses of these variables through multiple correlations and multiple regression led to the creation of two readability formulas: the reading ease formula and the human interest formula. The reading ease formula included the average sentence length (*sl*) and average word length (*wl*) elements and a constant: Reading Ease = 206.835 – (846) (*wl*) – (1.015) (*sl*). The human interest formula consisted of the average percentages of personal words (*pw*) and personal sentences (*ps*): Human Interest = (3.635) (*pw*) + .314 (*ps*). The results from formulas are interpreted on a 100-point scale. Reading ease formula scores are interpreted as follows: 0 – 30 is *very difficult*; 30 – 50 is *difficult*; 50 – 60 is *fairly difficult*; 60 – 70 is *standard*; 70 – 80 is *fairly easy*; 80 – 90 is *easy*; and 90 – 100 is *very easy*. Human interest scores are interpreted as follows: 0 – 10 is *dull*; 10 – 20 is *mildly interesting*; 20 – 40 is *interesting*; 40 – 60 is *highly interesting*; and 60 – 100 is *dramatic*.

Flesch (1948) found that the reading ease formula showed a .70 correlation with the criterion (McCall-Crabbs Standard Test Lesson in Reading, 1926), which was only .04 lower than the correlation of his earlier (1943) formula. Conversely, the human interest formula yielded a .43 correlation with the criterion. Flesch, therefore, admitted that the

human interest formula contributed little to readability research. On the other hand, he reminded the reader that the human interest formula included only two human interest variables and that the correlation coefficient exclusively reflected the extent to which human interest would make a passage easier for a reader to understand.

Flesch (1948) tested the formulas with text passage samples similar to those that had been problematic for his 1943 formula. He applied the three formulas (i.e., the old formula, 1943; the reading ease formula, 1948; the human interest formula, 1948) to passages taken from *The* New Yorker and *Reader's Digest*. As expected, the old formula and the reading ease formula rated *Reader's Digest* as significantly more readable than *The New Yorker*. Conversely, the human interest formula rated *The New Yorker* as significantly more readable.

In a sample application of the formulas, Flesch (1948) applied the three formulas to two pieces of text that discussed the same topic. *Life* magazine and *The New Yorker* had both published articles about the nerve-block method of anesthesia. Flesch explained that the *Life* magazine passage was very straightforward, complex, and lacked human interest. Conversely, "*The New Yorker* passage is [was] part of a personality profile, vivid, dramatic, using all the tricks of the trade to get the reader interested and keep him in suspense" (p. 231). As expected, all three formulas rated *The New Yorker* passage as significantly more readable than the *Life* magazine passage.

Farr, Jenkins, and Patterson.

Farr, Jenkins, and Patterson (1951) created a simplified version of Flesch's reading ease formula (1948) to make it easier to use. They contended that syllable counts are difficult and may introduce error because analysts may make mistakes. According to the

authors, counts of one-syllable words are sufficient to replace counts of syllables per one hundred words. This would enable practitioners without knowledge of syllabification to more accurately and quickly use the formula.

To test their hypothesis, Farr et al. (1951) extracted 360 one-hundred-word samples from 22 General Motors employee handbooks. They applied Flesch's reading ease formula (1948) to the passages and counted the number of one-syllable words per passage. The authors then calculated correlations between: 1) the number of one-syllable words and the number of syllables per passage (r = -.91), 2) the number of syllables per 100 words and Flesch's formula (r = -.87), and 3) the number of one syllable words and Flesch's formula (r = .76). The Flesch reading ease index is: reading ease = 206.835 – (846) (*wl*) – (1.015) (*sl*). Farr et al's (1951) new reading ease index is: 1.5999 (number of one syllable words per 100 words) – 1.015 (*sl*) – 31.517.

Farr et al. (1951) applied the old and new reading ease indices to the 360 sampled passages and found that mean reading scores were essentially the same, but the new formula had less variability than the old formula: old formula score mean = 48.3, *SD*. = 15.7; new formula score mean = 47, *SD*. = 14.2. Old and new formula scores for the 360 one-hundred-word samples were highly correlated (r = .93). Because Flesch (1948) fashioned his reading ease formula for use with whole books, Farr et al. took several passages from each manual, applied both formulas to the passages, and established average readability scores for each manual. The correlation between the old and new average reading ease scores for the 22 passages was .95. Farr et al. (1951) contended that the correlation would have been higher but there was restriction of range in difficulty for the passages. That is, the average difficulty levels in the manuals were

very similar. On the 100-point scale, the mean difficulties ranged from 36 (difficult) to 57 (fairly difficult). According to Farr et al. (1951), had the reading ease averages ranged from very easy to very difficult, the correlation would have likely reached .99. Therefore, the authors held that their revised formula could be used more quickly, would require less knowledge of syllabification, and could be safely substituted for the old reading ease formula.

Dale and Chall.

During the same year that Flesch revised his original formula, Dale and Chall (1948) published the Dale-Chall readability formula. Their formula became one of the most widely used readability formulas in education (Klare, 1988). The popularity of this method was likely due to the validation studies of the Dale-Chall formula rendering more consistent results and higher reliabilities than any of the other formulas devised during this period (DuBay, 2004). The Dale-Chall formula was based on three hypotheses: 1) a larger word list (as compared to the Dale 796-word list) would offer an equal or better prediction of difficulty than counts of affixes; 2) counting personal references does not contribute much to predicting readability; and 3) a shorter, more efficient formula could be devised employing word and sentence structure factors.

Like Lorge (1939) and Flesch (1943), Dale and Chall (1948) used sample passages from the McCall-Crabbs *Standard Test Lessons in Reading* (1926). Their criterion was the grade-level from a group of students who correctly answered half of the multiple-choice questions. An important distinction between the Dale-Chall (1943) formula and the Lorge and Flesch formulas is Dale and Chall's creation and use of their own list of three-thousand words. To create this list, the authors tested fourth-grade students' reading

knowledge of approximately ten thousand words. The list was comprised of the most common words on Thorndike's (1931) list of ten-thousand words and Buckingham and Dolch's (1936) combined word list. Unlike the Thorndike list, which was based on frequency of appearance in printed material, the Dale list was a measure of familiarity.

Dale and Chall (1948) counted the relative number of words in the 367 passages (books two and five) of the McCall-Crabbs *Standard Test Lessons in Reading* (1926) that were not on the Dale list of 3,000 words. They found that number of words not on the list correlated .6833 with the criterion (i.e., grade-level of a group of students who correctly answered half of the multiple-choice questions on the McCall-Crabbs *Standard Test Lessons in Reading*). Sentence length offered the next highest correlation, r = .4681, with the criterion. Dale and Chall tested several combinations of the following factors: average sentence length, words outside the 3,000-word list, affix counts, personal reference counts, and words outside the Dale 769-word list. They found that the combination of words not on the Dale 3,000-word list (vocabulary load factor, X_1), average sentence length (sentence structure factor, X_2), and three constants (.1579, .0496, and 3.6365) provided the best prediction of readability: Readability = (.1579) (X_1) + (.0496) (X_2) + 3.6365. This combination of variables yielded a multiple correlation of .70 with the criterion.

Dale and Chall (1948) tested their formula with passages other than the McCall-Crabbs (1926). They compared the formula predictions to judgments made by experienced teachers and readability experts as well as readers' comprehension scores. The formula predictions correlated .92 with judgments of readability experts and .90 with comprehension scores of children and adults for fifty-five passages of health-education

materials. For seventy-eight passages from current-events magazines, government pamphlets, and newspapers, the formula prediction correlated .90 with judgments of experienced social science teachers. Dale and Chall (p. 18) provided a table of estimated corrected grade levels for formula scores (see Table 1).

After decades of monitoring use of the formula in research and practice, Chall and Dale revised their readability formula and published the new Dale-Chall readability formula in 1995. Although they contended that their original formula showed high levels of reliability and validity (Chall, 1955), they chose to make two revisions. First, they thought it was important to revise the formula based on a new set of criterion passages, an updated word list, and improved methods for measuring the word familiarity and sentence length factors. Second, they thought it was necessary to simplify essential computations and instructions.

Table 1

Dale-Chall (1948) corresponding grade levels for formula scores

Formula Score	Corrected Grade Levels
4.9 and below	4 and below
5.0 to 5.9	5-7
6.0 to 6.9	7-8
7.0 to 7.9	9-10
8.0 to 8.9	11-12
9.0 to 9.9	13-15 (college)
10.0 and above	16 + (college graduate)

To standardize their new formula, Chall and Dale (1995) used the cloze procedure on thirty-two passages from Bormuth (1971), thirty-six passages from Miller and Coleman (1967), eighty passages from MacGinitie and Tretiak (1971), and twelve passages from Caylor, et al., (1973). These passages ranged from third grade to college graduate reading levels. Chall and Dale retained their original syntactic variable, average sentence length. An updated word list was employed, *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981). Samples from reading material to be analyzed were shortened to exactly 100 words. Rules for counting headings were introduced. Through multiple correlations and multiple regression analyses, Chall and Dale created their new readability formula: Dale-Chall cloze = 64 – (.95) (number of unfamiliar words) – (.69) (average sentence length).

The use of the new Dale-Chall readability formula (1995) does not require a practitioner to calculate the Dale-Chall cloze formula. Chall and Dale developed cloze and reading level tables that have the number of familiar words along the Y axis and number of sentences in the sample along the X axis. The practitioner follows the following steps for each 100-word sample: 1) count the number of complete sentences; 2) count the number of unfamiliar words; 3) obtain a cloze score, via the cloze table, with the counts of sentences and unfamiliar words; 4) obtain a reading level score, via the reading level table, with the counts of sentences and unfamiliar words. These steps are repeated for each passage and the cloze and reading levels are then averaged.

The Dale-Chall cloze formula yields a cloze score that can be converted to reading level. Cloze scores indicate the percentage of deleted words in a passage that can be correctly identified by readers. Passages with higher cloze scores are estimated to be

more readable: passages with cloze scores above 57 are the easiest and those with scores below 15 are the most difficult. The authors contended that cloze scores may be preferable to reading levels in research settings because cloze scores offer a wider range and more precise measurement. Cloze scores might also be more useful in differentiating the difficulty levels of different texts and for use with adult-level reading material.

Reading levels range from 1 (approximately 1st-grade) to 16 (college graduate level). Reading levels 1-4 correspond with their respective grades and levels, whereas 5 through 16 depict ranges of reading levels (i.e., 5-6, 7-8, 9-10, 11-12, 13-15, and 16+). Chall and Dale (1995) provided a table for the conversion of cloze scores to reading levels and reading levels to a range of cloze scores. The correspondence of reading levels (RL) to cloze score ranges (CS) are included in Table 2. The authors provided an additional conversion table for obtaining exact cloze scores based on reading level. Chall and Dale explained that reading level values might be preferable to cloze scores when the intent is to match a reader's ability to text difficulty.

Although Chall and Dale (1995) designed their formula for use with several 100-word samples, they also offered a set of amended instructions for use with samples shorter than 100 words. The number of sentences and number of unfamiliar words should be converted to percentages. This requires dividing the number of sentences by the number of words in the sample and dividing the number of unfamiliar words by the number of words in the sample, respectively. The tables for cloze and reading level scores should then be used in the same manner as with the regular formula. Chall and Dale's attention to the matter of shorter selections of texts is helpful, but they did not test this amended version against any criteria.

Table 2

Reading level to cloze score range correspondence for Chall and Dale's formula (1995)

Reading Levels	Cloze Score Range
1	58+
2	57-54
3	53-50
4	49-45
5-6	44-40
7-8	39-34
9-10	33-28
11-12	27-22
13-15	21-16
16+	15 and below

Gunning.

Gunning (1952) was one of the first researchers to address readability concerns in the workplace. After years of working as a readability consultant for large newspapers and magazines, he published, *The Technique of Clear Writing* (1952) in which he presented a readability formula for adults, the FOG Index (DuBay, 2004). This formula was widely used by several government agencies for their writing manuals (e.g., Army, Navy, Air Force, and the Department of Agriculture).

Gunning (1952) considered the readability formulas developed prior to 1952 to be too complex and difficult for practical use. Therefore, he attempted to create a formula that was easy to use, would render reliable results, and would focus writers' attention on

factors that cause readers the most difficulty. Gunning (1952) identified two factors that he thought contributed most to reading levels: average sentence length and number of hard words (more than two syllables).

Like readability researchers before him, Gunning (1952) used passages from the McCall-Crabbs *Standard Test Lessons in Reading* (1926) to create his formula. His criterion for grade-level estimates was much more stringent than those of his predecessors: he identified the average sentence length and percentage of hard words in passages for which students from grade levels 6, 8, 10, and 12 correctly answered 90% of the comprehension questions. He used a regression equation to transform the variables into grade levels. Gunning's equation is simpler than those of earlier readability formula authors: Grade level = .4 (average sentence length + percentage of hard words). Each complete thought in a sentence is treated as a separate sentence. Because Gunning's (1952) criterion was so much higher than those of other readability researchers, his index tends to render readability estimates higher than those of other formulas (e.g., reading ease, 1948; Dale-Chall readability formula, 1948; DuBay, 2004). Validation studies for the FOG Index have never been published, but according to DuBay's calculations, the FOG Index correlates .93 with the normed passages used by Chall, Bissex, Conard, and Harris-Sharples (1996).

Spache.

According to Spache (1953), the abundance of readability formulas was developed to address the difficulty levels of adult reading materials (e.g., Flesch, 1948 and Dale-Chall, 1948). Therefore, Spache devised a formula intended for primary-grade (i.e., below grade 4) materials. Following Dale-Chall's (1948) lead, he employed average sentence length

and the Dale list of 769 words in his formula. He extracted 224 one-hundred-word sample passages from 152 books that were commonly used in grades one, two, and three. Spache assigned grade levels to these books according to their classroom use: 1.2, pre-primer; 1.5, primer; 1.8, 1st-grade; 2.1, 2nd-grade; and 3.3, 3rd-grade.

The multiple correlation for the combined variables of sentence length and percentage of hard words (not on the Dale 769 list) with predicted grade levels of books was .818. In particular, Spache found that sentence length (r = .751) was more closely related to difficulty in primary texts than was vocabulary load (r = .683). This is contrary to the findings of Lorge (1944), Flesch (1948), and Dale and Chall (1948), who established that vocabulary load was the most important factor in predicting readability. Spache reconciled this difference by explaining that primary materials are constructed differently than higher-level texts. Specifically, authors of primary-level books are more cautious about sentence length. Through multiple regression, Spache arrived at a formula to predict the readability of primary-level materials: grade level = .141 (average sentence length per 100 words) + .086 (percent of words outside the Dale "easy word" list of 769 words) + .839.

Powers, Sumner, and Kearl.

In 1958, Powers, Sumner, and Kearl recalculated the Flesch reading ease formula (Flesch, 1948), the Dale-Chall readability formula (Dale & Chall, 1948), the Farr-Jenkins-Patterson formula (Farr, Jenkins, & Patterson, 1951), and the FOG index (Gunning, 1952) with the revised McCall-Crabbs *Standard Test Lessons in Reading* (1950). Powers et al. thought that the formulas required recension because they were based on the outdated 1926 version of the McCall-Crabbs and the original formula

authors did not include a standard error figures. The authors measured the following variables in the 383 passages of the McCall-Crabbs (1950): 1) average grade score of pupils who correctly answered 50% of comprehension questions; 2) average number of words per sentence; 3) number of syllables per 100 words; 4) percentage of words in each passage not included in the Dale list of 3,000 words (1948); 5) percent of monosyllables (one-syllable words) per passage; and 6) percent of polysyllables (words with more than one syllable) per passage. Through regression analysis of the five measures and comparisons of scores from the four formulas applied to 113 samples of text from various sources, Powers et al. established revised versions of each formula. See Table 3 for original and recalculated formulas.

Table 3

Original and recalculated formulas (Powers et al.1958)

Formulas	Original	Recalculated
Flesch Reading Ease	= 206.835 – (846) (*wl*) – (1.015) (syllables per 100 words)	= -2.2029 + (.0778) (*sl*) + (.0455) (syllables per 100 words)
Dale-Chall Readability	= (.1579) (% non-Dale words) + (.0496) (*sl*) + 3.6365	= 3.2672 + (.0596) (*sl*) + (.1155) (% non-Dale words)
Farr-Jenkins-Patterson revised Reading Ease	= 1.5999 (number of monosyllables) – 1.015 (*sl*) – 31.517	= 8.4335 + (.0923) (*sl*) – (.0648) (% monosyllables)
Gunning FOG Index	= .4 (*sl* + percentage of hard words)	= 3.0680 + (.0877) (*sl*) + (.0984) (% polysyllables)

Note. "wl" = word length*; "sl"* = average sentence length.

Powers et al. (1958) calculated coefficients of multiple determination to establish the variance in difficulty accounted for by the style variables included in each formula. They found that the recalculated Flesch formula accounted for 40% of variance in difficulty of the McCall-Crabbs tests (1951); the recalculated Dale-Chall formula accounted for 51%, and the recalculated Farr-Jenkins-Patterson and FOG formulas accounted for 34%. The error terms for the recalculated formulas are: Flesch, .85 grade levels; Dale-Chall, .77 grade levels; Farr-Jenkins-Patterson and FOG, .90 grade levels. Conversions into grade-level figures and inclusion of standard error practices (i.e., range plus or minus two standard errors) resulted in the following error ranges for the recalculated formulas: Flesch, 1.71 grade levels; Dale-Chall, 1.55 grade levels; and Farr-Jenkins-Patterson and FOG, 1.80 grade levels. These results indicated that the recalculated Dale-Chall formula provided the most accurate results.

To appraise the practical utility of the recalculated Flesch and Dale-Chall formulas, Powers, et al. (1958) applied the original and recalculated formulas to 47 sample passages from a variety of sources. Both recalculated formulas generated lower difficulty scores than the original versions (Dale-Chall, .94 grades; Flesch, .85 grades). The authors then applied the four recalculated formulas to 113 sample passages from 15 magazines and compared. The average discrepancy between the recalculated Dale-Chall and recalculated Flesch was .54 grade levels, whereas a comparison between the original Dale-Chall and Flesch resulted in grade-level differences of .87. In addition, all four recalculated formulas showed better agreement with each other (deviations: Dale-Chall and Flesch, .54; Flesch and Gunning, .44; Dale-Chall and Gunning, .56; Flesch and F-J-P,

.50; Dale-Chall and F-J-P, .66; and Gunning and F-J-P, .54) than the original Dale-Chall and Flesch did with each other (.87).

Coleman.

Coleman (1965) was the first scholar to employ Taylor's (1953) cloze procedure, instead of the traditional McCall-Crabbs *Standard Test Lessons in Reading* (1926, 1950) or judges' rankings, to develop a readability formula (Dubay, 2004; Klare, 1974-1975). Coleman's research project was sponsored by the National Sciences Foundation and the report is not available to the public. Therefore, secondary sources are cited in this section. Coleman devised four formulas that included: percentage of correct cloze completions (C%); number of one-syllable words per 100 words (*w*); number of sentences per 100 words (*s*); number of pronouns per 100 words (*p*); and number of prepositions per 100 words (prep; Dubay, 2004; Klare, 1974-1975). Coleman's four readability formulas take the following form:

$$C\% = 1.29w - 38.45$$

$$C\% = 1.16w + 1.48s - 37.95$$

$$C\% = 1.07w + 1.18 + .76p - 34.02$$

$$C\% = 1.04w + 1.06s + .56p - .36prep - 26.01$$

Coleman (1965) found high multiple correlations among his formulas and cloze completion scores: .86, .89, .90, and .91, respectively (Dubay, 2004; Klare, 1974-1975). In a cross-validation study, Szalay (1965) confirmed Coleman's findings with only marginally weaker multiple correlations: .83, .88, .87, and .89, respectively. Use of cloze scores clearly showed higher validation coefficients than the use of the McCall-Crabbs (1926, 1950) multiple-choice scores (DuBay, 2004; Klare, 1974-1975). Coleman's study

marked a turning point in readability research. Readability scholars from that point on began to primarily employ cloze procedures in their research.

Bormuth.

During the 1960s, Bormuth published a series of studies that has been referred to as "the most extensive readability research to date" (Felker, 1980, p. 79). Bormuth's 1966 research was not conducted to develop a new readability formula; instead, his work focused on the viability of cloze techniques for readability formula calibration and the impact of additional predictor variables. Bormuth's (1966) research revealed several findings concerning the utility and influence of additional variables on reading comprehension.

Bormuth (1966) used 20 sample passages of 275 to 300 words from literature, history, geography, biology, and physical science instructional texts. He selected these passages to render a generally equal distribution of readability levels from 4.0 to 8.0 grade levels, according to Dale-Chall's readability formula (1948). Bormuth created five cloze tests with these passages by deleting every fifth word and starting at five different points. He administered the cloze tests to students from grades four to eight.

Bormuth (1966, p. 124) contended that the cloze technique "solved the problem of validity" because its use allowed a more powerful and flexible means of measuring difficulty. His results specifically revealed findings related to the following: 1) linearity of regressions, 2) variable strength as a function of reading ability, 3) predictive difficulties of small language units, 4) validities of readability formulas, and 5) new linguistic variables. Each set of findings are briefly discussed below.

Bormuth's (1966) conducted *F* tests of linearity and discovered that at the word level all existing correlations were curvilinear. He therefore contended that scholars should use quadratic equations to predict difficulty at that level. At the independent clause level, the Dale-Chall 3,000 word list significantly departed from linearity and several other factors approached significance (e.g., word frequency and word depth). Inspection of scatter plots led him to contend that curvilinearity was most notable at the extreme ends of the difficulty distribution. He asserted that the low power of *F* tests of linearity might have been responsible for the absence of significant results and proposed that more powerful methods of analysis should be employed in future investigations. Bormuth's results were similar at the passage level. Although scatter plots suggested curvilinearity at the extreme ends of the difficulty distribution, none of the *F* tests of linearity reached significance. Once again, Bormuth implicated the insufficient power of the statistical tests for the failure to find significance.

Using analysis of variance to investigate variable strength as a function of reading ability, Bormuth (1966) found that a linguistic variable offered equivalent predictions of difficulty for readers of different ability levels. He consequently concluded that a single readability formula could be reliably used for participants of varying reading abilities and that the formulas could be used at higher reading levels than previously thought.

As part of his investigation of whether reliable predictions of readability could be made from small language units, Bormuth (1966) measured multiple correlations between small language units (i.e., individual words, independent clauses, and sentences) and comprehension difficulty. He found multiple correlations of .51 for individual words; .67 for independent clauses; and .68 for sentences. Once again, his inspection of scatter

plots revealed curvilinearity. According to standards used in the past, readability formulas with validity levels from .5 to .7 are useful. Bormuth thus asserted that the small language units he examined were of use but could be markedly improved by devising a method of addressing the curvilinear relationships at word, sentence, and prose levels.

Bormuth (1966) also addressed whether the validity of readability formulas based solely on linguistic variables could be improved. He calculated two multiple regressions at the passage level of analysis that resulted in multiple correlations of .93 and .81. The new linguistic variables originated by Bormuth entered the equation at higher levels, nearly without exception, than linguistic variables employed in previous research. From these results, Bormuth concluded that readability formulas could be markedly improved by including new linguistic predictor variables.

Finally, Bormuth (1966) addressed the question of whether the use of new types of linguistic variables could offer improvements in the accuracy and reliability of readability predictions. Bormuth's investigation of 47 predictor variables resulted in an abundance of findings. Here, only the three findings he deemed most important are discussed (see Bormuth, 1966 for further details of the findings). First, although sentence length and complexity were highly correlated, each showed a significant relationship with difficulty. Second, difficulty was significantly correlated with part of speech variables. Third, a number of previously employed predictor variables were significantly improved with minor refinements.

In a more comprehensive investigation funded by the United States Department of Health, Education, and Welfare, Bormuth (1969) conducted a series of studies to gain information necessary to improve student comprehension of their instructional materials.

This series of studies concerned the analysis of linguistic variables, establishing cloze criterion scores comparable to traditional comprehension criterion scores, and to calculate readability prediction formulas. Bormuth used 2,600 4th- to 12th-grade students, their California 1963 Reading Achievement test scores, 330 100-word passages, and five cloze tests for each passage. He identified and determined the reliability of 164 variables related to vocabulary, syntactic structure, syntactic complexity, parts of speech, and anaphora and developed 24 readability formulas. Because the nature of syntactic structure, syntactic complexity, and anaphora variables is not readily apparent, explanations of these variables follow.

According to Bormuth (1969), *syntactic structure* potentially influences comprehension. He explained that according to transformational theory, deeper structures underlie sentences and represent semantic interpretations of them. The underlying forms of the structures in a sentence must be identified before the sentence can be understood. He, therefore, included in this portion of the research a syntactic structure analysis, which consisted of "identifying the basic structures occurring in English sentences and then counting the number of transformations required to derive the surface structure from the assumed underlying structures…." (Bormuth, p. 11).

Bormuth (1969) separately analyzed *syntactic complexity* variables because these variables correlate with passage difficulty. In addition, the complexity can be manipulated independent of types or numbers of structures in a sentence. His measures of syntactic complexity concerned the structural density of sentences (i.e., the proportion of structures per words, clauses, minimal punctuation units, and/or sentences); transformational complexity (i.e., density of the operations in a segment of prose

necessary to identify the underlying structure); structural complexity (i.e., the ratio of structures to words in a sentence); Yngve depth (i.e., a model used to predict reader/listener behavior and comprehension, see Yngve, 1960 for a complete description); and syntactic length variables (i.e., a measure of word length using letters, syllables, words, clauses, and minimal punctuation units).

Anaphors are similar to pronouns in that they include a pro element and an antecedent. They typically enable authors to state a complex idea and offer a shorter version of that idea to which the author can subsequently refer as a sort of shorthand. For example, in the sentence, "The boy took the book and read it", "it" refers to the book and is an anaphora. Bormuth (1969) analyzed *frequency*, *density*, and *distance* of anaphora variables. Frequency variables represent the proportion of occurrences of a particular type of anaphora to the total number of anaphora in a passage. Density variables are the proportion of anaphoras to the number of words in a passage. Anaphora distance concerns how many words occur between an anaphoric expression and its antecedent.

In his analysis of linguistic variables, Bormuth (1969) assessed the correlations between each of the predictor variables and passage difficulty and factor analyzed the linguistic variables for passage difficulty. The purpose of Bormuth's first step in the analysis was to identify a great number of linguistic variables that might be related to reading comprehension and determine which of them correlated significantly with passage difficulty. Ninety-five of the 164 linguistic variables related to vocabulary, structure portion, syntactic complexity, parts of speech, and anaphoras were significantly correlated with passage difficulty. Specifically, the numbers of significant correlations with passage difficulty were as follows: 8 of 8 vocabulary variables; 20 of 50 structure

portion variables; 34 of 38 syntactic complexity variables; 25 of 62 part of speech variables; and 8 of 11 anaphora variables. Bormuth explained that an even greater number of variables may have been significantly related to passage difficulty but the relationships were impossible to identify because of insufficient occurrences in the passages.

Bormuth (1969) clarified that the significant correlations between the linguistic variables and passage difficulty should not be construed to indicate that all of the linguistic variables cause passage difficulty. Specifically, the part of speech and syntactic length variables, although related to difficulty, could not be directly manipulated and therefore could not be implicated as actual causes of difficulty. Syntactic structure and anaphora variables, however, were directly manipulable and, therefore, could be inculpated as causes of passage difficulty.

Bormuth (1969) also factor analyzed the 95 linguistic variables that were significantly correlated with passage difficulty as well as two additional variables: ratio of lexical to structure words (WL/WS) and proportion of lexical words (WL/W). He defined lexical words as nouns, verbs, adjectives, and adverbs and structure words as pronouns, modal and auxiliary verbs, articles, and prepositions. Using principal component analysis with varimax rotation, Bormuth extracted 20 factors that accounted for 73.7% of variance. Two patterns of factor loadings emerged. First, almost all of the syntactic complexity variables loaded heavily on three factors with loadings ranging from .45 to .94. Three factors, therefore, explained the variance of 31 of the 34 syntactic complexity variables.

Second, the remaining 17 factors characterized primarily “one type of syntactic structure and one or more part of speech categories or anaphora which usually

accompany that structure" (Bormuth, 1969, p. 34). According to Bormuth, this second pattern of factor loadings suggested that there was very little variance shared within the part of speech, syntactic structure, and anaphora variables when each was considered separately. Therefore, he subjected 29 part of speech variables, 19 structure variables, and 8 anaphora variables to separate factor analyses using Joreskog's maximum likelihood method with the probability of a solution's fit set at .20. From the part of speech analysis (29 variables included), 12 factors emerged. Most of these factors had only one variable loading highly (e.g., .8 or .9) and any other variables that loaded on the factor had much lower loadings (e.g., .2 or .3). In addition, 13 of the 29 variables had unique variances of .7 or higher. Four factors surfaced in the analysis of the 19 structure variables and 14 of the 19 variables had unique variances of .7 or higher. Bormuth did not offer details of the anaphora factor analysis results but wrote that they were similar to those of structure variables. From the results of these factor analyses, Bormuth concluded that a simple structure does not likely underlie variables that are correlated with passage difficulty.

These sets of factor analyses results led Bormuth to further question how many of the emerging factors were required to sufficiently account for the variance in passage difficulty. That is, it is possible that some factors might not be correlated with passage difficulty. Therefore, Bormuth (1969) calculated correlations between ten of the factor scores and passage difficulty (he did not indicate how or why he chose those ten factors). He first calculated factor scores for the ten factors and then regressed them, using stepwise, polynomial, multiple regression on passage difficulty. All ten factor scores were significantly correlated with passage difficulty, but none accounted for more than 26% of the variance in difficulty alone. An orthogonal rotation was performed in the

initial factor extraction; therefore, the correlations between each factor score and passage difficulty should be regarded as a partial correlation. That is, it is the correlation between a factor score and passage difficulty after partialing out the effects of the other factors.

From the results of the first phase of his investigation, Bormuth (1969) concluded that explaining the language comprehension process was a more complex endeavor than he had anticipated. Specifically, many more variables showed significant relationships with passage difficulty than he had predicted and many more variables might be uncovered in future research. Syntactic complexity appeared to affect comprehension and the effects were independent of syntactic structure effects. Syntactic complexity also revealed itself to be more complicated than Bormuth surmised. According to Bormuth, future measures of complexity should necessarily be devised to "take into account the possibility that comprehension involves the memory of structures which are not yet completed at a given point in the sentence as well as the anticipation of structures begun but not yet completed" (p. 41). Clause lengths used to measure syntactic length also showed differential effects. When syntactic length was measured in syllable units, the correlation with passage difficulty was higher than when the syntactic length was measured in letter or word units. Bormuth held that this offered evidence that words have a complexity similar to that of sentences. Because of the complexity of the first phase of his investigation, Bormuth indicated that when designing a readability formula, one must balance the need for face validity, practical utility, and predictive validity.

Bormuth (1969) conducted another set of studies to develop a set of readability formulas for use with scientific materials, machine analyses, manual analysis by skilled users, and manual analysis by unskilled users. He intended these formulas to consider

difficulties of passages and individual words and sentences. He focused on individual words and sentences, as well as passages, because readability formula users had previously used readability formulas designed for passages to determine the readability of smaller units of text (e.g., sentences). According to Bormuth, the use of readability formulas designed to assess whole passages on smaller units of text led to erroneous conclusions. This inappropriate use likely introduced systematically biased estimates. Specifically, although average language counts tend to be normally distributed, most individual language unit counts are skewed and leptokurtic. Therefore, employing formulas based on the average measures is inappropriate.

Bormuth used 330 100-word passages, five cloze tests for each passage, and 35%, 45%, and 55% cloze criterion scores in his investigation. He first scaled the passages to assign grade-placement scores for the 35%, 45%, and 55% cloze criterion and to calculate and plot a general function (i.e., passage grade-placement formula). This general function produced passage grade-placement scores when any of the three criterion scores and a cloze mean (estimated by one of the formulas he created) was entered into the equations. He created readability formulas that estimated cloze means and formulas to estimate grade-placement scores for all three criterion scores because he was unsure which of the criterion scores (e.g., 35%, 45%, or 55%) was most appropriate.

To assign grade-placement numbers to each of the passages, he first analyzed each passage independently by correlating students' cloze percentage scores with their reading achievement scores. Bormuth (1969) employed a stepwise polynomial regression model because some of the regressions were curvilinear. Then he used the polynomial regression equation to determine predicted grade-placement scores that corresponded to

the cloze percentages, whereby he obtained grade-placement scores related to the 35%, 45%, and 55% criterion scores for each passage.

Bormuth (1969) then computed grade-placement formulas for each passage. This formula delivered the grade-placement for a passage given its cloze mean and the chosen cloze criterion. Three sets of scores were associated with each passage: cloze mean (M), criterion scores (C), and grade-placement scores (GP) that corresponded to each criterion score. He created the formulas by calculating stepwise multiple regressions: GP scores were the dependent variable and M and C scores and the powers of their cross-products were the independent variables.

Bormuth (1969) found curvilinear relationships between cloze and reading achievement scores for most of the passage regressions (i.e., 303 of 330 passages). The 35%, 45%, and 55% criterion grade-placement scores provided significant estimates of passage difficulty as shown by their intercorrelations, none of which were below .915. The passage grade-level placement formula fit the data well: r = .978, SE = .61 for grade-placement scores calculated with the formula and the grade-placement scores calculated from the cloze and achievement test scores. The equation was: GPest = 4.275 + 12.881M – 34.934 M^2 + 20.388 M^3 + 26.194C – 2.046 C^2 – 11.767 C^3 – 44.285MC + 97.620$(MC)^2$ – 59.538$(MC)^3$. With this set of analyses, Bormuth satisfactorily established dependent variables for calculating readability equations for cloze criterion scores of 35%, 45%, and 55%.

Bormuth (1969) went on to calculate passage-level, sentence-level, and word-level readability formulas using stepwise multiple regression. Many of the linguistic variables

that Bormuth included showed curvilinear relationships with difficulty. Therefore, he included the linguistic variables and their squares, cubes, and first powers.

Bormuth (1969) created four sets of passage-level formulas: 1) unrestricted, 2) short form of the unrestricted, 3) manual computation, and 4) machine computation. He calculated the first set (unrestricted) with only statistical restrictions for variables entering the equation. The other three sets of formulas (short form unrestricted, manual computation, and machine computation) were created for use by people with different levels of technical skills, available equipment, and materials. For each of the four sets of passage-level formulas, Bormuth created four separate formulas. One formula was for estimating cloze means for passages and the other three were to estimate the grade-placement (GP) difficulty scores derived by scaling passages according to the 35%, 45%, and 55% criterion scores. Formula users could then use any of the latter three formulas to estimate readability based on their own choice of criterion. All four unrestricted formulas were linearly related to the difficulty levels of the passages upon which they were calculated.

Bormuth (1969) created short forms of the unrestricted formulas because the unrestricted formulas were very long and included many variables (i.e., 19 variables for cloze mean, 20 variables for GP 35%, 18 variables for GP 45%, and 15 variables for GP 55%). For practical use, shorter formulas were likely to introduce less error due to mistakes made by a practitioner. He selected 10 linguistic variables for inclusion in the short forms according to their correlations with difficulty, the number of unrestricted formulas they entered, and how frequently they occurred in the passages (where relevant). For the sentence-level formulas Bormuth (1969) excluded anaphora and

structure frequency variables used in the passage-level formulas because they were not appropriate for use at the sentence level. In addition, he collapsed some of the parts of speech variables to create a smaller number of categories because not all parts of speech occur with enough frequency at the sentence level. He collapsed 61 parts of speech into 15 variables.

Bormuth (1969) originally planned to create four sentence-level formulas: unrestricted, short form of the unrestricted, manual computation, and machine computation. The short form of the unrestricted formula was adequate for machine computation and the two formulas were moderately correlated (r = .645). In addition, Bormuth did not think it was appropriate to calculate formulas for estimation of sentence grade placement. That would have required each student to receive a score on each sentence and each of those scores would be based on a limited number of responses. Therefore, the results would not have been reliable. Consequently, Bormuth created a total of three formulas designed to estimate cloze mean: 1) unrestricted; 2) machine computation/unrestricted, short form; and 3) manual computation.

Bormuth (1969) found that minimal punctuation unit formulas and sentence-level formulas were redundant. That is, the formulas created for minimal punctuation units were almost identical to the sentence-level formulas. Sentence-level formulas, on the other hand, had higher validity levels than minimal punctuation unit formulas. This was likely due to minimal punctuation units being comprised of fewer words than sentence-level variables and therefore having lower reliabilities.

It was necessary for Bormuth (1969) to create two types of word-level formulas. In the first type, he considered contexts of sentences, which required considering the

syntactic context of words, the syntactic functions of words, word positions within a sentence, and characteristics of words. With the second type of word-level formula, Bormuth addressed word difficulty without consideration to context.

To create a word-level formula that addressed context, Bormuth (1969) collapsed all of the part of speech variables into two categories: structural and lexical words. Both formulas were moderately valid and showed correlations with difficulty of .532 for words with context and .522 for isolated words.

According to Bormuth (1969) his passage-level formulas were significantly more accurate than the Dale-Chall (1948) formula, which had been previously regarded as the best predictor of difficulty. He asserted that his passage-level formulas still required improvement because the best formula accounted for 85% of variability in difficulty. He thought that the other 15% should be accounted for in future research. In addition, he contended that the formulas, regardless of their accuracy, lacked the validity necessary to ensure that the results were unquestionable. He wrote,

> For example, the machine computation formula seems to assert that passages containing short words which all appear on the Dale List of 3,000 Easy Words and which contain only short sentences not incorporating modal verbs will necessarily be easy to understand. Yet nearly any experienced writer can easily produce passages which fit all of these specifications yet which are extremely difficult to understand. (p. 72)

Bormuth, therefore, contended that readability formulas, based on the assertion that short words and sentences result in more readable passages, had the potential to produce misleading results.

Bormuth (1969) warned that his sentence-level formulas should be used with great caution for two reasons. First, formula validity was not tested. Second, the sentences used to create the formulas were parts of larger passages and using the formulas to determine the readability of sentences in isolation would be done without empirical or logical support. In addition, Bormuth explained that sentence readability predictions should not be converted to grade-placement scores. The grade-placement-transformation formula was devised for use with passages and is therefore not suited for the transformation of sentence-level scores. He extended all of these sentence-level cautions to the word-level formulas.

Fry.

While serving a lectureship in Uganda, Edward Fry created a readability formula and corresponding readability graph to assist teachers and editors in helping people learn to read. The graph was intended to be used as a tool to assist in the selection of texts of appropriate difficulty levels. The original formula was published in a British Journal, *Teacher Education* (1963), which is unavailable to the public. In later research, Fry continued to improve his graph and described the simple method he used to devise it.

Fry (1968) offered a simplistic explanation of how he created his uncomplicated graph. He explained that to design his *Graph for Estimating Readability*, he plotted the grade levels (according to publishers) of a number of books (he did not indicate how many), found clusters, and then smoothed the curve. He then adjusted grade levels according to the results of correlational studies. Unfortunately, in none of his writings did Fry offer more specific details about his design methodology.

The readability graph that Fry first published in 1963 and discussed in further detail in 1968, was designed to estimate the readability of grade 1 to grade 12 books. With "considerable trepidation" (Fry, 1977, p. 251), Fry later extended his graph to level 17 through extrapolation based on the preexisting levels 10, 11, and 12. He extended the graph to include these higher levels in response to requests for a measure suitable for college-level materials. Because he had no data to assist in determining actual differences between level 13, 14, and 15, Fry cautioned that estimates at those levels should not be considered normed scores. Rather, they should only be used to determine relative difficulty between higher-level texts. He explained the determining college norms was especially problematic because texts at that level tend to be highly content specific and motivation levels might play a greater role at the college level.

Fry's (1963, 1968, 1977) readability graph includes two variables for the estimation of readability: average number of sentences per 100 words and average number of syllables per 100 words. Average sentence length offers an estimate of syntactic difficulty, while word length (measured by Fry with syllable counts) offers an estimate of vocabulary difficulty. The average number of syllables per 100 words is represented along the X-axis of the graph and ranges from 108 to 172. Average number of sentences per 100 words is represented along the Y-axis and ranges from 3.6 to 25.0.

To estimate the readability level of a book using the graph, three passages of 100 words are extracted from near the beginning, middle, and end of the book. The sentence and syllable variables are then measured for each passage and an average is determined for both variables. The corresponding values on the X and Y-axes are then located and

the point on the graph where the two converge signifies the estimated difficulty level of the book.

Fry (1968) contended that his graph was accurate "probably within a grade level" (p. 514) and explained that he viewed validation of readability formulas to be particularly difficult because there is no established standard to identify what constitutes difficulty for a specific grade level. He pointed out that publishers and educators have a general agreement about grade-level designations based on test data, but that even standardized test data differ in their designations. According to Fry, the most desirable alternative is to rank order texts based on comprehension test scores. This approach, however, is limited by the possibility of the texts themselves having differing difficulty levels, but nonetheless offers the most attractive alternative.

Fry (1968) offered validity evidence through the results of a comparative investigation conducted by one of his graduate students, Kistulentz (1967). Kistulentz analyzed 10 tenth-grade English class books and constructed comprehension tests for those books. He calculated rank-order correlations between the results of the Fry graph, Dale-Chall formula, and Flesch formula (among others). The formulas correlated well with each other (Fry and Dale-Chall: $r = .94$; Fry and Flesch: $r = .96$; Dale-Chall and Flesch: $r = .95$; $p < .01$) and with the results of the comprehension tests (Fry, $r = .93$; Dale-Chall, $r = .90$; Flesch, $r = .94$; $p < .01$).

The Dale-Chall formula tended to rank the books as moderately more difficult than the Fry graph. Fry (1968) originally surmised that this was because the Dale-Chall formula was devised 20 years prior to the Fry graph and that readers were less skilled at that time. Later, Fry (1977) reported that the reason for the more difficult ratings of those

books was that he had, in 1968, erroneously advised that proper nouns not be counted. When proper nouns were included in readability estimates using the Fry graph, the results corresponded more closely with results from the Dale-Chall formula.

Fry (1968, 1977) admitted that his graph tended to result in slightly less accurate results than the Dale-Chall and Flesch formulas. Nevertheless, he contended that, regardless of the slight loss of accuracy, the Fry graph might still be preferable to the other formulas because of ease of use. He cited Klare (1974-1975) who wrote,

> Unless the user is interested in doing research, there is little to be gained from choosing a highly complex formula. A simple 2-variable formula should be sufficient, especially if one of the variables is a word or semantic variable and the other is a sentence or syntactic variable…If the count is to be made by hand, counting syllables in some fashion…is somewhat faster than using most word lists (p. 244).

McLaughlin.

McLaughlin (1969) published *SMOG Grading—A New Readability Formula,* in which he presented a readability formula that he contended was even simpler to use than Fry's (1963, 1968) readability graph. He agreed with the readability scholars before him that semantic (word length) and syntactic (sentence length) variables held the most predictive power for readability estimations. Like Gunning (1952) and Flesh (1948), McLaughlin employed syllable counts to measure semantic difficulty.

Although McLaughlin (1969) attended to the same variables in his approach to readability estimation, his view of the relationship between these variables, how they affected readability, the form that the formulas should take, and the methods that should

be used to measure the variables differed from scholars before him. Specifically, he held that semantic and syntactic variables were not isolated constructs and instead interacted with one another. McLaughlin wrote, "A slight difference in word or sentence length between two passages does not indicate the same degree of difference in difficulty for hard passages, as it does for easy passages" (p. 640). He, therefore, contended that the usual form of readability formulas (i.e., $a + b$ (word length) + c (sentence length)) was inappropriate. McLaughlin thought that formulas would more appropriately take the following form: $a + b$ (word length x sentence length). This type of formula was simpler than what had been previously used: it had one fewer constant.

McLaughlin (1969) went further and devised a method to eliminate the need for multiplication of the semantic and syntactic variables. He explained that instead of measuring each variable and multiplying them by one another and then by a constant (b) and adding another constant (a), one could simply count out a fixed number of sentences and count the number of syllables within those sentences. He supported this idea by pointing out that an average number of syllables per word would increase as sentence length increased and as sentence length increased word length would increase. McLaughlin (1969) wrote:

> For any given average number of syllables per word, the count will increase if the sentence length is increased; likewise, for any given average number of words per sentence, the count will be greater if the word length is increased. (p. 641)

In addition, McLaughlin (1969) proposed a simpler method to count syllables than had been used by scholars before him (e.g., Flesch, 1948). He devised a means of establishing the number of syllables in a passage without counting each one. Instead, he

counted the number of words comprised of more than three syllables (polysyllabic) in a passage, multiplied that number by three, and added 112. This offered a practical alternative to the time-consuming task of counting out each syllable in a passage.

McLaughlin (1969) also contended that the constant *b* could be eliminated by selecting a specific number of sentences, instead of words, to be counted. Through trial and error he established that 30 sentences were appropriate. This was in contrast to the 100-word samples that had been used by the majority of readability scholars in their readability estimations. With formulas that called for 100-word samples, several samples were necessary. Whereas McLaughlin's 30-sentence sample was taken in three groups of 10 consecutive sentences from different parts of a text and more than 600 words were typically included. This larger sample negated the necessity for several samples and, according to McLaughlin, increased the reliability of estimations.

Returning to his newly devised method for syllable counts (i.e., counting the number of polysyllabic words in a passage, multiplying that value by three, and adding 112) McLaughlin (1969) recognized that the value obtained by his method needed to be converted into a number that would be meaningful for formula development. To that end, he used 390 passages from McCall-Crabbs *Standard Test Lessons in Reading* (1961) and their respective comprehension questions. Scholars before him had typically employed 50% (e.g., Dale-Chall, 1948 & Powers et al., 1958) and 75% (e.g., Thorndike, 1916) correct responses from a respective grade level as indicators of adequate comprehension. McLaughlin elected to use "complete comprehension" (p. 642), or 100% comprehension scores, as an indicator of reading difficulty.

McLaughlin (1969) created four regression equations that related the polysyllabic word counts of each passage to the mean grade score of students who had successfully completed 100% of the corresponding comprehension questions. The first equation, $g = 6.2380 + 0.0785\ p$ (p = polysyllabic word count), resulted in predictions that correlated with the criterion at $r = .71$. Regardless of the high correlation with the criterion, this equation was only suitable for predicting readability above the 6th-grade level and involved a multiplication operation that was more difficult than McLaughlin desired. The second equation, $g = 4.1952 + 0.8475\ \sqrt{p}$, also involved a multiplication operation that was more complicated than what McLaughlin had in mind. The third equation, $g = 2.8795 + .9986\ \sqrt{p} + 5$, had a simpler multiplication operation but required more addition. McLaughlin, therefore, established a fourth equation that was a compromise between the second and third equations: $g = 1.0430\ (3 + \sqrt{p})$ or $g = 3.1291 + 1.0430\ \sqrt{p}$. For practical purposes, he simplified this equation to the following: $g = 3 + \sqrt{p}$. This was a far less complex than any formula that had been previously devised.

McLaughlin (1969) was satisfied with the final regression equation for two reasons. First, the standard error of estimate was 1.5, which offered sound validity evidence. Second, it was so easy to use that it merely required about nine minutes to estimate the readability level of a 600-word passage. According to McLaughlin, this was considerably more efficient than the Dale-Chall formula, which he estimated took about the same time to estimate the readability level of a single 100-word passage.

McLaughlin (1969) tested the predictive power of his polysyllabic word counts and the formula he devised that included them. In so doing, he had 64 university students read eight 1,000-word passages from periodicals and complete a free recall test of content in

each passage. Participant responses were rated for comprehension on a scale from 0 to 10. He monitored, but did not control, reading time and adjusted comprehension scores accordingly. Specifically, because participants who took longer to read the passages tended to perform better on the recall task, McLaughlin divided participant comprehension scores by the time they took to read the corresponding passage. The results showed a perfect negative rank correlation between polysyllabic word counts and reading efficiency (i.e., comprehension score divided by time). The SMOG grade levels yielded from each passage also corresponded to reading efficiency. McLaughlin interpreted these results to indicate that the count of polysyllabic words in a fixed number of sentences offered an accurate index for the relative difficulty of texts and that his final formula offered acceptable results.

Caylor et al.

Caylor et al. (1973) were involved in research focused on determining literacy skill requirements for US Army occupations. As part of their work with the Army they developed the FORCAST readability formula to estimate the readability of materials used during training and job performance. The Army's printed materials were different from any other materials for which readability formulas were previously created. Therefore, traditionally used readability formulas were not suitable. Caylor et al. explained that the other formulas were not suitable for two primary reasons: 1) the Army materials had a distinct style, format, and were laden with technical language and 2) most consumers of these materials were adult, male soldiers.

To create the formula, the researchers selected 12, 150-word passages from reading materials used by Army personnel in preparation for qualifying examinations for seven

jobs. They analyzed the passages according to 15 structural properties, including number of sentences, words per sentence, one-syllable words, letters per sentence, and independent clauses. To appraise reader comprehension of the 12 passages, Caylor et al. (1973) assessed the reading comprehension of 200 men on the passages using the cloze technique. With the data from the cloze tests and previously determined reading levels of the 200 participants, the researchers scaled the 12 passages according to reading grade level (RGL). Specifically, they established the lowest RGL at the point where 50% of the participants scored the standard 35% correct or better criterion on the cloze test for each passage.

The next phase of their research involved Caylor et al.'s (1973) development of a regression equation including the 15 structural properties to predict scaled RGLs for the passages. They analyzed the intercorrelations among the 15 individual and combined properties with the cloze results. With the results, the researchers determined that the number of one-syllable words per 150-word passage was as useful as any of the other, more difficult to apply, structure factors. The correlation between the number of one-syllable words per 150-word passage and the RGLs that corresponded to the 35% criterion was .87. Through regression analysis, Caylor et al. created a preliminary equation/readability formula: RGL = 20.43 – (.11) (number of one-syllable words per 150-word passage). The researchers were interested in developing a formula that was simple to use and, therefore, rounded 20.43 down to 20 and .11 to .10 and changed .10 to 1/10. They contended that this modification resulted in a minor, justifiable loss of fidelity. The final FORCAST formula, therefore, was: RGL = 20 – number of one-syllable words/10.

To illustrate the usefulness of the FORCAST formula, Caylor et al. (1973) applied it, the Dale-Chall readability formula, and the Flesch formula to the 12 passages. Correlations among the three formulas ranged from .92 to .97, which indicated that the formulas resulted in the similar rank orderings of the readability of the passages. The researchers also determined the correlations among the RGLs and the readability estimates for the passages derived from each of the formulas: Dale-Chall r = .93, Flesch r = .92, and FORCAST r = .87. The Dale-Chall and Flesch formulas overestimated the readability of the passages (as compared to the RGL) by 1.7 and 1.9 mean grade levels, respectively. The standard deviation of mean grade levels for the FORCAST formula was less than half the size of those corresponding to the Dale-Chall and Flesch formulas and .6 lower than that of the RGL.

Because the first study was conducted with the passages upon which the FORCAST formula was calibrated, Caylor et al. (1973) conducted a cross-validation study with a new set of passages and participants (they did not indicate how many participants or the text materials from which they drew the passages). The correlations among the results from the three readability formulas ranged from .94 to .98. The correlations among the formulas and the RGLs were as follows: Dale-Chall r = .86, Flesch r = .78, and FORCAST r = .77. In this case, the three readability formulas underestimated the readability of the passages (as compared to the RGL) by approximately one grade level. The standard deviation of the mean grade levels for the FORCAST was less than half the size of those corresponding to the Dale-Chall and Flesch formulas and .2 smaller than that of the RGL.

The initial study, in which they assessed the texts upon which the FORCAST formula was calibrated, resulted in a .87 correlation between the FORCAST formula readability estimates and RGLs. In their cross-validation study they used a new set of materials and participants and the correlation between the FORCAST formula readability estimates and the RGLs decreased to .77. Caylor et al. (1973) were not discouraged by this decrease. Instead, they contended that the results of the cross-validation study confirmed the validity of using the formula with job-related reading materials and indicated that it would be useful for pairing the reading ability of Army personnel and the reading demands of training and job-related texts.

Caylor et al. (1973) admitted that the FORCAST formula suffered from restriction of range. Specifically, if a text were comprised of all one-syllable words, the readability estimate would be grade 5. It was not possible for the formula to yield results at grade levels below grade 5. In addition, because the reading test that was used to calibrate the reading ability of the personnel was normed at an RGL of 12.9, this is the upper limit of readability estimates for the formula. Any estimates below 5.0 or above 12.9 would be derived through linear extrapolation. The researchers contended that the practical use of the FORCAST formula involves ordering texts according to difficulty level; therefore, the use of linear extrapolation should not be of concern.

Coleman and Liau.

Coleman and Liau (1975) published a readability formula that was designed to be machine scored. The primary purpose of their research was to illustrate that the previous methods used for syllable counts were not accurate or efficient. They, therefore, argued that predictor variables that lent themselves well to machine scoring were in order.

Coleman and Liau included measures of sentence complexity (average number of words per sentence) and word complexity (average number of syllables per word) because these variables had been shown to account for 60% to 80% of the variance in most readability formulas and could be reliably identified by an optical scanner.

To develop their prediction equation, Coleman and Liau (1975) used Miller and Coleman's (1967) 36, 150-word passages and data from their three cloze tests. With Miller and Coleman's data, Coleman and Liau computed equations with predictive variables of letters per 100 words (*L*) and sentences per 100 words (*S*). Their subsequent formula is: Estimated cloze % = 141.8401 - .214590*L* + 1.079812*S*. The multiple correlation for their equation and cloze percentage scores was .92. The authors explained that the high multiple correlation was not only due to the high predictive validity of the formula, but also to the large difficulty range of the passages (i.e., 1st-grade to very difficult prose).

Coleman and Liau (1975) recognized that some people (e.g., educators) would find it easier and more useful to interpret readability scores in terms of grade levels instead of cloze percentage scores. Therefore, the authors also provided a formula for transforming cloze percentage scores to grade levels: Grade level = -27.4004 estimated cloze % + 23.06395. According to Coleman and Liau, the correlation between cloze percentage scores and grade levels was -.88, hence little accuracy was lost in this useful transformation.

Homan, Hewitt, and Linder.

Not all texts lend themselves well to readability formulas, which generally require several 100-word passages for proper implementation (Allan, McGhee, & van Krieken,

2005; DuBay, 2004; Klare, 1984; Hewitt & Homan, 2004; Homan, Hewitt, & Linder, 1994; Oakland & Lane, 2004). For instance, readability formulas may not yield valid results for materials such as multiple-choice test items or documents with long word lists (Allan, McGhee, & van Krieken, 2005; Hewitt & Homan, 1991, 2004; Homan et al.; 1994). Hewitt and Homan (2004) and Homan et al. (1994), therefore, addressed the need for readability formulas for single sentences that occur in test items.

After a decade of work devoted to creating a readability formula to identify the readability level of single-sentence test items Homan et al. (1994) tested their formula. The authors asserted that test takers who are presented with multiple-choice questions (stems) and options written at readability levels potentially beyond their reading comprehension abilities cannot be assumed to understand "what is being asked" (p. 350) or to comprehend the correct and incorrect responses. Incongruence between the readability of items and test-takers' reading comprehension capacities could, therefore, influence item difficulty. Hence, Hewitt and Homan (2004) and Homan et al. (1994) contended that readability of multiple-choice items required consideration in test development.

The Homan and Hewitt formula included three variables: 1) number of difficult words (WUNF), 2) word length (WLON), and 3) sentence complexity (WNUM). The number of difficult words was determined by familiarity with *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981). Homan et al. (1994) considered a word *familiar* if it was familiar at the 4th-grade level for 80% of the students used to create the word list. All other words were considered unfamiliar. The second component in their formula, word length, was an indicator of vocabulary load.

Word length was established by counting how many words per sentence included more than six letters. The third component in their formula, sentence complexity, was determined by counting the average number of words per Hunt's T-Unit, which is a measure of syntactic complexity that considers the number of clauses per sentence.

Homan et al. (1994) used stepwise multiple regression with the three variables as predictors and readability level assigned to each sentence by their source as the criterion. The levels assigned by the sources were established through standardized norming procedures. The authors randomly selected 180 sentences from a 300-sentence sample. The authors did not report the source of this sample. The resulting formula was:

$$Y = 1.76 + (.15 * WNUM) + (.69 * WUNF) – (.51 * WLON).$$

The predictor variables were significantly related to the criterion variable (i.e., readability level of the 180 sentences): WUNF (unfamiliar words) R^2= .383; WNUM (average words per T-Unit) R^2= .460; and WLON (long words) R^2= .496. The 120 sentences not initially randomly selected from the 300-sentence sample were used for cross validation.

Homan et al's (1994) study was devised to validate their readability formula, which they contended could be used to accurately identify the readability of single-sentence test items. To test their formula, the authors used 1,172 2nd-, 3rd-, 4th-, and 5th-grade social studies students. In order to be selected for participation, students were required to pass a test of relevant content knowledge (social studies) and be able to read at grade level. The results from 782 of these students were used in the analysis and the results from the other 390 students were discarded because the students' reading levels were below average or their mastery of the material upon which the instruments were constructed was insufficient.

Homan et al. (1994) subjected each test item to the Homan and Hewitt readability formula. To do this, each option was combined with the stem and those combinations were considered separately with the formula. This resulted in four readability estimates (one per option) for each item. The average readability level for each item was designated as the mean readability level of four respective stem/option combinations. The four readability estimates were then averaged for each item to determine the average readability level of an item.

Homan et al. (1994) created four multiple-choice social studies tests from 84 items that consisted of 12 items at each of 7 readability levels (grades 2-8). They further divided the subgroups of 12 items into groups of 3 items that covered specific concept areas: taxes, scarcity, interest, and budget. Each test consisted of 48 items that represented four readability levels: 12 items at the student's grade level and twelve items from each of the three proceeding levels. Homan et al. balanced the items to ensure that items of a particular concept at higher and lower readability levels were of the same cognitive level. They determined that all of the items were at the knowledge or comprehension level. The authors divided the 2nd- and 3rd-grade tests into two 24-item tests (A and B) so that the tests could be given in two sittings because they were concerned about the ability of younger children to complete longer tests.

The authors used a two-factor mixed model analysis of variance with grade level as the between-subject factor and readability level of items as the within-subject factor to analyze the data (class means) from the 782 students retained in the study. The results revealed a significant interaction between grade level and readability level ($p < .0001$). That is, as readability level of test items increased, mean class scores at grade levels

decreased. Scheffé post hoc comparisons showed significant differences ($p < .05$) between all possible combination of readability levels and class mean performance.

Homan et al. (1994) interpreted the findings to suggest that the readability level at which an item is constructed directly affects student performance. That is, student performance is negatively affected by items being written at readability levels above which the students are operating. This lends support to the utility of the Homan and Hewitt formula as a readability measure for single-sentence test items.

T-units.

The Homan and Hewitt readability formula involves the division of passages into T-units, instead of sentences, to measure syntactic complexity. Because the Homan and Hewitt readability formula was the first to incorporate the use of T-units, a mere description of how T-units are defined will likely not elucidate why the authors selected it as their unit of measure. Therefore, this section includes a discussion of research related to the use of the average T-unit length as an index of syntactic complexity.

To determine syntactic, or sentence, complexity, Homan and Hewitt counted the average number of words per Hunt's (1965) minimal terminal unit (T-unit). T-Units are a measure of complexity that considers the number of clauses per sentence. Hunt (1965) introduced the concept of the T-unit in 1965. He explained a T-unit as "a grammatically discrete unit intervening in size between the clause and what is punctuated as a sentence" and further defined a T-unit as "one main clause plus the subordinate clauses attached to or embedded within it" (p. 49). Because the ways in which clauses have been defined differ among linguistic researchers, it is important to note that in his investigation, Hunt

defined a clause as "a structure containing a subject (or coordinating subjects) and a finite verb phrase (or coordinating verb phrases)" (p. 40).

Hunt (1965) conducted a quantitative study of grammatical structures and investigated developmental trends in the writings of 4th-, 8th-, and 12th-grade students. He collected 1,000-word, in-class writing samples from 54 average-intelligence 4th-, 8th-, and 12th-grade students (nine boys and nine girls from each grade). During his investigation, he also compared the writings of these school children to those of adults with superior writing abilities (i.e., authors of *Harper's* and *Atlantic* magazines) to identify how much and in what ways their writings differed.

For each student, Hunt (1965) collected the following data from the 1,000-word texts: 1) mean clause length (*w/c*), 2) mean T-unit length (*w/T*), 3) mean sentence length (*w/s*), 4) ratio of mean number of clauses per T-Unit (*c/T*), and 5) ratio of mean number of T-Units per punctuated sentence (*T/s*). Hunt's calculations for these variables are included in Table 4. Hunt conducted a 2x3 factorial analysis of variance for each dependent variable listed above with sex and grade as the between subject variables. He then analyzed the variables with chi-square tests and calculated contingency coefficients for variables that were significant at the .05 level. He used the contingency coefficient technique to determine the best indicator of student grade level.

Hunt (1965) found that all of the variables were statically significant for grade at a .05 level or better (no adjustment for α per comparison was made). Contingency coefficients were significant for all of the variables except ratio of T-units per sentence. The contingency coefficients indicated that average length of T-units was the best indicator of

mature writing (.694), followed by average length of clauses (.616), ratio of clauses per T-unit (.496), and finally average length of sentences (.489).

Table 4

Hunt's (1965) variables and calculations

Variable	Calculation
Mean clause length (*w/c)*	*N* of *W*s in a *S* / *N* of *C*s
Mean T-unit length (*w/T)*	*N* of *W*s in a *S* / *N* of T-units
Mean sentence length (*w/s)*	*N* of *W*s in a *S*
Ratio of mean number of clauses per T-unit *(c/T)*	*N* of *C*s in a *S* / the *N* of T-units
Ratio of mean number of T-Units per punctuated sentence (*T/s*)	*N* of T-units in a *S*

Note. N = Number; W = Words; S = Sentence; C = Clause

Hunt further investigated T-unit lengths of the three grade levels and identified three groups of T-unit lengths: short (1-8 words), middle (8-20 words), and long (20 or more words). Students at all three grade levels wrote approximately the same number of middle length T-units; 4th-grade students wrote the most short T-units; and 12th-grade students wrote the most long T-units. Short T-units accounted for 43% of 4th-grade students' writings; 21% of 8th-grade; and 10% of 12th-grade.

Hunt determined that the use of short T-units was a good indicator of grade level with a chi-square of 52.87 and contingency coefficient of .70. In addition, the number of short T-units correlated with average T-unit length (r = .902) for all three grades. Therefore, he proposed that counts of short T-units could offer a more time-efficient means of

determining passage complexity or the maturity of the writer. For writings of older authors, he recommended increasing the cutoff for short T-units to 9 or 10 words.

In the second phase of his study, Hunt (1965) extended his investigation of developmental trends in writing to include texts written by "superior adults". He added to his study 1,000-word excerpts from articles in *Harper's* and *Atlantic* magazines. He collected data for the same variables and extended the statistical analysis he used in the first phase of his investigation to include the new data from the magazines. Once again, he found that all of the variables were statically significant for grade at a .05 level or better. Contingency coefficients were significant for all of the variables except ratio of T-units per sentence; but, with the inclusion of the magazine article excerpts, the contingency coefficients were different. The contingency coefficients indicated that average length of T-units (.73) and average length of clauses (.73) were the best indicators of mature writing, followed by average length of sentences (.64), and ratio of clauses per T-unit (.51).

Hunt (1965) revisited the categorical groupings of T-units according to length with the extended data set. He determined that the same trend established with the school children continued with superior adults. Short T-units accounted for a scant 6% of superior adults' writings, which means that they wrote 59% as many short T-units as the 12^{th}-grade students. In addition, compared to the 12^{th}-grade students, the superior adults wrote 51% as many middle-length T-units and 169% as many long T-units.

The result of the statistical analysis of the data that included the magazine article excerpts indicated that average length of T-units and average length of clauses were equally good indicators of mature writing. Hunt (1965) inspected the percentage

increases for each variable among grade 4, grade 8, grade 12, and superior adults. He found that the largest percentage increase from 12th-grade to superior adult was for average T-unit length (40%) and that increase was largely due to the increase in average clause length (36%), not an increase in the use of subordinate clauses. He further identified that, in terms of percentages, the increase in average clause length from 12th-grade students to superior adults (36%) was greater than the increase from 4th-grade students to 12th-grade students (23%). On the other hand, the increase in average T-unit length from 12th-grade students to superior adults (40%) was equal to the increase from 4th-grade students to 12th-grade students. Therefore, Hunt contended that, although average T-unit length and average clause length were equally good indicators of maturity or complexity of writings, average clause length revealed the most notable developmental difference in writing samples from 12th-grade students to superior adults.

From the results of his preliminary study, Hunt (1965) concluded that clause-to-sentence factors (i.e., T-unit and clause length) could be useful measures for matching appropriately difficult texts with readers. In addition, he contended that T-unit or clause length may be better indicators of syntactic complexity than sentence length in readability formulas.

In 1970, Hunt conducted a follow-up investigation to his 1965 study that included school children, average adults, and superior adults. He collected data from 250 students from grades 4, 6, 8, 10, and 12 (50 students from each grade); 25 men who had graduated high school but had no college education and were employed as firefighters (mean age = 32, median age = 29); and 25 adults who had published articles in either *Harper's* or *Atlantic* magazines. The school children were selected according to their scores on

standardized intelligence and achievement tests. Hunt's objective was to represent an approximately normal distribution of academic/ability level for each grade (1970a). For each grade, 17 students were assigned to the high academic/ability level (ranges: I.Q. = 116.9 – 117.5; mean score percentiles = 82.6 – 83.6), 16 students were assigned to the middle level (ranges: I.Q. = 100 – 101.3; mean score percentiles = 48.1– 50.2), and 17 students were assigned to the low level (ranges: I.Q. = 79.4 – 84.4; mean score percentiles = 16.8 – 18.4).

Instead of taking samples of free writing, Hunt (1970a; 1970b) had each participant engage in a rewriting activity developed by O'Donnell (1967). Restricting the topic of the writing in this way enabled Hunt to control what they said without affecting how they said it (Hunt, 1970b). Therefore, he was able to control for differences in content and focus the investigation on differences in the output of the writers according to how they recomposed the original text. The original text (O'Donnell) was a 32-word discourse about the manufacturing of aluminum. The sentences were as short as possible and contained only single clauses. Participants were asked to write the passage in a "better way" without deleting any information (Hunt, 1970b).

Hunt (1970a; 1970b) collected data for the same variables he used in his 1965 study: 1) mean clause length (*w/c*), 2) mean T-unit length (*w/T*), 3) mean sentence length (*w/s*), 4) ratio of mean number of clauses per T-Unit (*c/T*), and 5) ratio of mean number of T-Units per punctuated sentence (*T/s*). For the school children's writings, he conducted five 3x5 factorial analyses of variance with academic/ability level and grade as the between subject variables. Hunt (1970a) used Newman-Keul's post hoc test to follow-up statistically significant grade-level differences.

The following variables were statistically significant at $p < .01$: 1) mean clause length (*w/c*) for grade and academic/ability; 2) mean T-unit length (*w/T*) for grade and academic/ability; 3) mean sentence length (*w/s*) for grade; 4) ratio of mean number of clauses per T-Unit (*c/T*) grade, and academic/ability; 5) ratio of mean number of T-Units per punctuated sentence (*T/s*) for grade. The following variables were statistically significant at $p < .05$: 1) ratio of mean number of clauses per T-Unit (*c/T*) for grade by academic/ability interaction and 2) ratio of mean number of T-Units per punctuated sentence (*T/s*) for grade by academic/ability interaction and academic/ability. These findings should be interpreted with caution because Hunt (1970a, 1970b) may have had an inflated type I error rate. He did not report conducting correlations between any of the dependent variables prior to his analysis of variance and did not adjust his alpha levels for multiple comparisons.

Average length of T-units and average length of clauses showed to be the best indicators of mature or complex writing in Hunt's (1965) study. Sentence length in this study (Hunt, 1970a; 1970b) showed an irregular pattern from grade to grade. Hunt (1970a; 1970b) did not clearly indicate the results of his follow-up tests for every variable. Therefore, only the results for average length of T-units, average length of clauses, and average length of sentences are discussed in detail here.

Sentence length was significant for grade level at the .01 level, but the variable did not show an even correspondence from grade to grade. For instance, overall (combining all ability groups) mean sentence lengths were greater for grades 8, 10, and 12 than they were for grades 4 and 6, but grade 8 means were higher than grade 10 means and grade 4

means were greater than grade 6 means. Therefore, Hunt (1970a; 1970b) contends that sentence length is an unreliable indicator of grade level or mature writing.

Average T-unit length or average number of words per T-unit increased across all grade levels. In addition, the only academic/ability level interval that did not show an increase was between low and middle academic/ability level 4th graders. That is, average T-unit length for low and middle academic/ability level 4th graders were nearly identical. Considering the steady increase of T-units and the irregular increase of sentence length, Hunt (1970a) contended that average T-unit length was a better indicator of syntactic maturity or complexity than average sentence length.

Average clause length or average words per clause increased across all grade and academic/ability levels. Wilcoxon rank sum tests showed that the differences between high and low academic/ability level groups were statistically significant at the .05 level for every grade. Newman-Keuls post hoc tests indicated that the differences between each grade were statistically significant. Hunt, therefore, contended that with the use of a rewriting instrument, average clause length was an "extremely sensitive measure of some factor which is closely related to chronological age and mental ability" (1970a, p. 18).

To determine whether the trends he established with school children continued with superior adults, Hunt (1970a, 1970b) had 25 writers from *Harper's* and *Atlantic* magazines complete the same rewriting task. He analyzed their writings according to the same variables. Through inspection of means, Hunt found that the superior writers carried on the trends for all five variables. He did not report statistical findings for this part of the investigation.

The firefighters' writing also progressed in the same direction as the school children's writings. The 25 firefighters' average clause lengths and T-unit lengths were higher than the average 12th grader, but not significantly so. Conversely, according to the results of Wilcoxon rank sum tests ($p < .01$), the firefighters scored statistically significantly lower ($p < .01$) than the superior adults on the same two variables.

Hunt (1970a; 1970b) concluded that with his instrumentation, of the five variables considered, average clause length was the best indicator of chronological age and academic/ability level. He pointed out that it was sensitive enough to make statistically significant distinctions between each grade level and between high and low academic/ability levels.

Since Hunt's (1965, 1970a; 1970b) studies, T-units have been applied in a variety of research endeavors. For instance, linguistic researchers have used T-unit length as a measure of syntactic complexity (e.g., Baines, 1975; Golub, 1974; O'Donnell, 1975) and writing proficiency and growth for students for whom English is a second language (e.g., Ho-Peng, 1983;) as well as native English-speaking populations (e.g., Maimon & Nodine, 1978). Researchers have also used T-units as division points for the analysis of abstractness of a text (e.g., Dilworth, 1978; Freedman, 1980).

Lexiles.

The previous section included a discussion of several readability formulas developed by reading researchers throughout the 20th century. Measurement researchers have forged another line of readability research. The Lexile Framework is one of most well-known and respected methods of assessing readability from a measurement perspective. The Lexile Framework, in its development and constituent parts, is significantly more

sophisticated than readability formulas devised by reading researchers. In essence, the Lexile Framework involves two primary elements: construct-specification equations and Rasch model calibrations. To explain the Lexile Framework and how it functions, a description of the background research that led to the validation of construct-specification equations as well as an explanation of the Rasch model will be helpful. Therefore, in the next sections Stenner and Smith's (1982) and Stenner, Smith, and Burdick's (1983) research regarding construct definitions/specifications is first outlined. This includes a brief overview of the Rasch model. Then how construct-specification and Rasch model calibration are used in conjunction to estimate readability within the Lexile Framework are outlined. In that discussion a full explanation is offered for: 1) the components assessed, 2) the calibration equation, 3) the Lexile scale, 4) research conducted to test the Lexile equation, 5) interpretation of Lexile measures, 6) methods used to forecast comprehension rates using the Lexile Framework, and 7) error rates for text measures, reader measures, comprehension forecasting, linking standards, and how the errors combine. Finally, a recent development in the Lexile Framework that addressed error introduced by construct misspecification is discussed.

Construct specification.

Stenner and Smith (1982) devised and tested the use of construct-specification equations as a means to assess the construct validity of psychological instruments. According to the researchers, the influence of item-score variation on construct validity deserved attention that it had not been given in previous research. Stenner and Smith wrote, "Until the developers of a psychological instrument can adequately explain

variation in item scores (i.e., difficulty), the understanding of what is being measured is unsatisfyingly primitive" (p. 415).

According to Stenner and Smith (1982), construct theory testing had previously been largely approached with the study of between-person variation. With the exception of notable work done in cognitive psychology (e.g., Carroll, 1976; Pellegrino & Glaser, 1982; Sternberg, 1977. 1980; and Whitely, 1981), relationships between item characteristics and item scores were grossly neglected. Stenner and Smith concluded that prominent test developers such as Thurstone, Binet, Terman, and Goodenough had neglected the item characteristic and item score relationship through "historical accident and tradition" (p. 452).

Stenner and Smith (1982) discussed three advantages to analyzing item-score variation in the construct-validation process: 1) stating falsifiable theories; 2) higher generalizability of independent and dependent variables; and 3) enabling experimental manipulation. First, the authors explained that constructs measured by psychological instruments are generally given verbal descriptions. These verbal descriptions are typically inadequate for precisely defining constructs and determining whether they are appropriately measured. These verbal descriptions do not lend themselves to scrutiny or refutation. However, a deductive theory that emphasizes item-score variation could be delineated in a construct-specification equation that could, in turn, be confirmed or falsified.

Second, Stenner and Smith (1982) outlined that item scores tend to be more generalizable (i.e., reliable) than person scores. This is because when the person is measured the error term is divided by the number of items, whereas when the item is

measured the error term is divided by the number of people. Typically, psychometric studies involve a greater number of people than items; therefore, focusing measurement efforts on items rather than people reduces error.

Third, Stenner and Smith (1982) contended that analyzing item-score variation in the construct-validation process offers the advantage of having items, rather than people, serve as subjects. This makes experimental manipulation possible. The researchers pointed out that items are "docile and pliable" (p. 452) subjects that can be manipulated and measured without informed consent.

According to Stenner and Smith (1982), a particular instrument does not, in and of itself, operationally define the construct meant to be measured; instead, a corresponding construct-specification achieves that task. Therefore, their goal was to create and test the usefulness of a construct-specification equation. The researchers explained that such an equation, created through the regression of item scores on specified item characteristics, would represent a theory concerning the regularity with which a measurement instrument/procedure yielded consistent results. They contended that a construct-specification equation would offer a theory regarding item-score variation and offer a means to confirm or falsify the theory. In addition, the equation would offer a vehicle to test alternate theories. A construct-specification equation would supply two sources of information critical to determining the degree of construct validity related to an instrument/procedure: 1) the amount of variance in item scores explained by the model (R^2) and 2) a regression equation that identifies item characteristics useful in predicting item scores.

The degree of fit between the measurement observations and construct-specification equation predictions would allow one to ascertain the degree of construct validity represented in score interpretation. Specifically, confidence in the validity of score interpretations would be increased if the construct-specification equation explained a suitable amount of variance in item scores. A high R^2 would support the construct theory, while a low R^2 would provoke doubt in the theory under investigation.

As with any statistical analysis, residuals play an important role in Stenner and Smith's (1982) construct-specification equation model. These item residuals offer information useful in evaluating items and modifying the specification equation. Small item residuals would indicate that item scores were suffering from little confounding or unwanted ancillary-variable-influence on item-score variability; whereas, large item residuals would suggest that unspecified variables were unacceptably contributing to item-score variability. Item residuals determined through the construct-specification equation can inform decisions about which items to retain or discard. In addition, the item residuals can inform construct theory modifications that would potentially improve the construct-specification equation.

To illustrate the usefulness of construct-specification equations in providing an objective method of clarifying the elements that account for the complexity of an item set, Stenner and Smith (1982) analyzed data from the Knox Cube Test. The Knox Cube Test was designed to measure visual attention and short-term memory and requires participants to replicate an experimenter's sequence of block tapping. Four blocks are affixed two-inches apart on a board. Participants are asked to repeat two to seven block taps that vary according to sequence.

Stenner and Smith (1982) outlined the causes for information loss from short-term memory: interference and time decay. Interference occurs when new information is introduced and old information is pushed out of short-term memory (lost). Time decay refers to the idea that the longer a piece of information inactively resides in the short-term memory system, the more likely that information will suffer from decay or be lost. Stenner and Smith attempted to identify item characteristics (i.e., sequence and number of taps) that significantly contributed to difficulty and converged on: 1) number of taps (2 to 7); 2) number of reversals; and 3) distance covered.

The researchers computed item scores for 101 subjects ages 3 to 16. They ordered the items from least to most difficult (determined by Rasch item difficulties) and then examined each item to determine if it differed from others according to the above described characteristics (i.e., number of taps; number of reversals; and distance). Stenner and Smith (1982) calculated zero-order correlations with the item difficulties and item characteristics to determine the extent to which the defined item characteristics accounted for item difficulty. The results indicated that as difficulty increased, the number of taps increased (r = .94), the number of reversals increased (r = .87), and the distance covered increased (r = .95). They also discovered multicolinearity among the item characteristics: number of taps and distance covered, r = .90; number taps and number of reversals, r = .82; and distance covered and number of reversals, r = .90.

To generate and refine a construct-specification equation for the Knox Cube Test, Stenner and Smith (1982) conducted several regression analyses. Their first analysis involved hierarchical stepwise regression with the main effects for the hypothesized item characteristics entered into the equation first, followed by the three two-way interactions.

The main effects accounted for 93% of variance in item difficulty. The interactions did not significantly contribute to variance. They only explained an additional 3% of variance. In their second analysis, Stenner and Smith used stepwise regression, which revealed that distance covered and number of taps significantly contributed to item difficulty accounting for 93% of variance. Therefore, the researchers concluded that the construct-specification equation for the Knox Cube Test required inclusion of distance covered and number of taps; number of reversals did not make a significant contribution.

Stenner and Smith's (1982) regression analysis results offered statistical evidence that corroborated the hypothesized causes of information loss from short-term memory: interference and time decay. The distance-covered variable corresponds with interference, while number of taps corresponds with time decay. The researchers, therefore, interpreted the results to indicate that the construct-specification equation for the Knox Cube Test provided satisfactory evidence that the test was measuring what was intended.

In a follow-up investigation, Stenner, Smith, and Burdick (1983) further discussed and illustrated the usefulness of construct-specification equations as a means of establishing construct validity and made the first step toward the development of a readability measure that was based on measurement theory. They held, as did Stenner and Smith (1982), that the equations would offer a test fit between theory and observations (i.e., model and data). That is, if a construct-specification equation were to account for significant variation in item scores, then validity of the instrument as an operationalization of the construct theory could be inferred. On the other hand, if a construct theory delineated in a construct-specification equation failed to account for

significant variation in item scores, then the operationalization of the construct theory for that instrument would be questionable. This would limit the applicability of the theory.

Stenner, et al. (1983) explained that with their model it would be possible to define a construct with a specification equation; but, instruments are a compilation of items that are "bound by the equation" (p. 4). The researchers held that two tests can be assumed to measure the same construct if a fit can be established between a single specification equation and the respective observations of the test (i.e., scores). This would be case regardless of test names, presentation method, scoring procedures, scaling, or superficial appearances. In turn, tests that are purported to measure the same construct, might require different specification equations to explain significant variance in scores.

To illustrate the usefulness of the construct theory definition, Stenner, et al. (1983) applied their model to a theory for receptive vocabulary, which applies to pictorial representations of primary level English noun, verb, adjective, and adverb meanings. The receptive vocabulary theory centers largely on the notion that word knowledge is gained through contextual exposure. That is, words that most frequently appear in written or spoken language are the most likely to be known by examinees and words that tend to be localized to particular domains and are not widely used across domains are more difficult and less likely to be known. When frequency and dispersion across domains are equal, difficulty can be predicted according to whether the words are concrete or abstract. Specifically, according to receptive vocabulary theory, difficulty of vocabulary items can be ascribed to three characteristics of stimulus words: 1) common logarithm of word frequency in samples of written material; 2) the likelihood of encountering a word across multiple domains; and 3) the type of word (i.e., concrete or abstract). Based on the

construct theory that Stenner, et al. described for receptive vocabulary, one would predict that vocabulary could be scaled from easy to difficult; a construct-specification equation including the above variables could predict the location of a word on the scale; and variables that represent language exposure would correlate with person scores.

To test their construct theory for the receptive vocabulary theory, Stenner, et al. (1983) incorporated a modified Rasch model. Rasch is one of several probabilistic latent trait response models based on the logistic cumulative distribution. To establish trait level, examinee responses are not simply scored and summed. Instead, Rasch involves a search process in which, according to the characteristics of the items and how the characteristics likely influence behavior, the trait level that best explains examinees' response patterns is identified. The use of Rasch requires the assumption that all items are equally discriminating and participant guessing is not significant. Item difficulty (b_i) is the only nuisance parameter considered in the estimation of the parameter of interest: examinee trait level. Item difficulty is defined as the point on the ability scale at which an examinee at the same position on the continuum as the item has a 50% probability of answering the item correctly. The Rasch model represents examinee and item characteristics on the same scale; therefore, with its use the Lexile Framework positions reader ability and text readability on the same developmental scale (Stenner, Burdick, Sanford, & Burdick, 2006). For the Lexile Framework, the Rasch model was modified, whereby text difficulty was defined as the point on the reader ability scale at which an examinee at the same position on the continuum as the item would have a 75% probability of answering the comprehension item correctly.

Stenner, et al. (1983) used forms L and M of Dunn and Dunn's (1981) *Peabody Picture Vocabulary Test-Revised* (PPVT-R) to illustrate their construct definition theory. The authors of the PPVT-R contend that the instrument measures receptive vocabulary for Standard English. Each item in the test includes four black and white illustrations. The test administrator speaks a word to an examinee and asks him/her to select the picture that best represents the meaning of that word.

Stenner, et al.'s (1983) dependent variable was the Rasch item scale values for 350 words from the PPVT-R and their predictor variables were word frequency, dispersion, and abstractness. They established word frequency and dispersion values with reference to Carroll, Davies, and Richman's (1971) list of 5,088,721 words selected from schoolbooks used in 3rd- through 9th-grade. Carroll, et al.'s word list identifies how frequently each word appears in text according to category (e.g., mathematics, literature, art). Instead of using the log frequency of stimulus words from the list, Stenner, et al. used the log frequency of the "word family". Word families include the stimulus words as well as their 1) plurals; 2) adverbial forms; 3) comparatives and superlatives; 4) verb forms; 5) past participles; and 6) adjective forms. Dispersion was a measure of distribution of word frequencies over 17 subject categories. The authors scored dispersion on a scale from 0 to 1.0, where lower values indicated that the frequency of a word tended to be concentrated in fewer subject categories. Abstractness was scored dichotomously by two independent raters: tangible objects were identified as *concrete* and words that denoted concepts were identified as *abstract*.

Regression analysis results revealed that the construct-specification equation that Stenner, et al. (1983) created for form L of the PPVT-R explained 72% of variance in

item scale values. Frequency and dispersion were highly correlated (r = .848), indicating that higher frequency words tended to be more widely dispersed in subject content areas, whereas lower frequency words tended to be more concentrated in fewer subject content areas. Abstractness was moderately related to item scale values (r = .352) and was not related to frequency (r = -.033) or dispersion (r = -.081). The analysis results for the PPVT-R form M were very similar (R^2= .712).

Because the analyses of form L and M yielded nearly identical results, Stenner, et al. (1983) combined the data from the forms in an additional regression analysis. The regression analysis of the combined data yielded results similar to the individual form analyses: the construct-specification equation accounted for 71% of variance in item scale values. An additional benefit yielded by combining the data sets was a reduction in the standard error because the combined data set was twice the size of the individual data sets.

Stenner, et al. (1983) examined 50 additional variables for inclusion in the specification equation to determine if variance explained could be improved. They increased the number of variables to 8, 10, and 12, and found only negligible improvements in variance explained. This offered further support that they had identified the most critical variables in their specification equation.

Estimating readability under the Lexile Framework.

The work of Stenner and Smith (1982) and Stenner, et al. (1983) offered the foundation for the Lexile scale. Stenner and Burdick (1997) outlined how the Lexile Framework was devised to use construct-specification and Rasch model calibration in conjunction to estimate readability. The Lexile Framework marries the one-parameter

Rasch model and a readability formula (Stenner, Burdick, Sanford, Burdick, 2006). In their discussion, Stenner and Burdick (1997) explained: 1) the components assessed, 2) the calibration equation, 3) the Lexile scale, 4) research conducted to test the Lexile equation, 5) interpretation of Lexile measures, 6) methods used to forecast comprehension rates using the Lexile Framework, and 7) error rates for text measures, reader measures, comprehension forecasting, linking standards, and how the errors combine. The following section includes a brief overview of these Lexile Framework characteristics.

Components of the Lexile framework.

Stenner and Burdick (1997) explained that the Lexile Framework components were, in part, based on previous work of readability scholars (e.g., Chall, 1988; Carroll, Davies, & Richmond, 1971; Klare, 1963) as well as the work of measurement scholars (i.e., Stenner, Smith, & Burdick, 1983). According to the Lexile Theory, readability is influenced by the familiarity of semantic units and the complexity of syntactic structures. The Lexile Framework, therefore, incorporates semantic and syntactic measures: 1) word frequency and 2) sentence length, respectively.

To determine the best measure of word frequency, Stenner and Burdick (1997) used Carroll et al.'s (1971) word list (5,088,721 words). They calculated the mean word frequency of 66 of the reading comprehension test passages from Dunn and Markwardt's (1970) *Peabody Individual Achievement Test*. Through correlations between algebraic transformations of means and rank orders of items according to difficulty, they determined that log word frequency served as the best predictor for word frequency.

Therefore, log word frequency serves as the semantic component (word frequency) in the Lexile Framework.

To identify the best predictor of syntactic complexity, Stenner and Burdick (1997), once again, used 66 reading comprehension test passages from the Peabody Individual Achievement Test (Dunnand & Markwardt, 1970). They conducted algebraic transformations of the mean sentence lengths and correlated them with item rank order (according to item difficulty). Through their analysis they concluded that the best predictor of syntactic complexity (word length) was the log of the mean sentence length. Therefore, the log of the mean sentence length serves as the syntactic component in the Lexile Framework.

Calibration equation.

Stenner and Burdick (1997) then created a provisional (calibration) regression equation with the log of word frequency (semantic component) and the log of mean sentence length (syntactic component). They regressed data from the Peabody Individual Achievement Test (Dunnand & Markwardt, 1970) using the provisional regression equation and found that 85% of variance in the rank order of test items (according to difficulty) was accounted for by the semantic and syntactic component variables.

Stenner and Burdick (1997) then used the provisional regression equation to assign theoretical difficulties to 400 pilot test items. They then ordered the items according to difficulty and administered them to 3,000 students at grade levels two through twelve. The researchers used Rasch analysis to identify misfitting items, which they discarded. They then established observed logit difficulties for the remaining 262 items using Rasch analysis. Stenner and Burdick (1997) used the observed logit difficulties of the remaining

262 items to determine the final regression equation. They regressed word frequency and sentence length components on the observed logit difficulties and found a .97 adjusted correlation between the observed logit difficulties and the theoretical calibrations. The resulting equation was: Theoretical Logit = (9.82247 x LMSL) – (2.14634 x MLWF) – constant (LMSW = log of the mean sentence length; MLWF = mean of the log word frequencies).

Lexile scale.

In their description of the Lexile Scale, Stenner and Burdick (1997) explained that with the use of the MSCALE program for Rasch calibration, the mean item difficulty for a test is located at zero on the logit scale. If an item were moved to a test with a different mean item difficulty, the item would shift in its location on the logit scale, which violates *general objectivity*. General objectivity requires that a "scale value of a single object is independent of conditions" (Stenner & Burdick, p. 5). To meet general objectivity, scores earned on different tests must be tied to a common zero. Therefore, the researchers transformed the theoretical logit difficulties they obtained from the above equation.

In a series of five steps, Stenner and Burdick (1997) established a scale with a fixed zero. First, they identified low and high anchor points. Text from seven basal primers (1st-grade reading level) served as the low point and text from the Electronic Encyclopedia (workplace level; Grailer, 1986) served as the high point. Second, the researchers used the above equation to establish logit difficulties of the low and high anchor texts. The mean logit difficulty for the low anchor text was -3.3 and +2.26 for the high anchor text.

Stenner and Burdick's (1997) third step was to establish a unit size for the Lexile scale: 1/1000. A Lexile unit, or Lexile, equals 1/1000th of the difference between the readability of the low anchor and high anchor texts. Fourth, the researchers assigned a scale value to the lower anchor: 200. They elected not to use zero as the lower anchor to avoid negative Lexile values.

Fifth, with the information assembled in steps one through four, Stenner and Burdick (1997) established a linear equation to convert logit difficulty values to Lexile scale values (CF = conversion factor): (logit score + constant) x CF + 200 = Lexile text measure. The researchers then plugged the mean logit difficulties for the low and high anchor texts and their corresponding Lexile scores into the equation and solved for the constant (3.3) and conversion factor (180). Their final equation for transforming logit difficulties into Lexile units took the following form: [(Logit + 3.3) x 180] + 200 = Lexile text measure.

Testing the Lexile equation.

In the next phase of their research, Stenner and Burdick (1997) tested the final Lexile regression equation described above. They applied the equation to texts using a computer program designed my MetaMetrics (1995), which analyzed the prose according to semantic and syntactic components and reported Lexile measures. Stenner and Burdick analyzed 1,780 reading comprehension test items from nine nationally normed tests. They obtained Lexile calibrations for the reading comprehension items with the MetaMetrics program and correlated those calibrations with the empirical item difficulties provided by the test publishers. The empirical item difficulties provided by the publishers were derived in one of three ways: 1) observed logit difficulties from

Rasch or three-parameter analysis; 2) logit difficulties estimated from item p-values, raw scores means, and raw score standard deviations; or 3) difficulty rank order of items.

The researchers plotted and correlated the empirical item difficulties and Lexile calibrations and assessed the plots for curvalinearity and high residuals. They observed that the Lexile equation did not fit poetry or noncontinuous prose test items and, therefore, determined that the Lexile equation should only be used with continuous prose. The researchers removed all noncontinuous prose and correlated the continuous prose empirical item difficulty and Lexile calibrations.

Stenner and Burdick (1997) then realized another model misspecification problem: restriction of range in item difficulties. Some of the tests from which they extracted data covered a narrow range of reading levels, which resulted in restriction of range and deflated correlations between item difficulties and Lexile calibrations. The researchers, therefore, used a method proposed by Thorndike (1949) to correct the correlations for restriction of range. The correlations between the two arrays offered evidence that, "most attempts to measure reading comprehension…measure the common comprehension factor specified by the Lexile theory" (Stenner & Burdick, p. 14). Raw correlations ranged from .65 to .95; correlations corrected for restriction of range ranged from .75 to .97; and correlations corrected for restriction of range and measurement error ranged from .77 to 1.0. The grand means, computed on Fisher Z transformed correlations, for raw correlations, correlations corrected for restriction of range, and correlations corrected for restriction of range and measurement error were .84, .91, and .93, respectively.

In a second study designed to test the Lexile equation, Stenner and Burdick (1997) identified Lexile calibrations for 11 basal readers. The researchers established observed

difficulties for the primers by rank ordering them, between and within grade levels, according to reading levels assigned by publishers. In other words, they assigned the first unit in the first book for the first-grade a rank of one and the last unit in the last book of the eighth-grade the highest rank order.

For each unit in the series, Stenner and Burdick (1997) calculated correlations between the Lexile calibrations and observed difficulties and made restriction of range corrections. Raw correlations ranged from .54 to .93; correlations corrected for restriction of range ranged from .94 to .99; and correlations corrected for restriction of range and measurement error ranged from .97 to 1.0. The grand means, computed on Fisher Z transformed correlations, for raw correlations, correlations corrected for restriction of range, and correlations corrected for restriction of range and measurement error were .86, .97, and .99, respectively.

Stenner and Burdick (1997) argued that the way in which Lexile theory accounts for the unit rank ordering of the basil readers offered sound support for the theory because the readers differed in prose selections, developmental range, continuous prose type, and emphasized objectives. The researchers further claimed that the Lexile theory could therefore be deemed useful for texts from pre-primer to graduate school level material (i.e., –200 to 1800 Lexiles).

Interpretation of Lexile measures.

In the next section of their paper, Stenner and Burdick (1997) explained how Lexile measures should be interpreted. The researchers touted that the Lexile Framework offered criterion-referenced, rather than norm-referenced, interpretations for every measure. The criterion-referenced interpretations offer information about what a student can and cannot

do rather than simply offering information about how a student's abilities compare to those of a normed group. This offers parents and teachers valuable information to inform future instruction.

These criterion-referenced interpretations for the measures work as follows. According to the Lexile theory, a student is predicted to have a 75% comprehension score for a text when his/her own measure is equal to the text calibration. For instance, if a reader earns a 75% comprehension score on a text with a Lexile calibration of 500, then that reader is assumed to be operating at that level. Stenner and Burdick (1997) explained that because the theory can be used to identify student and text reading levels, it is useful in the selection of level-appropriate reading materials.

Forecasting comprehension rates under the Lexile framework.

Stenner and Burdick (1997) also outlined how Lexile theory could be used to forecast comprehension rates. When a student's ability measure and the Lexile calibration of a text correspond (e.g., both are 700), then the student is expected to correctly respond to 75% of the corresponding comprehension questions. Reader and text calibrations cannot always be expected to perfectly match. The researchers, therefore, explained how comprehension rates could be forecasted when mismatch between reader and text calibrations exist. When a reader's measure is higher than that of the text, the reader is forecasted to have a better than 75% comprehension rate and vice versa.

The question remains, how much different from 75% will the comprehension rate be when mismatch exists between reader and text calibrations? Stenner and Burdick (1997) offered theoretical and computational explanations. They explained that to obtain the comprehension rate, after adding the 1.1 logit offset, the difference between the reader

and text calibrations could be converted to logits with the 180 conversion factor and subjected to Rasch model calibration. The adverse effect of this procedure is that it yields biased results because calibrations for each slice within a text are not equal and variability within, and average difficulty of, a prose section affects its comprehensibility.

To address the above bias concern, Stenner and Burdick (1997) changed the conversion factor from 180 to 225 and devised the following equation:

$$Rate = \frac{e^{Eld+1.1}}{1 + e^{Eld+1.1}}$$

Where *Eld* is the "effective logit difference" given by the following:

$$Eld = \frac{Person\ Lexile\ Meausure - Text\ Lexile\ Measure}{225}$$

Measurement error.

In their discussion of measurement error, Stenner and Burdick (1997) explained that they found reliability coefficients and standard errors of measurement to be inadequate for estimating error in the Lexile Framework and, therefore, used resampling theory and corresponding standard errors of measurement to analyze the Lexile Framework. The researchers addressed four types of measurement error related to the Lexile Framework: text measure error, reader measure error, error related to forecasted comprehension rates, and error in test linking.

Stenner and Burdick (1997) began their discussion of text measure error with an explanation of how text calibrations are conducted in the Lexile Framework. To obtain a Lexile calibration for a book, they randomly sample 20 pages from the text and concatenate them into a text file. That file is entered into the Lexile Analyzer computer

software program, which divides or "slices" the text files into passages of 125 words and computes Lexile calibrations for each slice. The Lexile calibrations are then subjected to an equation that solves for the Lexile measure with a 75% comprehension rate. The program uses the calibrations for the 125 word slices as test item calibrations and estimates the measure for a 75% raw comprehension score.

The specific operations executed by the Lexile Analyzer are (Stenner, Burdick, Sanford, Burdick, 2006):

1. An auto-edit routine is performed on the text to remove unfamiliar characters, figures, tables, and other nontext features;
2. The text file is "sliced" into standard-sized paragraphs of 125 words;
3. Each word in the slice is looked up in a frequency dictionary based on a 550-million-word corpus and the mean of the log word frequencies is computed for the slice;
4. The log of the mean sentence length is computed for the slice;
5. The two variables (from steps 3 and 4 above) are entered into an equation that returns a Lexile calibration for the slice;
6. This process is repeated for each slice in the text file;
7. The test is then treated as a virtual test with the number of test items equal to the number of slices and the item calibrations equal to the slice calibrations;
8. A measure is then returned that answers the question, "How well would a reader have to read (in Lexiles) to answer correctly 75% of the imagined test items comprising this text?";

9. The answer to the above question is the text measure assigned to the text. (p. 312)

To determine the reliability of Lexile text measures the researchers used resampling procedures to simulate repeated measurements on the same book. To do this, the researchers sampled 49 calibrations (with replacement) from the 49 sliced calibrations and solved for the Lexile measure. With this method, each of the 49 sets of resampled slices differs from the original 49 slices. A replicate text measure is then yielded from each replication. The standard deviation of all of the replicate text measures is the standard error of measurement. The resultant standard error of measurement can then be used to determine the level of uncertainty associated with the location of the book in the Lexile Framework. Most texts measured by the Lexile Framework have standard errors of 30 to 40 Lexile units.

In the Lexile Framework, reader measurement error is also determined using resampling theory. Stenner and Burdick (1997) explained that with the use of resampling to determine reader measurement error, person-specific error values are not affected by other people's performance. The authors did not explain how they used resampling to estimate this source of error, but did contend that this method allowed them to account for method (items), moment (occasion and context), and method by moment interaction sources of error.

Stenner and Burdick (1997) also discussed the error involved in forecasted comprehension rates. The difference between text and reader Lexile scores can be used to forecast reading comprehension rates. As with text and reader error estimations, the Lexile Framework uses resampling theory to identify error in forecasted comprehension

rates. The researchers explained that because reader measures and text measures are involved in forecasting comprehension rates, error associated with both of those measures must be considered in the estimation of error rates for forecasting comprehension rates. It follows that reader and text measure errors in resampling are aggregated and contribute to variability in the resampling of forecasted comprehension rates. A confidence band is established around a forecasted comprehension rate by resampling a text and reader measure replicant and using those data to forecast a comprehension rate replicant.

In their discussion of linking standard errors, Stenner and Burdick (1997) explained how they derived the equation for converting target scores to Lexile measures. The researchers administered the North Carolina End of Grade (NCEOG) test and a Lexile test to 956 students. They counterbalanced the order and administered both tests within a two-week period. The researchers transformed the NCEOG scores to three-parameter item response scores and Lexile measures to one-parameter Rasch measures. Stenner and Burdick then plotted the transformed scores and fit a *sd* line (geometric mean of the two regressions) to the data. They used the *sd* line equation to establish the correspondence between the two sets of scores.

To determine the error involved in the linking of score scales, Stenner and Burdick (1997) used resampling theory. In their simulation, they fixed NCEOG items (not resampled) and let Lexile items and people vary (resampled). The researcher fixed NCEOG items to imitate real practice where standardized test scores are linked to the Lexile Framework. In such a case, the standardized test items would remain in use for the life of the test (approximately seven years). The Lexile test items, on the other hand, would be likely to vary between studies.

Stenner and Burdick (1997) explained the five steps used in the resampling procedure to compute the linking standard error. First, sample (with replacement) 956 people from the 956 person data set. Each person's Lexile response record is resampled and their replicate Lexile measure is computed. NCEOG response records are resampled and replicate NCEOG measures are computed. Second, the NCEOG scale scores and resampled Lexile measures are plotted. Third, the *sd* line is computed and a table is constructed to illustrate NCEOG scores and their corresponding Lexile measures. Fourth, steps one, two, and three are repeated 100 times. Fifth, the standard deviations for the 100 Lexile measures are computed and reported as the linking standard error. Small linking standard errors warrant confidence in the correspondence between target scale scores and Lexile measures. Conversely, large linking standard errors lead to doubt in that correspondence.

Stenner and Burdick (1997) also discussed the circumstance under which different sources of error combine. Under the Lexile Framework, error combines in two ways: reader measurement with text measurement error in forecasting comprehension rates and reader measurement with linking error. When a reader is assessed with Lexile items, his/her reader error adequately reflects the uncertainty of the measure. On the other hand, when a non-Lexile test is administered and the score is then linked to the Lexile Framework, the linking standard error contributes to the reader measurement error. To forecast reading comprehension rates, the difference between reader and text measures is considered. Therefore, errors for both measures contribute to the error involved in the comprehension rate forecast.

Measurement Error due to Construct Misspecification.

Stenner, Burdick, Sanford, and Burdick (2006) addressed an additional source of error under the Lexile Framework: theory misspecification. In addition, they asserted that although several sources of error exist in the Lexile Framework, readability estimates derived by it result in less uncertainty than older readability estimation methods (e.g., Dale-Chall, Flesch-Kincaid). According to Stenner, et al. (2006), this improved accuracy stems from the incorporation of the *ensemble interpretation* and *whole-text* processing (whole books are analyzed with no sampling) into the Lexile Framework. The ensemble interpretation is rather complex, and, therefore, deserves explanation.

According to Kintsch and Van Dijk (1978), a passage is comprised of a several macro propositions. Stenner, et al. (2006) contended that any of those global propositions could be used to develop a comprehension test item. The conglomerate of items and their contexts make up a test. A difficulty value exists for each item (member of the ensemble) and these difficulty values can be averaged to establish an *ensemble mean*. According to Lexile theory, ensemble means can be predicted from the semantic and syntactic characteristics of a passage. Incorporating ensemble means removes the details of an ensemble member (item) by averaging the details. Stenner, et al. wrote, "The result of the averaging is a new concept (ensemble mean) removed from the particulars of its creation and is the unit of text readability predicted by the Lexile Theory" (p. 313).

Stenner, et al. (2006) contended that with the ensemble interpretation, irrelevant details that are associated with individual items and introduce variability are removed. They asserted that the variability of item difficulty stems from three sources: 1) item writers' choices of macro propositions (some of which may be sampled multiple times);

2) item location on a test form; and 3) item difficulties established according to the performance of an examinee sample. The ensemble mean is the average of the three sources of error across all items in set, "the ensemble mean taken over all persons, items, and contextualizations is seen as the function of the semantic and syntactic features of the text, as operationalized in the Lexile Analyzer" (Stenner, et al., p. 314).

Because the Lexile Analyzer used in the Lexile Framework uses ensemble means, Stenner, et al. (2006) claimed that its use results in more accurate readability estimations. They investigated level of uncertainty associated with Lexile text measures. Specifically, they estimated the standard deviation component corresponding to Lexile theory misspecification with the use of ensemble means.

To investigate the uncertainty of the Lexile text measures, Stenner, et al. (2006) used reading assessment data records for 2,867 3rd-, 4th-, and 5th-grade students. Three item-writing teams developed comprehension questions for 30 text passages. Each team wrote a question for each of the passages, resulting in a total of 90 items. Stenner, et al. spiraled the items into three test forms that most closely corresponded to Rasch model theoretical item calibrations: grade levels three, five, and eight. The researchers then ascertained the correspondence between theoretical text calibrations and the 30 ensemble means to determine the level of theory misspecification and its effects on text measure standard errors.

Stenner, et al. (2006) regressed observed ensemble means on text calibrations and obtained a root mean squared error (RMSE) of 110L (i.e., 110 Lexiles). Because ensemble means were derived based on three items, they were not well estimated. Therefore, Stenner, et al. simulated data and added an error term to each theoretical value

[distributed ~$N(0,\sigma = 64L)$]. The researchers regressed the "true" (i.e., simulated) ensemble means on the text calibrations and obtained an estimated RMSE of 64L ($110^2 - 89^2 = \sqrt{4,308} = 64$). The RMSE of 64L indicated the average error at the passage/slice level when the Lexile theory was used to predict "true" ensemble means. Texts are comprised of a number of passages/slices (i.e., 125 word samples); therefore, a text of n_i passage/slices would have an expected error of $64/\sqrt{n_i}$. According to this formula, shorter passages will have larger standard error of measurement values. For example, if a text consisted of 625 words, its standard error of measurement would be 29L ($64/\sqrt{5} = 28.62$); whereas, if a text were made up of 3,625 words, its standard error of measurement would equal 12L ($64/\sqrt{29} = 11.88$). The researchers also illustrated the interpretation of the ensemble mean reduction in standard error by showing the differences in standard errors with the older Lexile method and their ensemble interpretation method. When applied to the same 12 texts, the older, resampling method resulted in standard errors ranging from 70 to 268; whereas the newer, ensemble mean method resulted in standard errors ranging from 2 to 9.

Stenner et al. (2006) concluded that the ensemble mean interpretation offered more accurate predictions than the previously used Lexile method that involved the use of raw item difficulties. Stenner, et al. showed that raw item difficulties used in early Lexile research (e.g., Stenner, et al., 1983; Stenner & Burdick, 1997) were insufficient for ascertaining the predictive power of construct theories. With their use of ensemble means to address error introduced by theory misspecification, Stenner, et al. removed details of ensemble members (items). By so doing, they removed variability introduced by irrelevant details associated with individual items. Consequently, Stenner et al. reduced

standard error to single digits and held that the error was small enough that uncertainty in text measures can be disregarded in many applications of the Lexile Framework.

When data fit the model (i.e., when reader and text data fit the Lexile Framework) the Lexile Framework enables reader and text variables to exist on the same scale. In addition, because the Lexile Framework uses Rasch model calibrations, it enjoys a benefit that is the cornerstone of item response theory: the property of invariance. Specifically, although they are measured on the same scale, reader and text characteristics are independent on one another.

The Proposed Study

The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999) clearly address issues related to the readability of test items. Standard 9.8, "In employment and credentialing testing, the proficiency level required in the language of the test should not exceed that appropriate to the relevant occupation or profession" (p. 99) and Standard 7.7, "In testing applications where the level of linguistic or reading ability is not part of the construct of interest, the linguistic or reading ability demands should be kept to the minimum necessary for the valid assessment of the intended construct" (p. 82 – 83) are particularly relevant. These two standards, taken together, focus attention on the degree to which the linguistic characteristics of test items may introduce construct irrelevant variance into a testing situation. Seldom, however, has the issue of readability been directly addressed and currently an industry-established method for determining the readability of test items does not exist.

The difficulty lies not in ignorance of the importance of readability issues in testing, but in the complexities surrounding the acquisition of valid estimates of the readability of

test items. Although readability formulas are useful for determining text difficulty, not all texts lend themselves to formulas use because the formulas generally require several 100-word passages for proper implementation (Allan, McGhee, & van Krieken, 2005; DuBay, 2004; Hewitt & Homan, 2004; Homan, Hewitt, & Linder, 1994; Klare, 1984; Oakland & Lane, 2004). Readability formulas, therefore, do not yield valid results for materials such as multiple-choice test items or documents with long word lists (Allan, McGhee, & van Krieken, 2005; Hewitt & Homan, 1991, 2004; Homan et al., 1994).

Readability estimates for licensure or certification examination items are necessary to establish that student/candidate learning/training materials, examination materials, and occupational materials are of equivalent readability levels. Incongruity among the readability levels of these sets of materials likely reduces measurement precision (i.e., increases measurement error). Estimations of these readability levels could identify incongruity or provide evidence of congruity among the readability levels of the learning, examination, and occupational materials. Therefore, these estimations could provide further construct-related validity evidence to credentialing testing programs.

There is currently no readability formula suitable for the occupational-specific language included in credentialing learning/training, examination, and occupational materials or the multiple-choice format of credentialing-examination items. Existing formulas cannot be confidently and reliably applied to learning/training and occupational materials related to credentialing because the materials generally include occupational-specific language that would likely skew the results. In addition, existing formulas cannot be confidently and reliably applied to credentialing examination items for three reasons. First, existing formulas (with the exception of the Homan-Hewitt) are only suitable for

several samples of continuous prose of 100 or more words. Second, the Homan-Hewitt formula is suited for multiple-choice items, but it was specifically developed for elementary-school-level text. Third, credentialing examination and related materials include occupational-specific content vocabulary that has the potential to skew readability estimates. In other words, job-related vocabulary for some occupations includes polysyllabic and, typically, unfamiliar words that would likely result in unduly high readability estimates. If the text were to be posed to a person outside the respective occupation, the high estimates would be appropriate; but, candidates taking credentialing examinations should be familiar with the occupational-specific vocabulary.

The purpose of this study is to develop a set of procedures to establish readability, including an equation, that accommodates the multiple-choice item format and occupational-specific language related to credentialing examinations. The procedures and equation should be appropriate for learning materials, examination materials, and occupational materials. To this end, variance in readability estimates accounted for by combinations of semantic and syntactic variables were explored, a method was devised to accommodate occupational-specific vocabulary, and new-model readability formulas were created and calibrated. Existing readability formulas were then recalibrated with the same materials used to calibrate the new-model formulas. The new-model and recalibrated formulas were then applied to examination items from a dental licensing program and the results were compared.

A three-phase investigation was conducted to create a new model appropriate for measuring credentialing materials: learning, occupational, and examination. Phase I, *Variables in the model*, involved identifying semantic and syntactic variables for

inclusion in the new model. During Phase II, *Formula calibration*, four new-model formulas were calibrated and three existing readability formulas were recalibrated with the same materials used to calibrate the new-model formulas. Phase III, *External validity and reliability evidence*, was designed to investigate how the new-model formulas performed with credentialing-examination materials.

The objective of the first phase of the investigation was to determine the variables to be retained for the second phase of the investigation. The Miller and Coleman (1967) passages and their corresponding total cloze scores were analyzed according to their semantic and syntactic variables. The semantic variable, number of unfamiliar words, was selected a priori but was further specified during this phase of the investigation. Specifically, regression techniques were used to investigate the effects of identifying unfamiliar words according to *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) at grade levels 4, 6, 8, 10, 12, 13, and 16. The Miller and Coleman (1967) passages were analyzed according to number of unfamiliar words at each of the grade levels. Those values were then regressed against Miller and Coleman passage total cloze scores to determine the amount of variance in total cloze scores accounted for by the number of unfamiliar words at each grade level.

Regression techniques were also used to determine the syntactic variables to be retained for the second phase of the investigation. The syntactic variables analyzed for the Miller and Coleman (1967) passages included: 1) number of T-units; 2) T-unit length (i.e., average number of words per T-unit); 3) number of clauses; 4) clause length (i.e., average number of words per clause); 5) number of sentences; 6) sentence length (i.e., average number of words per sentence); and 7) voice (i.e., percent of passive sentences

and percent of passive verb phrases). These syntactic-variable values were then regressed against the Miller and Coleman total cloze scores. The variables that accounted for significant variance in total cloze scores were retained for use in the second phase of the investigation.

The objectives of the second phase of the study were to calibrate the new-model formulas and recalibrate the existing Dale-Chall (1995), FOG, and Homan-Hewitt formulas with the Miller and Coleman (1967) passages. For the calibration of the new-model formulas, different combinations of semantic and syntactic variables retained from the first phase of the study were explored. Stepwise multiple regression techniques were used with Miller and Coleman passage semantic and syntactic values as the independent variable and the respective total cloze scores as the dependent variable. Several formulas were created and four were retained for further investigation and use in the third phase of the study.

Simple-linear, stepwise-multiple, and hierarchical-multiple regression techniques were used to recalibrate the existing formulas. The Miller and Coleman passages were analyzed according to the predictor variables for each existing formula. The predictor variables for each formula served as the independent variables and the total cloze scores served as the dependent variables. This resulted in recalibrated formulas for each of the four existing readability formulas, which were used in the third phase of the investigation.

The objective of the third phase of the study was to collect external validity and reliability evidence to support the use of the new model with credentialing materials. Materials related to a dentistry licensing program were used. Specifically, samples were

collected for analysis from examination materials. The materials included actual test items (N = 48) and related item difficulty data from an administration of a dental licensing test. Methods were devised and used to convert the examination items into pseudo-continuous prose prior to analysis.

The new-model and recalibrated formulas were used to assess the estimated readability of the examination materials. Correlational, non-parametric-rank comparisons, and regression analysis methods were used to compare the estimated readability values across formulas. The correlational analyses were used to determine how well the results of the new-model and recalibrated formulas corresponded. Freidman's two-way analysis of ranks and Sign tests were used to compare the formula results. The materials were subjected to regression analyses to determine whether differences among the new-model and recalibrated formula results were systematic and potentially due to the existence of, and the recalibrated formulas not accounting for, occupational-specific vocabulary.

According to the results of the analyses conducted in Phase III, one new-model formula was identified as the most stable of the four new-model formulas. This formula was selected for retention and included in post-hoc analyses. Specifically, the occupational specific vocabulary list was used with the recalibrated formulas and additional Sign tests were conducted and the order in which the recalibrated and new-model formula fell were compared.

CHAPTER 3

METHODS

Variables in the Model

To determine variables, procedures, and supplementary instruments to be included in the model, those used in the most popular and well-validated readability formulas were considered. Previous research has unanimously revealed that semantic and syntactic characteristics of texts are reliable and valid indicators of readability. All existing readability formulas include semantic variable(s) and virtually all formulas include syntactic variable(s). Therefore, this new model included semantic and syntactic measures.

To identify which semantic and syntactic variables to address and determine the most appropriate measures of those variables, the work of the most popular, well-established, and well-validated readability formulas (e.g., Bormuth, 1969; Chall & Dale, 1995; Dale & Chall, 1948; Flesch, 1948) were incorporated. The work of Hewitt and Homan (2004) and Homan et al. (1994) is particularly relevant as the researchers were able to establish a formula suitable for the multiple choice format. Although there have not been extensive validation studies for this formula, initial investigations have shown the formula to be reliable (Hewitt & Homan, 2004; Homan et al., 1994).

Semantic Characteristics

The readability formulas created and validated by Dale and Chall (1948); Chall and Dale (1995); Bormuth (1969), and Homan and Hewitt (1983, 1989) include measures of vocabulary load that involve the use of lists of familiar words (e.g., The Dale-Chall list of 3,000 familiar words [Dale & Chall, 1943] and *The Living Word Vocabulary: A National*

Vocabulary Inventory [Dale & O'Rourke, 1981]). Although word lists have been useful in identifying vocabulary load in the estimation of readability levels, the exclusive use of existing word lists is unsuitable for the purpose of the proposed model. Occupational-specific terminology, which is likely to affect readability estimates, is not included in existing lists of familiar words. Therefore, the proposed model involved the use of two words lists to estimate syntactic complexity, or vocabulary load: 1) *The Living Word Vocabulary: A National Vocabulary Inventory (*Dale & O'Rourke, 1981) and 2) an occupational-specific word list. *The Living Word Vocabulary* list offered a general measure of vocabulary load or semantic complexity. The occupational-specific word list allowed common job-related terms, that would otherwise be deemed high-level and difficult, to be considered familiar and treated in the same way as words included in *The Living Word Vocabulary* list.

The first word list, *The Living Word Vocabulary: A National Vocabulary Inventory (*Dale & O'Rourke, 1981) was used in the Dale-Chall readability formula (1995) and the Homan-Hewitt readability formula (1983, 1989; 2004; Homan et al. 1994). The corpus of 44,000 words offers grade-level familiarity scores for multiple meanings of each included word. Familiarity scores are offered for students in grade school through college (i.e., grades 4-16). For each grade level, the authors offer the word definitions with which students at that grade level should be most familiar as well as the percentage of students at that grade level who should be familiar with the meaning (DuBay, 2004).

Second, an occupational-specific word list was created and included more than 4,900 terms assumed to be familiar to students of dentistry (see Appendix 1). The dentistry occupational-specific word list was created by referencing 26 dental textbooks. Once an

exhaustive list of sources and words that appeared to be common to the dentistry field was created, it was submitted to a subject-matter expert who is a practicing dentist and teaches dentistry courses at a University. The subject-matter expert provided feedback on the word list and the sources from which the words were drawn. Amendments to the word list and inventory of sources from which the words were derived were made according to the subject-matter expert's input.

Syntactic Characteristics

Determining which syntactic characteristics to measure was more complex and required careful consideration of numerous variables that may or may not have been useful in the estimation of readability for the present text types. The following variables for the measurement of syntactic complexity were considered: 1) number of T-units; 2) T-unit length (i.e., average number of words per T-unit); 3) number of clauses; 4) clause length (i.e., average number of words per clause); 5) number of sentences; 6) sentence length (i.e., average number of words per sentence); and 7) voice (i.e., percent of passive sentences and percent of passive verb phrases).

T-unit and clause length were considered because these variables have shown to be appropriate indices of syntactic complexity and mature writing (Hunt, 1965, 1970a, 1970b). This approach is similar to that used by Gunning with his FOG index (1952), where each complete thought in a sentence was treated as a separate sentence. Hunt (1965) found that for school aged children, T-units were the best indicator of syntactic complexity. When he included the writings of superior adults in his analysis, he determined that clause length was an equally good index of syntactic complexity. In his follow-up study, which included school-aged children, average adult writers, and superior

adult writers, Hunt (1970a, 1970b) substantiated his 1965 findings. In addition, Homan and Hewitt (1983, 1989, 2004) used T-unit length as a measure of syntactic complexity in the readability formula they devised for multiple-choice examination items.

T-units and clauses are typically shorter than sentences, yet they possess, at minimum, a subject and a verb. Definitions of a clause differ among scholars. In this investigation, a clause was defined as Hunt (1965) defined it, "a structure containing a subject (or coordinating subjects) and a finite verb phrase (or coordinating verb phrases)" (p. 40). T-units are larger than a single clause, but smaller than sentences. Hunt introduced the T-unit in 1965 and defined it as, "a grammatically discrete unit intervening in size between the clause and what is punctuated as a sentence" and further described a T-unit as "one main clause plus the subordinate clauses attached to or embedded within it" (p. 49). Because T-units and clauses are shorter than sentences, it was possible to more precisely divide a small text than would be possible with the use of sentences. This offered more data points for investigation.

An example of how texts are divided into T-units and clauses according to Hunt's guidelines is provided. Below is a single sentence written by a 4th-grade student who participated in Hunt's (1965) study. Following the sentence is the division of the sentence into T-units and clauses (Hunt, 1965). Each T-unit is numbered, begins with a capital letter, and ends with a period. Clauses are delineated with backslashes.

> I like the movie we saw about Moby Dick the white whale the captain said if you can kill the white whale Moby Dick I will give this gold to the one who can do it and it is worth sixteen dollars they tried and tried but while they were trying they killed a whale and used the oil for the lamps they almost caught the white whale

1. I like the movie / we saw about Moby Dick, the white whale.
2. The captain said / if you can kill the white whale, Moby Dick, / I will give this gold to the one / that can do it.
3. And it is worth sixteen dollars.
4. They tried and tried.
5. But / while they were trying / they killed a whale and used the oil for the lamps.
6. They almost caught the white whale. (p. 20)

This passage includes eleven clauses, six T-units, and one sentence.

Sentence length has been successfully used as a syntactic measure in the majority of existing readability formulas (e.g., Bormuth, 1969; Chall & Dale, 1995; Dale &Chall, 1948; Flesch, 1948; Gunning, 1952; Spache, 1953). Existing readability formulas typically require several samples of 100 or more words to reliably estimate readability of a text. Although it would be possible to obtain samples of this size for learning and occupational materials, multiple-choice examination items tend to be shorter. Therefore, sentence length was not expected to be appropriate for multiple-choice examination items, but deserved consideration.

Because the number of T-units per passage was explored as an independent variable, the number of sentences per passage was also explored. Like sentence length, number of sentences was not expected to be appropriate for multiple-choice examination items, but was worthy of consideration.

Although passive versus active voice has received limited attention by readability researchers, it deserved consideration. The voice of verb phrases has shown to affect

comprehension, especially for English language learners (Abedi, 2006, 1995; Abedi & Lord, 2001). Therefore, the percentage of passive sentences, as well as the percentage of passive verb phrases per passage, was investigated to determine if voice accounts for significant variance in passage difficulty.

Formula Calibration

This section includes a discussion of the new-model readability formula calibration and the existing readability formula recalibration. The materials and data that were used for formula calibrations and recalibrations are discussed in the first section. The methods that were used to investigate the usefulness of the variables under consideration and identify variables worthy of retention and further investigation are explained in the second section. The methods that were used to determine appropriate weightings of each retained variable to develop the new-model readability formula and how the existing readability formulas were recalibrated are discussed in the third section.

Materials

Miller and Coleman's (1967) 36, approximately 150-word passages were used to calibrate the formulas. These passages range in difficulty from 1st-grade to technical material. Miller and Coleman constructed and administered three types of cloze tests for the 36, 150-word passages to 479 college students. Coleman and Miller (1968) used data from 20 undergraduate students to establish Information Gain (IG) scores for each of the 36 passages. IG refers to "the efficiency with which a passage transmits new information" (Coleman & Miller, p. 371).

Aquino (1969) established significant relationships between Miller and Coleman's findings (CT I and CT III) and word-for-word recall as well as judgments of difficulty.

For word-for-word recall, Aquino had 14 participants, who were employed in an educational research laboratory, read each passage and attempt to recall the passage word-for-word. For judgments of difficulty, the author had the same subjects arrange the passages according to difficulty. Aquino found that his measures were significantly correlated with CT I and CT III scores established by Miller and Coleman and resulted in similar rank orderings.

Miller and Coleman (1967) did not include their passages or report their mean cloze percentage scores for the passages and tests in their research report. Aquino (1969), on the other hand, offered these passages in his study designed to determine the validity of Miller and Coleman's scale. In addition, Aquino included Miller and Coleman's mean cloze percentage scores for each test, total value scores for the three tests (CT I, II, and III) combined, and Coleman and Miller's (1968) IG scores for each passage. It was not possible to locate any other published version of Miller and Coleman's passages. Therefore, the passages and related scores were accessed from Aquino's work.

Procedures

This section includes a discussion of the methods that were used to investigate the usefulness of the variables considered, identify variables for further retention and analysis, and create and calibrate the new-model formulas.

The 36 passages calibrated for complexity by Miller and Coleman (1967) were analyzed according to the chosen syntactic and semantic variables. Specifically, the syntactic analysis for each passage included determining: 1) number of T-units; 2) T-unit length (i.e., average number of words per T-unit); 3) number of clauses; 4) clause length (i.e., average number of words per clause); 5) number of sentences; 6) sentence length

(i.e., average number of words per sentence); 7) percentage of passive sentences, and 8) percentage of passive verb phrases. To analyze the passages for semantic complexity, the number of words not included in *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) were determined for grade levels 4, 6, 8, 10, 12, 13, and 16.

Because not all of the Miller and Coleman (1967) passages included exactly 150 words and ranged from 149 to 152, variable measures were adjusted for exactly 150 words. For example, Miller and Coleman passage 9 included 151 words and 8 sentences. The number of sentences was adjusted by dividing the actual number of sentences by the total number of words and multiplying that product by 150 [i.e., (8/150)*150 = 7.947].

Phase I: Usefulness of Variables

The usefulness of occupational-specific vocabulary list was not investigated with the Miller and Coleman (1967) passages. Although it would have been possible to identify words that appeared to be technical terms related to the respective fields in the two most difficult passages, and thereby create an occupational-specific word list, it would not have been appropriate to treat the terms as familiar. The two most difficult passages included technical language. The second most difficult passage concerned how the investigation of scientific theory contributes to the establishment of empirical law in the psychological sciences. The passage includes terms that would likely be included on a list of familiar words for social scientists (e.g., empirical, variables, phenomena, psychological). The most difficult passage was a discussion regarding nerve division experiments. The passage included terms that would likely be included on a list of familiar words for medical sciences (e.g., volar, anesthetic, cutaneous, algesiometer). The

words that appeared to be technical terminology could have been treated as though they were part of an occupational-specific vocabulary list, but the cloze scores were based on responses from an audience for whom these terms should not be familiar. Miller and Coleman used the responses of undergraduate college students to scale the passages. Therefore, in this phase of the investigation, the usefulness of the semantic variable did not include an occupational-specific vocabulary list.

Simple linear regression analysis was used to investigate the variance in cloze scores accounted for by the semantic and syntactic variables under examination. Simple linear regression analysis was conducted to determine the usefulness of the variables. Miller and Coleman's (1967) total cloze scores (i.e., the sum of CT I, CT II, and CT III scores for each passage; Aquino, 1969) was the dependent variable and 1) number of familiar words, 2) number of T-units; 3) T-unit length (i.e., average number of words per T-unit); 4) number of clauses; 5) clause length (i.e., average number of words per clause); 6) number of sentences; 7) sentence length (i.e., average number of words per sentence); 8) percentage of passive sentences, and 9) percentage of passive verb phrases were the independent variables. The regression analyses allowed the identification of predictor variables that accounted for a statistically significant amount of variance in cloze scores while controlling for the effects of the other variables. Data and standardized residuals for predictor variables were plotted to facilitate the identification of potential curvilinearity. The results from the simple linear regression analyses were used to identify the variables to be retained in the next phase of the investigation.

Phase II: Formula Creation and Calibration

An exploratory regression approach was also used to create and calibrate four new-model formulas. Stepwise multiple regression was used to refine the variable selection and determine appropriate weightings. Several syntactic variables accounted for statistically significant amounts of variance in cloze scores during the first phase and were, therefore, retained for the second phase of the investigation. The usefulness of each variable, and how much variance they accounted for when they were combined with semantic variables, was explored. Details of these variable combinations are included in the results section of this study.

Dale-Chall (1995), FOG, and Homan-Hewitt readability formulas were recalibrated. Specifically, multiple regression techniques were used with Miller and Coleman (1967) passage total cloze test (CT) scores as the dependent variable and existing formula components as the dependent variables. The recalibration of these formulas with the retention of their established variables provided a consistent comparison of the existing formula and new model results during the third phase of this investigation.

Phase III: External Validity and Reliability Evidence

This section includes an explanation of the methods that were used to collect and analyze external validity and reliability evidence for the new model. The discussion begins with a description of the materials that were used. Data collection procedures are then outlined. The comparisons that were made and expected consistencies and differences are described next. Finally, the statistical methods that were used to analyze the results are outlined.

Materials

This phase of the investigation involved the use of examination materials related to a dentistry licensure program. The first subsection includes a brief discussion of the licensure program and the stakes involved for the candidates, program owners, and general population. The subsequent subsection includes an explanation of the relevant examination materials that were used in this portion of the investigation.

Dental licensure program.

The dentistry professional licensure program is owned by a board of dentistry and is mandatory for the practice of dentistry in a specific region of North America. Candidates must pass a two-part (knowledge and practical) multiple-choice examination to be licensed to practice. The learning materials related to this examination consist of a variety of textbooks and professional journal articles that students are required to read during schooling. The examination is comprised of 300 knowledge-based questions and 92 practical questions. The occupational materials consist of textbooks, continuing education materials, professional journal articles, dental association monthly news packets, and instructional manuals for products and equipment.

The dentistry professional licensure examination is high stakes for the candidates. They have had several years of schooling with the goal of becoming a dentist. Passing the examination is a principal requirement to be licensed to practice dentistry in this geographic region. Candidates pay approximately $1,500 in registration and examination fees every time they take the exam. They are eligible to take the test three times within 60 months of graduation. If they fail it all three times, they have to take and pass a qualifying course to be eligible to retake the exam.

The stakes are also high for the dentistry professional licensure board. They have several years and millions of dollars invested in their program and are a trusted authority and governing body charged with identifying dentistry students who are ready to enter the field. They must have enough confidence in the validity and reliability of the examination results to assert that candidates who pass it have the prerequisite knowledge and skills necessary to enter the field and not do harm to the public. Unqualified candidates passing the examination could damage the credibility of the board and its individual members. In addition, candidates who believe they have been unjustly failed can contest the examination results and even pursue law suits against the licensing body. Therefore, if the licensing board cannot offer sound validity evidence for the examination results, they may be subject to legal costs.

For the general public, or dental patients, the stakes of the examination are high. Incompetent people working in most health-related fields can pose significant risks to public safety. Candidates who pass the dental licensing examination are certified to practice endodonics (e.g., root canals), basic oral surgery (e.g., tooth extraction), periodontal surgery (e.g., root planning), placement of fixed prosthetics (e.g., crowns), operatives (e.g., amalgam and composite fillings of lesions), and administer anesthetics. In addition, practicing dentists must be aware of life-threatening issues such as drug interactions (C.W. Buckendahl, personal communication, July 30, 2008).

Procedure

The readability of examination items for the licensure program was investigated. The new-model, recalibrated Dale-Chall (1995), recalibrated FOG, and recalibrated Homan-Hewitt formulas were applied to the examination materials. The following subsections

include an explanation of the procedures that were used to estimate the readability of these materials.

Examination items.

Test items and related data (i.e., item reliability, discrimination, and difficulty values) for 100 candidates were provided by the dentistry professional licensure program. Stratified and systematic sampling procedures were used to select examination items for inclusion in the investigation. Forty-eight examination items were selected from the two 150-item components (i.e., Book 1 and Book 2) of the knowledge-based portion of the dentistry examination: 24 examination items from Book 1 and 24 items from Book 2. The difficulty values, calculated according to the percentage of candidates who correctly answered an item, were considered in the selection of items. Equal numbers of high, medium, and easy items were selected. Details of the sampling procedures are provided in the results section of this study. The new-model formulas, as well as the recalibrated Dale-Chall (1995), FOG, and Homan-Hewitt formulas, were then applied to the 48 selected items.

Estimating the readability of the multiple-choice examination items required the creation of a method for converting the items into pseudo-continuous prose. Therefore, the 48 multiple-choice examination items were converted into pseudo-continuous prose with a method similar to that used by Plake (1988). Familiarization with terminology related to the components of multiple-choice items is essential to understanding the procedures that were used. Therefore, three key terms are defined prior to the explanation of the procedures that were used: 1) scenarios, 2) stems; and 3) options. *Scenarios* include background information or hypothetical situations presented to the candidates to

consider when they answer the question. *Stems* are the actual questions posed to the candidate. *Options* include the keyed response(s) and distractors from which the candidate has to choose. Below are the guidelines that were followed to create pseudo-continuous prose from each examination item:

1) If the stem was an incomplete sentence and each of the options completed the sentence, the stem and each option were combined to create individual sentences.
2) If the stem was a complete sentence and the options were not complete sentences, the stem and options were combined to create individual sentences.
3) If the stem and each option were complete sentences, each was considered an individual sentence.
4) If an item included a scenario, the scenario was not combined with the stem or options. The scenario stood alone and each sentence in a scenario was counted once and measured along with the other components of the item.
5) If an item included instructions, such as those indicating that a reference image should be considered, the instructions were counted in the same way as scenarios. If a set of instructions applied to a group of items, the instructions were added to each question and added to their pseudo-continuous prose.
6) Where the stem included options and the options actually referred back to the choices in the stem, the elements were combined to create as many complete sentences as possible.

If each multiple-choice item included a minimum of four options, the methods devised for converting the items into pseudo-continuous prose yielded texts of at least four sentences each. Below are examples of how the guidelines were used. For each

guideline, a multiple-choice item obtained from websites related to certification and licensure and the pseudo-continuous prose that would be extracted for the respective items are offered.

Guideline 1:

The most important organelle or component of a cell for oxidative processes is the

A. nucleus.

B. nucleolus.

C. mitochondrion.

D. Golgi complex.

E. endoplasmic reticulum.

Retrieved from http://www.ada.org/prof/ed/testing/nbde01/nbde01_candidate_guide_2008.pdf

The pseudo-continuous prose for the above item would consist of the following sentences:

1) The most important organelle or component of a cell for oxidative processes is the nucleus.
2) The most important organelle or component of a cell for oxidative processes is the nucleolus.
3) The most important organelle or component of a cell for oxidative processes is the mitochondrion.
4) The most important organelle or component of a cell for oxidative processes is the Golgi complex.

5) The most important organelle or component of a cell for oxidative processes is the endoplasmic reticulum.

Guideline 2:

Which of the following enzymes catalyzes the formation of uric acid from purines?

A. Urease

B. Uricase

C. Xanthine oxidase

D. Aspartate transcarbamoylase

E. Carbamoyl-phosphate synthetase

Retrieved from http://www.ada.org/prof/ed/testing/nbde01/nbde01_candidate_guide_2008.pdf

The pseudo-continuous prose for the above item would consist of the following sentences:

1) Which of the following enzymes catalyzes the formation of uric acid from purines- Urease?
2) Which of the following enzymes catalyzes the formation of uric acid from purines- Uricase?
3) Which of the following enzymes catalyzes the formation of uric acid from purines- Xanthine oxidase?
4) Which of the following enzymes catalyzes the formation of uric acid from purines- Aspartate transcarbamoylase?

5) Which of the following enzymes catalyzes the formation of uric acid from purines- Carbamoyl-phosphate synthetase?

Guideline 3:

If electrolyte from a lead-acid battery is spilled in the battery compartment, which procedure should be followed?

A. Apply boric acid solution to the affected area followed by a water rinse.

B. Rinse the affected area thoroughly with clean water.

C. Apply sodium bicarbonate solution to the affected area followed by a water rinse.

D. Rinse the affected area thoroughly with clean water followed by a sodium bicarbonate rinse.

Retrieved from http://www.faa.gov/education_research/testing/airmen/test_questions/media/amg.pdf

The pseudo-continuous prose for the above item would consist of the following sentences:

1) If electrolyte from a lead-acid battery is spilled in the battery compartment, which procedure should be followed?
2) Apply boric acid solution to the affected area followed by a water rinse.
3) Rinse the affected area thoroughly with clean water.
4) Apply sodium bicarbonate solution to the affected area followed by a water rinse.
5) Rinse the affected area thoroughly with clean water followed by a sodium bicarbonate rinse.

Guideline 4:

You would like to protect your corporate intranet from hacker attacks through the Internet.

Which two methods would help to accomplish this? (Choose two.)

A. Install a second network adapter.

B. Remove TCP/IP as the protocol used on IIS.

C. Restrict access through the use of permissions.

D. Implement IPX as the protocol between IIS and the intranet.

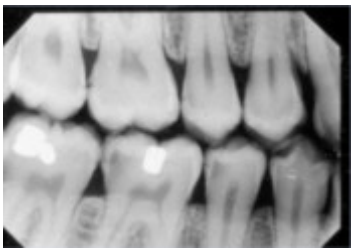
Retrieved from http://mcpmag.com/Features/article.asp?EditorialsID=103

The pseudo-continuous prose for the above item would consist of the following sentences:

1) You would like to protect your corporate intranet from hacker attacks through the Internet.
2) Which two methods would help to accomplish this?
3) Choose two.
4) Install a second network adapter.
5) Remove TCP/IP as the protocol used on IIS.
6) Restrict access through the use of permissions.
7) Implement IPX as the protocol between IIS and the intranet.

Guideline 5:

Using the print of the radiograph labeled Sample1, answer the following question on the answer score sheet.



(Select ONE OR MORE correct answers.)

There is radiographic evidence of caries on the

A. distal of tooth 4.3.

B. mesial of tooth 4.4.

C. distal of tooth 4.4.

D. mesial of tooth 4.5.

E. distal of tooth 4.5.

F. mesial of tooth 4.6.

G. distal of tooth 4.6.

H. mesial of tooth 4.7.

I. distal of tooth 4.7.

J. mesial of tooth 4.8.

K. distal of tooth 4.8.

Retrieved from http://www.ndeb.ca/en/accredited/osce_examination.htm

The pseudo-continuous prose for the above item would consist of the following sentences:

1) Using the print of the radiograph labeled Sample1, answer the following question on the answer score sheet.
2) Select ONE OR MORE correct answers.
3) There is radiographic evidence of caries on the distal of tooth 4.3.

4) There is radiographic evidence of caries on the mesial of tooth 4.4.

5) There is radiographic evidence of caries on the distal of tooth 4.4.

6) There is radiographic evidence of caries on the mesial of tooth 4.5.

7) There is radiographic evidence of caries on the distal of tooth 4.5.

8) There is radiographic evidence of caries on the mesial of tooth 4.6.

9) There is radiographic evidence of caries on the distal of tooth 4.6.

10) There is radiographic evidence of caries on the mesial of tooth 4.7.

11) There is radiographic evidence of caries on the distal of tooth 4.7.

12) There is radiographic evidence of caries on the mesial of tooth 4.8.

13) There is radiographic evidence of caries on the distal of tooth 4.8.

Guideline 6:

The washing of hands must be performed before putting on and after removing gloves because it

1. reduces the number of skin bacteria which multiply and cause irritation.

2. completely eliminates skin bacteria.

3. minimizes the transient bacteria which could contaminate hands through small pinholes.

4. allows gloves to slide on easier when the hands are moist.

A. (1) (2) (3)

B. (1) and (3)

C. (2) and (4)

D. (4) only

E. All of the above.

Retrieved from

http://www.ndeb.ca/en/accredited/documents/2006ReleasedEnglishBookII.pdf

The pseudo-continuous prose for the above item would consist of the following sentences:

1) The washing of hands must be performed before putting on and after removing gloves because it reduces the number of skin bacteria which multiply and cause irritation, completely eliminates skin bacteria, minimizes the transient bacteria which could contaminate hands through small pinholes.
2) The washing of hands must be performed before putting on and after removing gloves because it reduces the number of skin bacteria which multiply and cause irritation and minimizes the transient bacteria which could contaminate hands through small pinholes.
3) The washing of hands must be performed before putting on and after removing gloves because completely eliminates skin bacteria and allows gloves to slide on easier when the hands are moist.
4) The washing of hands must be performed before putting on and after removing gloves because it allows gloves to slide on easier when the hands are moist.
5) All of the above [are correct].

Comparisons and expectations.

Obtaining readability estimates for the materials according to the new-models, recalibrated Dale-Chall (1995), recalibrated FOG, and recalibrated Homan-Hewitt readability formulas enabled result comparisons for the recalibrated, existing readability formulas and new-model formulas. Correlational analyses were conducted to examine the

relationships between the readability estimates for the dental materials derived with each formula. Non-parametric analyses were used to compare the readability estimates across formulas. Regression techniques were used to determine whether differences among the results of the new-model and recalibrated formula readability estimates were related to the unfamiliar and multisyllabic occupational-specific terms in the passages.

Systematic differences in the rankings determined according to the recalibrated readability formulas and the new model were expected. More specifically, it was expected that the formulas that incorporate lists of familiar words (i.e., Dale-Chall, Homan-Hewitt) for measures of semantic complexity would yield readability estimations indicating more difficult-to-read text than the new model because job-related terminology would be counted as unfamiliar in the existing formulas and would be considered familiar with the new model. *The Living Word Vocabulary: A National Vocabulary Inventory (*Dale & O'Rourke, 1981), which is used with the Dale-Chall and Homan-Hewitt formulas, does not include occupational-specific vocabulary terms. Occupational terminology that would be appropriately deemed familiar to the respective populations of interest would be treated as unfamiliar, or difficult, in the Dale-Chall and Homan-Hewitt formulas. Therefore, it was expected that divergence of the results of the new-model and recalibrated Dale-Chall and Homan-Hewitt formulas would be related to occurrences of occupational-specific terminology in the materials.

Systematic differences between the results of the new-model formulas and those of the recalibrated FOG formula were also anticipated. The FOG formula involves the use of syllable counts as a measure of semantic complexity. Specifically, it requires counting the number multisyllabic words in a sample. The dentistry occupational-specific terms

tend to be comprised of many multisyllabic words; but those words should be considered familiar to the audience. Therefore, it was expected that divergence of the results of the new-model and recalibrated FOG formulas would be related to occurrences of occupational-specific terminology in the materials.

Analysis of external validity and reliability data.

Parametric and non-parametric statistical methods were used to analyze the readability data. Correlational analyses were conducted for each set of materials to determine the relationships between the results derived with each new-model and recalibrated formula. To determine whether the new model resulted in passage rankings that were significantly different from the passage rankings of the other formulas, Friedman two-way analysis of ranks and Sign tests were conducted with readability formula as the independent variable and readability estimates as the dependent variables.

The readability estimates derived with the recalibrated existing formulas were further examined according to the percentage or number of occupational-specific vocabulary terms in the passages that were identified as unfamiliar, long (more than six letters), or multisyllabic (more than three syllables). Simple linear and stepwise multiple regression techniques were used to determine whether relationships existed between the readability estimates determined according to the recalibrated Dale-Chall formula, which required the use of a list of familiar words, and the number of occupational-specific vocabulary terms that appeared in the passages and had been identified as unfamiliar. The same methods were used to investigate whether relationships existed between the readability estimates determined according to the recalibrated Homan-Hewitt formula, which required the identification of long (more than 6 letter) words and the use of a list of

familiar words and the number of occupational-specific vocabulary terms that appeared in the passages and had been identified as long or unfamiliar. Regression techniques were also used to investigate relationships between the readability estimates determined according to the recalibrated FOG formulas, which required syllable counts, and the number of occupational-specific vocabulary terms that appear in the passages and had been identified as multisyllabic.

The results of the planned statistical analyses revealed the need for subsequent, post-hoc analyses of the data. Additional correlational analyses and Sign tests were conducted. Post-hoc analysis results facilitated the interpretations of the planned analysis results and are described in the results section.

CHAPTER 4

RESULTS

This section is comprised of four major components. Figure 1 provides a graphic representation of the general layout of this section. The first three components correspond directly to the three phases in the investigation as outlined in the methods section and the fourth component includes a summary of the findings from Phase III and additional post-hoc analyses of the data. The first component, Phase I: Usefulness of variables, includes the results of exploratory regression analyses that were conducted to determine the amount of variance in (Miller & Coleman, 1967) total cloze test (CT) scores accounted for by the syntactic and semantic variables under investigation for the calibration passages. These analyses were conducted to determine which syntactic and semantic variables should be retained for further consideration in the second phase of the investigation. All eight syntactic variables accounted for a significant amount of variance in total cloze scores; however, only seven were retained for further investigation. The semantic variable, number of unfamiliar words according to *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) accounted for a significant amount of variance in total CT scores at five of the six grade levels and these levels were retained for further investigation.

The second component, Phase II: Formula creation and calibration, includes the results of regression analyses that were conducted to explore the variance in total cloze scores accounted for by all possible combinations of the retained syntactic and semantic variables. These analyses were conducted to create and calibrate the new-model formula. The results showed that four new-model formulas were worthy of retention and further

investigation. This component also includes the results of exploratory regression analyses that were conducted to recalibrate three existing readability formulas: the same total CT scores served as the dependent variable and each of the components from existing formulas served as the independent variables. These regression analyses resulted in five recalibrated formulas, because three recalibrated formulas were created for one of the existing formulas due to difficulties encountered during the recalibration process.

During Phase III: External validity and reliability evidence, the four new-model and five recalibrated formulas were applied to the examination materials. This resulted in readability level estimates derived with each formula for each individual passage as well as overall readability-level averages for the materials. The results section for this phase of the investigation includes the results of correlations, Friedman two-way analysis of ranks tests, Sign tests, and regression analyses that were conducted to investigate how the four new-model and five recalibrated formulas performed when applied to credentialing-examination materials.

The Phase III component is divided into subsections according to analyses that were conducted. The first subsection Step I: Relationships between formula results, includes the results of correlational analyses that were conducted to determine the relationships between the formulas. These analyses were conducted to determine how well the results of the new-model formulas correlated with the results of the recalibrated formulas and to explore how well the results of the recalibrated formulas correlated with one another. These initial correlation analyses revealed that one of the new-model formulas (TUL8) significantly correlated with the results of one recalibrated formula (FOG3). No other relationships between the results of new-model and recalibrated formulas reached

significance. When the occupational-specific vocabulary list was used with the recalibrated formulas, the results of all of the new-model formulas and recalibrated formulas were significantly correlated.

Post-hoc correlational analyses reported in the Phase III component were conducted to address the weak and non-significant correlations initially observed between the new-model and recalibrated formula results, which were assumed to be due to the inclusion of occupational-specific vocabulary as contributors to increases in semantic complexity with the recalibrated formulas. The recalibrated formulas were once again applied to the materials, but modifications were made to account for the occupational-specific vocabulary in the passages. Specifically, during the calculation of the semantic variable for each recalibrated formula, occupational-specific vocabulary terms were removed from the totals. By this, the occupational-specific vocabulary terms were treated in a manner consistent with the way they were treated in the new-models. It was expected that the correlations between the new-model and recalibrated formula results would be stronger when the occupational-specific vocabulary was treated the same way across all formulas. This expectation was met: the correlations between the four new-models and all recalibrated formulas increased to significance.

Step II: Differences between formula results includes the results of comparisons made between formula results. Friedman two-way analysis of ranks tests and Sign tests were employed. The results of these analyses within material sets (i.e., combined Books 1 and 2, Book 1, and Book 2) were not considered as support of, or evidence against, the utility of the new models. Instead, the comparisons were used to explore how the results of all formulas corresponded. Friedman two-way analysis of ranks test and Sign test results

revealed significant differences between the results of all but two new-model formulas. In addition, the Sign tests conducted to compare the results of the new-model and recalibrated formulas revealed significant differences between 15 of the 20 possible formula pairs for combined Books 1 and 2, 13 of the 20 possible formula pairs for Book 1, and 12 of the 20 possible formula pairs for Book 2.

The occupational-specific vocabulary list was then used with the recalibrated formulas and post-hoc Sign tests were conducted to compare the results to the results of a new-model formula. Specifically, the new-model TUL8 results were compared to the results of the recalibrated formulas that were derived with the use of the occupational-specific vocabulary list. These results were then inspected and compared to the results observed when the occupational-specific vocabulary list was not used with the recalibrated formulas. If fewer significant differences were observed between the results of the new-model and recalibrated formulas once the occupational-specific vocabulary list was used with the recalibrated formulas, there would be evidence to suggest that the differential treatment of occupational-specific vocabulary was largely a source of the previously observed significant differences.

Step III: Determining whether differences were systematic includes the results of regression analyses conducted to determine how much variance in the readability estimates derived with the recalibrated formulas was due to the existence and frequency of occupational-specific vocabulary in the passages. These analyses were conducted to determine whether the differences between readability estimates derived with the new-models and recalibrated formulas could be attributed to occupational-specific vocabulary in the passages. With the procedures used in the recalibrated formulas, these vocabulary

terms were identified as contributors to increases in semantic complexity. Conversely, with the procedures used in the new-models, these vocabulary terms were not considered to contribute to an increase in semantic complexity because these terms should be familiar to the respective reading audience. To determine how much variance in the readability estimates derived with the use of the recalibrated formulas was due to occupational-specific vocabulary being identified as contributors to semantic complexity, the readability estimate for each recalibrated formula served as the dependent variable. The number or percentage of words that were identified as contributors to semantic complexity and were included in the occupational-specific vocabulary list served as the independent variables. The occupational-specific vocabulary words that were identified as contributors to semantic complexity accounted for a significant amount of variance in readability estimates derived with the recalibrated formulas.

The fourth major component, Results of external validity and reliability analyses, includes a comprehensive summary of the results of Phase III of the investigation as well as additional post-hoc analyses results. The post-hoc analysis includes an examination of how the overall readability estimates ranked for each formula. The order in which the formula results fell was then compared across the two books of examination items. The results revealed that the order in which the formulas fell were perfectly consistent for Books 1 and 2 of the examination materials when the occupational-specific vocabulary list was not used with the recalibrated formulas. However, when the occupational-specific vocabulary list was used with the recalibrated formulas, the order in which the recalibrated formulas fell differed, although the mean values were not significantly different from one another.

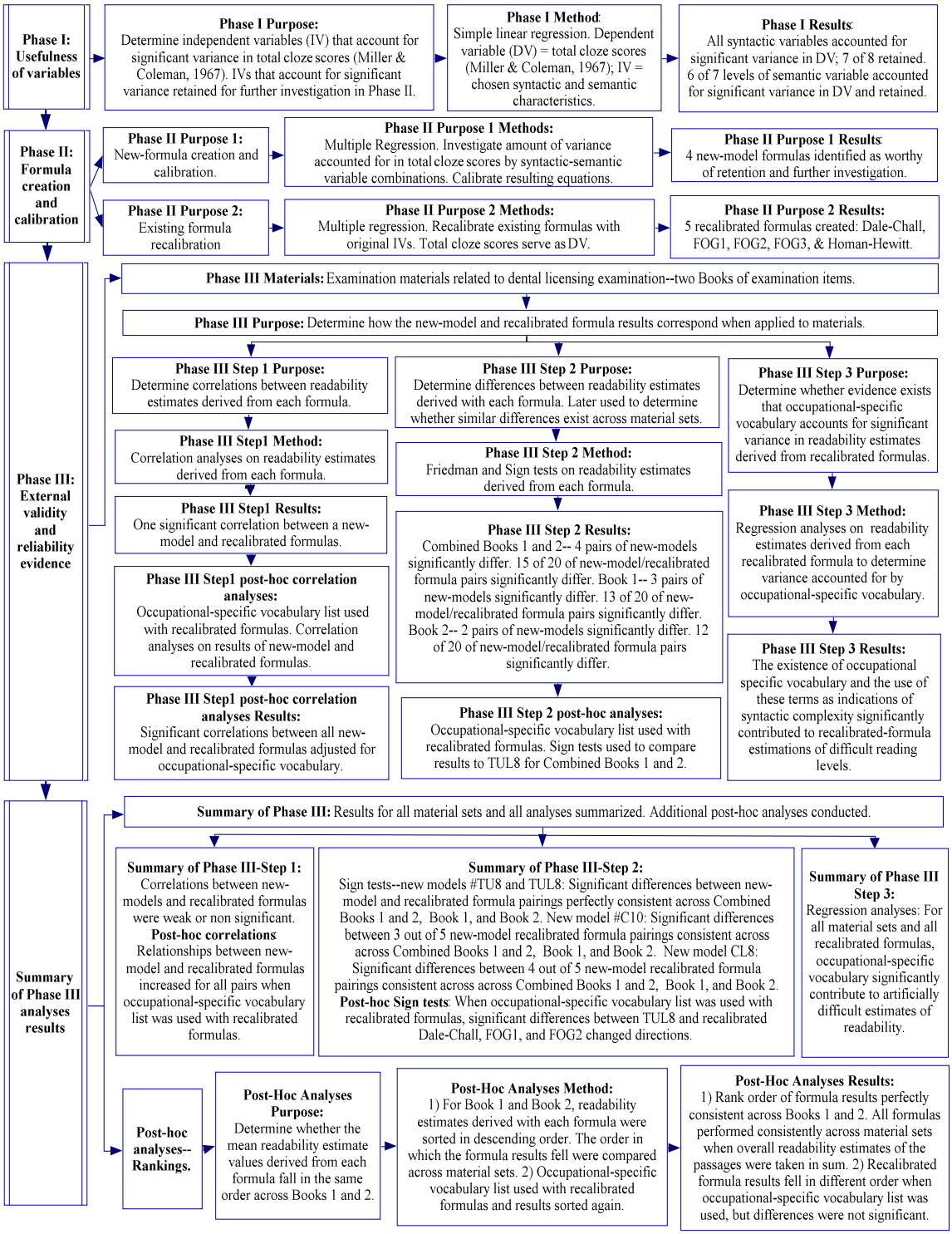


Figure 1. Graphic representation of the organization of the results section.

Phase I: Usefulness of Variables

The 36 passages calibrated for complexity by Miller and Coleman (1967) were analyzed according to the chosen syntactic and semantic variables. Specifically, the syntactic analysis for each passage included determining 1) number of T-units; 2) T-unit length (i.e., average number of words per T-unit); 3) number of clauses; 4) clause length (i.e., average number of words per clause); 5) number of sentences; 6) sentence length (i.e., average number of words per sentence); 7) percentage of passive sentences, and 8) percentage of passive verb phrases. To analyze the passages for semantic complexity, the number of words not included in *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) was determined for grade levels 4, 6, 8, 10, 12, 13, and 16.

Identifying T-units and clauses is neither a straightforward nor simplistic task. Therefore, three raters independently identified clauses and T-units for each set of passages. The T-unit and clause identification data were then analyzed to determine the inter-rater agreement. The initial T-unit identification agreement among the three raters for the Miller and Coleman (1967) passages (raters 1 and 2: r = .950, raters 2 and 3: r = .951; raters 1 and 3: r = .964) were acceptable. The initial clause identification agreement among the three raters was also acceptable (raters 1 and 2: r = .927, raters 2 and 3: r = .944; raters 1 and 3: r = .895).The overall inter-rater reliability among the three raters for the T-unit and clause identifications for all sets of passages were r = .984 and r = .972, respectively. Where discrepancies existed, the author of the study made the final decision.

The number of words for the Miller and Coleman (1967) passages ranged from 149 to 152. Therefore, variable measures were adjusted for exactly 150 words. For example, passage 9 included 151 words and 8 sentences. The number of sentences was adjusted by dividing the actual number of sentences by the total number of words and multiplying that product by 150 [i.e., (8/151)*150= 7.947].

Exploratory regression analysis was used to investigate the variance in total CT scores accounted for by the semantic and syntactic variables under examination. These analyses were conducted to determine which syntactic and semantic variables should be retained for further consideration in the second phase of the investigation. Simple linear regression analyses were conducted with the Miller and Coleman total CT scores (i.e., the sum of CT I, CT II, and CT III scores for each passage; Aquino, 1969) as the dependent variable and: 1) number of unfamiliar words, 2) number of T-units; 3) T-unit length (i.e., average number of words per T-unit); 4) number of clauses; 5) clause length (i.e., average number of words per clause); 6) number of sentences; 7) sentence length (i.e., average number of words per sentence); 8) percentage of passive sentences, and 9) percentage of passive verb phrases as the independent variables. The regression analyses allowed the identification of predictor variables that accounted for a statistically significant amount of variance in cloze scores while controlling for the effects of the other variables. Data and standardized residuals for predictor variables were also plotted to facilitate the identification of instances of curvilinearity. The results from the standard multiple regression analyses were used to identify the variables to be used in the next phase of the investigation.

The simple regression analyses indicated that all of the syntactic variables accounted for a statistically significant amount of variance in total CT scores: 1) number of T-units, $b = 36.1$, $t_{(34)} = 7.503$, $R^2 = .623$, $p < .0005$; 2) T-unit length (i.e., average number of words per T-unit), $b = -28.019$, $t_{(34)} = -5.587$, $R^2 = .479$, $p < .0005$; 3) number of clauses, $b = 28.721$, $t_{(34)} = 5.865$, $R^2 = .503$, $p < .0005$; 4) clause length (i.e., average number of words per clause), $b = -42.293$, $t_{(34)} = -5.005$, $R^2 = .424$, $p < .0005$; 5) number of sentences, $b = 32.956$, $t_{(34)} = 5.983$, $R^2 = .513$, $p < .0005$; 6) sentence length (i.e., average number of words per sentence), $b = -19.96$, $t_{(34)} = -4.52$, $R^2 = .375$, $p < .0005$; 7) percentage of passive sentences, $b = -541.587$, $t_{(34)} = -3.654$, $R^2 = .282$, $p < .001$; and 8) percentage of passive verb phrases, $b = -277.836$, $t_{(34)} = -2.851$, $R^2 = .192$, $p < .007$ (see Table 5).

Table 5

Correlations for syntactic variables

	#TU	TUL	#C	CL	#S	SL	PPS	PPVP
TCT	.790**	-.692**	.709**	-.651**	.716**	-.613**	-.531**	-.439**
#TU	--	-.901**	.791**	-.660**	.951**	-.864**	-.510**	-.412**
TUL	--	--	-.666**	.589**	-.811**	.929**	.486**	.299
#C	--	--	--	-.894**	.681**	-.554**	-.488**	-.403*
CL	--	--	--	--	-.532**	.486**	.452**	.361*
#S	--	--	--	--	--	-.856**	-.452**	-.357*
SL	--	--	--	--	--	--	.473**	.216
PPS	--	--	--	--	--	--	--	.733**

Note. TCT = Total Cloze test score; #TU = Number of T-units; TUL = T-unit length; #C = Number of Clauses; CL = Clause length; #S = Number of sentences; SL = Sentence length; PPS = Percentage of passive sentences; and PPVP = Percentage of passive verb phrases. ** Correlation significant at .01 level (two-tailed); * Correlation significant at .05 level (two-tailed).

The simple regression analyses also indicated that the semantic variable (number of unfamiliar words) at all levels accounted for a statistically significant amount of variance in total CT scores: 1) level 4, b = -10.68, $t_{(34)}$ = -9.009, R^2 = .705, p < .001; 2) level 6, b = -16.141, $t_{(34)}$ = -8.426, R^2 = .676, p < .001; 3) level 8, b = -26.023 , $t_{(34)}$ = -7.493, R^2 = .623, p < .001; 4) level 10, b = -34.799, $t_{(34)}$ = -7.033, R^2 = .593, p < .001; 5) level 12, b = -40.98, $t_{(34)}$ = -3.819, R^2 = .300, p < .001; 6) level 13, b = -37.849, $t_{(34)}$ = -2.991, R^2 = .208, p < .005; and 7) level 16, b = -27.575, $t_{(34)}$ = -2.03, R^2 = .108, p < .050 (see Table 6 for correlation coefficients).

Table 6

Correlations for number of unfamiliar words

	Level 4	Level 6	Level 8	Level 10	Level 12	Level 13	Level 16
Total CT	-.839**	-.822**	-.789**	-.770**	-.548**	-.456**	-.329
Level 4	--	.974**	.900**	.867**	.616**	.506**	.358*
Level 6	--	--	.943**	.898**	.633**	.533**	.392*
Level 8	--	--	--	.976**	.739**	.657**	.502**
Level 10	--	--	--	--	.837**	.757**	.600**
Level 12	--	--	--	--	--	.952**	.890**
Level 13	--	--	--	--	--	--	-.959**

Note. ** Correlation significant at .01 level (two-tailed); * Correlation significant at .05 level (two-tailed).

Through the simple linear regression results it was determined that all syntactic independent variables accounted for a statistically significant amount of variance in total CT scores. Seven of the eight original syntactic independent variables were retained for further analysis in the next phase of the investigation. Percentage of passive verb phrases only accounted for 19.2% of variance in total CT scores. Percentage of passive sentences accounted for more variance in total CT scores (28.2%) and was strongly correlated with percentage of passive verb phrases (r = .733). It was not necessary to include more than one measure of voice, especially because they were strongly correlated. Therefore, based on variance explained, percentage of passive sentences was retained for further investigation and percentage of passive verb phrases was not retained.

Through the simple linear regression results it was also determined that the semantic independent variable (number of unfamiliar words) at levels 4, 6, 8, 10, 12, and 13 accounted for a statistically significant amount of variance in total CT scores. Numbers of unfamiliar words at those levels were retained for further consideration in the next phase of the investigation. The number of familiar words at level 16 only accounted for 10.8% of variance in, and was not significantly correlated with, total CT scores. The semantic variable at level 16 was, therefore, not retained for further investigation.

Phase II: Formula Creation and Calibration

Phase II of the investigation had two primary purposes. The first purpose was to create and calibrate a new-model formula that incorporated variables that were retained

from the first phase of the study. The second purpose was to recalibrate three existing readability formulas with the same materials used to calibrate the new-model formula.

This component of the results section includes the results of exploratory regression analyses that were conducted to investigate the variance in total CT scores accounted for by all possible combinations of the retained syntactic and semantic variables and, thereby, to create and calibrate a new-model formula. The results showed that four new-model formulas were worthy of retention and further investigation. This component also includes the results of regression analyses that were conducted to recalibrate three existing readability formulas: the same total CT scores served as the dependent variable and the respective components of each existing formula served as the independent variables. This resulted in six recalibrated formulas, because three recalibrated formulas were created for one of the existing formulas due to difficulties encountered during the recalibration process.

The first part of this component includes the details of the new-model formula calibrations. First, the variables retained for this phase of the investigation are identified. Then the 36 syntactic and semantic independent variable combinations that were created and explored and the methods used to analyze these variable combinations are explained. Next, an explanation of how outliers were identified and treated is offered. Subsequent subsections include the results of the analyses for the syntactic and semantic independent variable combinations. The subsections are organized according to the syntactic variable under consideration. The removal of particular Miller and Coleman (1967) passages are then identified and the rationale for their removal is explained. Next, the criteria for selecting new-model formulas for retention and further investigation are outlined. Then,

the new-model formulas that were selected for retention and inclusion for further analyses are identified.

The second part of this component includes the details and results of exploratory regression analyses conducted to recalibrate the existing formulas. The existing formula recalibrated subsection is organized according to formula type. The recalibrated versions of the existing formulas that were retained for further investigation are then identified. A graphic representation of the layout of this entire component is offered in Figure 2.

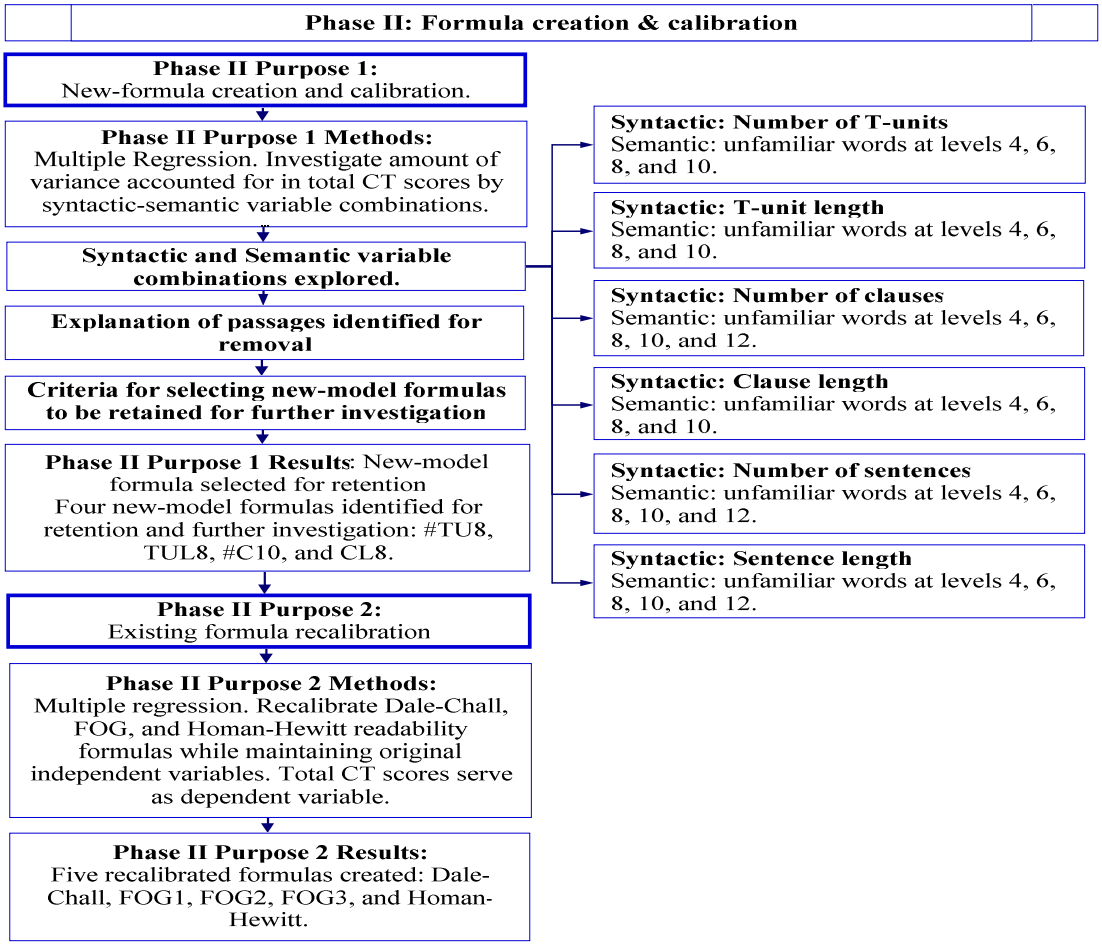


Figure 2. Graphic representation of layout of Phase II results section.

All eight syntactic variables accounted for statistically significant amounts of variance in total CT scores during Phase I and seven of them were retained for this phase of the investigation. Five of the six levels of the semantic independent variable (number of unfamiliar words) accounted for statistically significant amounts of variance in total CT scores and were retained for this phase of the investigation. The usefulness of syntactic variables and how much variance they account for when they were combined with the semantic variable at each level was explored.

Exploratory stepwise multiple regression was conducted to determine the variance in the dependent variable (i.e., Total CT scores) accounted for by syntactic and semantic variable combinations. The syntactic variables were 1) number of T-units; 2) T-unit length (i.e., average number of words per T-unit); 3) number of clauses; 4) clause length (i.e., average number of words per clause); 5) number of sentences; 6) sentence length (i.e., average number of words per sentence); and 7) percentage of passive sentences. The semantic variable was number of familiar words according to the Living Word Vocabulary (Dale & O'Rourke, 1981) at grade levels 4, 6, 8, 10, 12, and 13. The resulting variable combinations were explored with stepwise multiple regression. A regression analysis was conducted for each syntactic variable coupled with the semantic variable at the five retained levels and the voice variable (percentage of passive sentences). This resulted in five possible variable combinations for each syntactic variable. Table 7 outlines these variable combinations. The variable combinations were explored and the regression analyses conducted for the new-model calibrations were conducted with attention to the correlation matrices reported in Phase I. However, lower levels of the semantic variable accounted for more variance in total CT scores .

Therefore, once a particular level of the semantic variable included in a combination failed to account for enough variance to enter the equation along with the syntactic variable, no further analyses were conducted to explore the respective syntactic variable combined with higher levels of the semantic variable.

For all variable combinations explored in the creation and calibration of the new model, the first analysis included all 36 Miller and Coleman (1967) passages. For the second analysis, four passages were removed because, based on Total CT scores, they were the easiest passages. Total CT scores for these four passages were .75 standard deviations above the mean total CT score (M = 1004.278, SD = 184.82). For each subsequent analysis, standardized and studentized residuals were inspected to identify outliers warranting deletion. Passages with high standard residuals were inspected and deleted one at a time until the data set included only passages that had reasonable standardized residuals.

Outliers are typically identified as data cases that have standardized residual values greater than two and they should be examined (Pedhazur, 1997). This common practice was used for each regression analysis, but the distribution of residuals was also inspected. The calculation of standardized residuals is based on the assumption that all residuals have the same variance; whereas, the calculation of studentized residuals does not require this assumption (Pedhazur, 1997). Therefore, studentized residual scatter plots were also inspected. The studentized residual scatter plots showed almost identical distributions of residuals as the standardized residuals. Therefore, scatter plots of standardized residuals and studentized residuals were considered and the case wise diagnostic values for standardized residuals were used for identification of outliers.

Table 7

All potential variable combinations

Syntactic	Semantic— unfamiliar words	Voice
Number of T-units	Level 4	Percentage of passive sentences
Number of T-units	Level 6	Percentage of passive sentences
Number of T-units	Level 8	Percentage of passive sentences
Number of T-units	Level 10	Percentage of passive sentences
Number of T-units	Level 12	Percentage of passive sentences
Number of T-units	Level 13	Percentage of passive sentences
T-unit length	Level 4	Percentage of passive sentences
T-unit length	Level 6	Percentage of passive sentences
T-unit length	Level 8	Percentage of passive sentences
T-unit length	Level 10	Percentage of passive sentences
T-unit length	Level 12	Percentage of passive sentences
T-unit length	Level 13	Percentage of passive sentences
Number of clauses	Level 4	Percentage of passive sentences
Number of clauses	Level 6	Percentage of passive sentences
Number of clauses	Level 8	Percentage of passive sentences
Number of clauses	Level 10	Percentage of passive sentences
Number of clauses	Level 12	Percentage of passive sentences
Number of clauses	Level 13	Percentage of passive sentences
Clause length	Level 4	Percentage of passive sentences
Clause length	Level 6	Percentage of passive sentences
Clause length	Level 8	Percentage of passive sentences

Syntactic	Semantic— unfamiliar words	Voice
Clause length	Level 10	Percentage of passive sentences
Clause length	Level 12	Percentage of passive sentences
Clause length	Level 13	Percentage of passive sentences
Number of sentences	Level 4	Percentage of passive sentences
Number of sentences	Level 6	Percentage of passive sentences
Number of sentences	Level 8	Percentage of passive sentences
Number of sentences	Level 10	Percentage of passive sentences
Number of sentences	Level 12	Percentage of passive sentences
Number of sentences	Level 13	Percentage of passive sentences
Sentence length	Level 4	Percentage of passive sentences
Sentence length	Level 6	Percentage of passive sentences
Sentence length	Level 8	Percentage of passive sentences
Sentence length	Level 10	Percentage of passive sentences
Sentence length	Level 12	Percentage of passive sentences
Sentence length	Level 13	Percentage of passive sentences

The following six subsections include the results of the regression analyses described above. These subsections are organized according to syntactic variable. Within each syntactic-variable subsection, the regression results obtained from coupling the respective syntactic variable each level of the semantic variable is discussed in turn. First, the results of regression analyses that included number of T-units as the syntactic variable and its coupling with the semantic variable at each of its levels are described. Second, the results of regression analyses that included T-unit length as the syntactic variable and its

coupling with the semantic variable at each of its levels are discussed. Third, the results of regression analyses that included number of clauses as the syntactic variable and its coupling with the semantic variable at each of its levels are reported. Fourth, the results of regression analyses that included number clauses as the syntactic variable and its coupling with the semantic variable at each of its levels are outlined. Fifth, the results of regression analyses that included number of sentences as the syntactic variable and its coupling with the semantic variable at each of its levels are described. Sixth, the results of regression analyses that included sentence length as the syntactic variable and its coupling with the semantic variable at each of its levels are discussed.

Number of T-units as Syntactic Variable

Four sets of regression analyses were conducted to determine the variance in total CT scores accounted for by the combination of number of T-units, percentage of passive sentences, and number of unfamiliar words (at each grade level). In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. With all 36 passages included in the regression analysis, unfamiliar words at level 4 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2=.843, $F_{(2,33)}$ = 88.764, p < .0005; see Table 8). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 4 accounted for a significant amount of variance in total CT scores. Removing passages with unreasonably high residuals did not allow the syntactic variable (number of T-units) to enter the equation.

Table 8

Regression results for number of T-units as the syntactic variable

Semantic Variable	R^2	Adj R^2	F	β #TU	β UFW	Regression Equation
UFW-4*	.843	.834	88.764	.457	-.575	Y′=866.73-(7.316*UFW)+(20.872*#TU)
UFW-6*	.845	.835	89.740	.488	-.559	Y′=840.40-(10.97*UFW)+(22.30*#TU)
UFW-8**	.828	.815	67.296	.257	-.746	Y′=916.646-(18.506*UFW)+(13.544*#TU)
UFW-10**	.789	.774	52.434	.279	-.708	Y′=905.945-(24.218*UFW)+(14.665*#TU)

Note. *All passages included. **Four passages with highest total CT scores and outliers removed. UFW = unfamiliar words, #TU = Number of T-units. All analysis reported in this table are significant at the .001 level.

With all 36 passages included in the regression analysis, unfamiliar words at level 6 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .845, $F_{(2,33)}$ = 89.740, p < .0005; see Table 8). When the four passages with the highest total CT scores were removed, only unfamiliar words at the 6th grade level accounted for a significant amount of variance in total CT scores. Removing passages with unreasonably high residuals did not allow the syntactic variable (number of T-units) to enter the equation.

With all 36 passages included in the regression analysis, unfamiliar words at level 8 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .827, $F_{(2,33)}$ = 78.719, p < .0005). When the four passages with the highest total CT scores were removed, unfamiliar words at level 8 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .781, $F_{(2,29)}$ = 51.565, p < .0005). When outlying passage 5 was removed, unfamiliar words at level 8 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .828, $F_{(2,28)}$ = 67.296, p < .0005; see Table 8).

With all 36 passages included in the regression analysis, unfamiliar words at level 10 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .801, $F_{(2,33)}$ = 66.278, p < .0005). When the four passages with the highest total CT scores were removed, unfamiliar words at level 10 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .726, $F_{(2,29)}$ = 38.367, p < .0005). When outlying passage 5 was removed, unfamiliar words at level 10 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .789, $F_{(2,28)}$ = 52.434, p < .0005; see Table 8). Unfamiliar words at level 10 combined with number of T-units accounted for less variance in total CT scores than unfamiliar words at level 8 combined with number of T-units. The correlations between CT scores and unfamiliar words at levels 12 and 13 are weaker than the correlation between CT scores and unfamiliar words at level 10; therefore, regression analyses were not conducted for unfamiliar words at levels 12 or 13 combined with number of T-units.

T-unit length as the Syntactic Variable

Four sets of regression analyses were conducted to determine the variance in total CT scores accounted for by the combination of number of T-unit length, percentage of passive sentences, and number of unfamiliar words (at each grade level). In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. With all 36 passages included in the regression analysis, unfamiliar words at level 4 and T-unit length accounted for a statistically significant amount of variance in total CT scores (R^2 = .752, $F_{(2,33)}$ = 50.065, p < .0005; see Table 9). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 4 accounted for a significant amount of variance in total CT scores. Removing outlying passages did not have an effect on the resulting regression equation.

With all 36 passages included in the regression analysis, unfamiliar words at level 6 and T-unit length accounted for a statistically significant amount of variance in total CT scores (R^2 = .750, $F_{(2,33)}$ = 49.627, p < .0005; see Table 9). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 6 accounted for a significant amount of variance in total CT scores. Removing outlying passages did not have an effect on the resulting regression equation.

With all 36 passages included in the regression analysis, unfamiliar words at level 8 and T-unit length accounted for a statistically significant amount of variance in total CT scores (R^2 = .747, $F_{(2,33)}$ = 48.605, p < .0005). When the four passages with the highest total CT scores were removed, unfamiliar words at level 8 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .788,

$F_{(2,29)}$ = 53.840, p < .0005). When outlying passage 5 was also removed, unfamiliar words at level 8 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .831, $F_{(2,28)}$ = 68.691, p < .0005; see Table 9).

Table 9

Regression results for T-unit length as the syntactic variable

Semantic Variable	R^2	Adj R^2	F	β TUL	β UFW	Regression Equation
UFW-4*	.752	.737	50.065	-.278	-.667	Y´=1281.862-(8.487*UFW)-(11.245*TUL)
UFW-6*	.750	.735	49.627	-.331	-.634	Y´=1300.213-(12.442*UFW)-(13.421*TUL)
UFW-8**	.831	.819	68.691	-.248	-.777	Y´=1192.242-(19.278*UFW)-(8.461*TUL)
UFW-10**	.787	.772	51.684	-.256	-.745	Y´=1198.431- (25.469*UFW)-(8.743*TUL)

Note. *All passages included. **Four passages with highest total CT scores and outliers removed. UFW = unfamiliar words, TUL = T-unit length. All analysis reported in this table are significant at the .0005 level.

With all 36 passages included in the regression analysis, unfamiliar words at level 10 and T-unit length accounted for a statistically significant amount of variance in total CT scores (R^2 = .717, $F_{(2,33)}$ = 41.906, p < .0005). When the four passages with the highest total CT scores were removed, unfamiliar words at level 10 and number of T-units accounted for a statistically significant amount of variance in total CT scores (R^2 = .729,

$F_{(2,29)} = 38.941$, $p < .0005$). When outlying passage 5 was removed, unfamiliar words at level 10 and number of T-units accounted for a statistically significant amount of variance in total CT scores ($R^2 = .787$, $F_{(2,28)} = 51.684$, $p < .0005$; see Table 9). Unfamiliar words at level 10 combined with T-unit length accounted for less variance in total CT scores than unfamiliar words at level 8 combined with T-unit length. The correlations between CT scores and unfamiliar words at levels 12 and 13 are weaker than the correlation between CT scores and unfamiliar words at level 10; therefore, further regression analyses were not conducted for unfamiliar words at levels 12 or 13 combined with T-unit length.

Number of Clauses as the Syntactic Variable

Five sets of regression analyses were conducted to determine the variance in total CT scores accounted for by the combination of number of clauses, percentage of passive sentences, and number of unfamiliar words (at each grade level). In all analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. This variable was, therefore, removed from further consideration. With all 36 passages included in the regression analysis, unfamiliar words at level 4 and number of clauses accounted for a statistically significant amount of variance in total CT scores ($R^2 = .798$, $F_{(2,33)} = 65.380$, $p < .0005$; see Table 10). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 4 accounted for a significant amount of variance in total CT scores. Two passages showed high standardized residuals, but there appeared to be no legitimate reason for removing them and conducting further analyses.

Table 10

Regression results for number of clauses as the syntactic variable

Semantic Variable	R^2	Adj R^2	F	β #C	β UFW	Regression Equation
UFW-4*	.798	.786	65.380	.363	-.645	Y′=1141.039-(-8.20 *UFW)+(14.70 *#C)
UFW-6*	.783	.770	59.458	.383	-.621	Y′=853.110-(12.195*UFW)+(15.529*#C)
UFW-8**	.786	.772	53.400	.233	-.752	Y′=929.636-(19.135*UFW)+(8.956*#C)
UFW-10**	.818	.805	60.672	.224	-.779	Y′=944.244-(26.154*UFW)+(8.424*#C)
UFW-12**	.448	.409	11.747	.459	-.342	Y′=747.509-(19.716*UFW)+(17.643 *#C)

Note. *All passages included. **Four passages with highest total CT scores and outliers removed. UFW = unfamiliar words, #C = Number of clauses. All analysis reported in this table are significant at the .0005 level.

With all 36 passages included in the regression analysis, unfamiliar words at level 6 and number of clauses accounted for a statistically significant amount of variance in total CT scores ($R^2 = .783$, $F_{(2,33)} = 59.458$, $p < .001$; see Table 10). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 6 accounted for a significant amount of variance in total CT scores. Removing outlying passages did not have an effect on the resulting regression equation.

With all 36 passages included in the regression analysis, unfamiliar words at level 8 and number of clauses accounted for a statistically significant amount of variance in total CT scores (R^2 = .762, $F_{(2,33)}$ = 52.752, p < .0005). When the four passages with the highest total CT scores were removed, unfamiliar words at level 8 and number of clauses accounted for a statistically significant amount of variance in total CT scores (R^2 = .786, $F_{(2,29)}$ = 53.40, p < .0005; see Table 10). When outlying passage 5 was removed, only unfamiliar words at level 8 accounted for a statistically significant amount of variance in total CT scores.

With all 36 passages included in the regression analysis, unfamiliar words at level 10 and number of clauses accounted for a statistically significant amount of variance in total CT scores (R^2=.747, $F_{(2,33)}$ = 48.734, p < .0005). When the four passages with the highest total CT scores were removed, unfamiliar words at level 10 and number of clauses accounted for a statistically significant amount of variance in total CT scores (R^2 =.743, $F_{(2,29)}$ = 41.917, p < .0005). When outlying passages 5 and 31 were removed, unfamiliar words at level 10 and number of clauses accounted for a statistically significant amount of variance in total CT scores (R^2 =.818, $F_{(2,27)}$ = 60.672, p < .0005; see Table 10).

With all 36 passages included in the regression analysis, unfamiliar words at level 12 and number of clauses accounted for a statistically significant amount of variance in total CT scores (R^2 = .576, $F_{(2,33)}$ = 22.392, p < .0005). When the four passages with the highest total CT scores were removed, unfamiliar words at level 12 and number of clauses accounted for a statistically significant amount of variance in total CT scores (R^2 = .448, $F_{(2,29)}$ = 11.747, p < .0005; see Table 10). Unfamiliar words at level 12 combined with number of clauses accounted for less variance in total CT scores than unfamiliar

words at level 10 combined with number of clauses. The correlations between CT scores and unfamiliar words at level 13 are weaker than the correlation between CT scores and unfamiliar words at level 12; therefore, further regression analyses were not conducted for unfamiliar words at level 13 combined with number of clauses.

Clauses Length as the Syntactic Variable

Four sets of regression analyses were conducted to determine the variance in total CT scores accounted for by the combination of clause length, percentage of passive sentences, and number of unfamiliar words (at each grade level). In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. With all 36 passages included in the regression analysis, unfamiliar words at level 4 and clause length accounted for a statistically significant amount of variance in total CT scores ($R^2 = .742$, $F_{(2,33)} = 47.440$, $p < .0005$; see Table 11). When the four passages with the highest total CT scores were removed, only unfamiliar words at the level 4 accounted for a significant amount of variance in total CT scores. When outlying passages 5 and 31 were removed, the equation included only unfamiliar words at level 4.

With all 36 passages included in the regression analysis, unfamiliar words at level 6 and clause length accounted for a statistically significant amount of variance in total CT scores ($R^2 = .727$, $F_{(2,33)} = 43.917$, $p < .0005$; see Table 11). When the four passages with the highest total CT scores were removed, only unfamiliar words at the level 6 accounted for a significant amount of variance in total CT scores. When all outliers were removed, the equation included only unfamiliar words at level 6.

Table 11

Regression results for clause length as the syntactic variable

Semantic Variable	R^2	Adj R^2	F	β CL	β UFW	Regression Equation
UFW-4*	.742	.726	47.440	-.239	-.698	Y′=1273.568-(8.885*UFW)-(15.516*CL)
UFW-6*	.727	.710	43.017	-.273	-.668	Y′=1281.468-(13.102*UFW)-(17.744*CL)
UFW-8**	.849	.838	75.934	-.180	-.818	Y′=1169.09-(19.92*UFW)-(9.597*CL)
UFW-10**	.814	.800	59.158	-.216	-.778	Y′=1190.825- (26.124*UFW)-(11.559*CL)

Note. *All passages included. **Four passages with highest total CT scores and outliers removed. UFW = unfamiliar words, CL = clause length. All analysis reported in this table are significant at the .0005 level.

With all 36 passages included in the regression analysis, unfamiliar words at level 8 and clause length accounted for a statistically significant amount of variance in total CT scores ($R^2 = .698$, $F_{(2,33)} = 38.068$, $p < .0005$). When the four passages with the highest total CT scores were removed, only unfamiliar words at the level 8 accounted for a significant amount of variance in total CT scores. When outlying passages 5 and 31 were removed, unfamiliar words at level 8 and clause length accounted for a statistically significant amount of variance in total CT scores ($R^2 = .849$, $F_{(2,27)} = 75.934$, $p < .0005$; see Table 11).

With all 36 passages included in the regression analysis, unfamiliar words at level 10 and clause length accounted for a statistically significant amount of variance in total CT scores (R^2 = .682, $F_{(2,33)}$ = 35.33, p < .0005). When the four passages with the highest total CT scores were removed, unfamiliar words at level 10 and clause length accounted for a statistically significant amount of variance in total CT scores (R^2 = .729, $F_{(2,29)}$ = 38.915, p < .0005). When outlying passages 5 and 31 were removed, unfamiliar words at level 10 and clause length accounted for a statistically significant amount of variance in total CT scores (R^2 = .814, $F_{(2,27)}$ = 59.158, p < .0005; see Table 11). Unfamiliar words at level 10 combined with clause length accounted for less variance in total CT scores than unfamiliar words at level 8 combined with clause length. The correlations between CT scores and unfamiliar words at levels 12 and 13 are weaker than the correlation between CT scores and unfamiliar words at level 10; therefore, further regression analyses were not conducted for unfamiliar words at levels 12 or 13 combined with clause length.

Number of Sentences as the Syntactic Variable

Five sets of regression analyses were conducted to determine the variance in total CT scores accounted for by the combination of number of sentences, percentage of passive sentences, and number of unfamiliar words (at each grade level). In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. With all 36 passages included in the regression analysis, unfamiliar words at level 4 and number of sentences accounted for a statistically significant amount of variance in total CT scores (R^2 = .843, $F_{(2,33)}$ = 88.608, p < .0005; see Table 12). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 4 accounted for a

significant amount of variance in total CT scores. When all outliers were removed, the equation included only unfamiliar words at level 4.

Table 12

Regression results for number of sentences as the syntactic variable

Semantic Variable	R^2	Adj R^2	F	β #S	β UFW	Regression Equation
UFW-4*	.843	.834	88.608	.419	-.647	Y′=924.589-(8.232*UFW)+ (19.269 *#S)
UFW-6*	.836	.826	84.320	.444	-.631	Y′=901.978-(12.378 *UFW)+(20.421 *#S)
UFW-8*	.811	.800	70.884	.475	-.597	Y′=885.812 -(19.694*UFW)+ (21.839*#S)
UFW-10*	.781	.768	58.900	.478	-.570	Y′=886.039 - (25.772*UFW)+ (21.994*#S)
UFW-12**	.448	.409	1.385	.349	-.463	Y′=817.620 - (26.542*UFW)+ (23.229*#S)

Note: *All passages included. **Four passages with highest total CT scores and outliers removed. UFW = unfamiliar words, #S = number of sentences. All analysis reported in this table are significant at the .0005 level.

With all 36 passages included in the regression analysis, unfamiliar words at level 6 and number of sentences accounted for a statistically significant amount of variance in total CT scores ($R^2 = .836$, $F_{(2,33)} = 84.320$, $p < .0005$; see Table 12). When the four

passages with the highest total CT scores were removed, only unfamiliar words at level 6 accounted for a significant amount of variance in total CT scores. When all outliers were removed, the equation included only unfamiliar words at level 6.

With all 36 passages included in the regression equation, unfamiliar words at level 8 and number of sentences accounted for a statistically significant amount of variance in total CT scores (R^2 = .811, $F_{(2,33)}$ = 70.884, p < .0005; see Table 12). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 8 accounted for a significant amount of variance in total CT scores. When all outliers were removed, the equation included only unfamiliar words at level 8.

With all 36 passages included in the regression analysis, unfamiliar words at level 10 and number of sentences accounted for a statistically significant amount of variance in total CT scores (R^2 = .781, $F_{(2,33)}$ = 58.90, p < .0005; see Table 12). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 10 accounted for a significant amount of variance in total CT scores. When all outliers were removed, the equation included only unfamiliar words at level 10.

With all 36 passages included in the regression analysis, unfamiliar words at level 12 and number of sentences accounted for a statistically significant amount of variance in total CT scores, (R^2 = .597, $F_{(2,33)}$ = 24.413, p < .0005). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 12 accounted for a significant amount of variance in total CT scores. When outlying passage 5 was removed, unfamiliar words at level 12 and number of sentences accounted for a statistically significant amount of variance in total CT scores (R^2 = .448, $F_{(2,28)}$ = 11.385, p < .0005; see Table 12). Unfamiliar words at level 12 combined with number of sentences

accounted for less variance in total CT scores than unfamiliar words at level 10 combined with number of sentences. The correlations between CT scores and unfamiliar words at level 13 are weaker than the correlation between CT scores and unfamiliar words at level 12; therefore, further regression analyses were not conducted for unfamiliar words at level 13 combined with number of sentences.

Sentence Length as the Syntactic Variable

Five sets of regression analyses were conducted to determine the variance in total CT scores accounted for by the combination of sentence length, percentage of passive sentences, and number of unfamiliar words (at each grade level). In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. With all 36 passages included in the regression analysis and in an analysis when the four passages with the highest total CT scores were removed, only unfamiliar words at level 4 accounted for a significant amount of variance in total CT scores. When all outliers were removed, the equation included only unfamiliar words at level 4.

With all 36 passages included in the regression analysis, unfamiliar words at level 6 and sentence length accounted for a statistically significant amount of variance in total CT scores ($R^2 = .724$, $F_{(2,33)} = 43.291$, $p < .0005$; see Table 13). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 6 accounted for a significant amount of variance in total scores. When all outliers were removed, the equation included only unfamiliar words at level 6.

With all 36 passages included in the regression analysis, unfamiliar words at level 8 and sentence length accounted for a statistically significant amount of variance in total

CT scores (R^2 = .708, $F_{(2,33)}$ = 40.080, p < .0005; see Table 13). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 8 accounted for a significant amount of variance in total scores. When all outliers were removed, the equation included only unfamiliar words at level 8.

Table 13

Regression results for sentence length as the syntactic variable

Semantic Variable	R^2	Adj R^2	F	β SL	β UFW	Regression Equation
UFW-6*	.724	.707	43.291	-.256	-.690	Y′=1253.468 -(13.544*UFW)-(8.329*SL)
UFW-8*	.708	.691	40.080	-.327	-.644	Y′=1289.865 -(21.240 *UFW)-(10.637 *SL)
UFW-10**	.772	.756	47.470	-.211	-.780	Y′=1175.387 -(26.657*UFW)-(5.990 *SL)
UFW-12**	.457	.418	11.783	-.355	-.481	Y′=1207.778 -(27.557 *UFW)-(10.086*SL)

Note. *All passages included. **Four passages with highest total CT scores and outliers removed. UFW = unfamiliar words, SL = sentence length. All analysis reported in this table are significant at the .0005 level.

With all 36 passages included in the regression analysis, unfamiliar words at level 10 and sentence length accounted for a statistically significant amount of variance in total CT scores (R^2 = .685, $F_{(2,33)}$ = 35.830, p < .0005). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 10 accounted for a

significant amount of variance in total scores. When outlying passage 5 was removed, unfamiliar words at level 10 and sentence length accounted for a statistically significant amount of variance in total CT scores (R^2 = .772, $F_{(2,28)}$ = 47.470, p < .0005; see Table 13).

With all 36 passages included in the regression analysis, unfamiliar words at level 12 and sentence length accounted for a statistically significant amount of variance in total CT scores (R^2 = .483, $F_{(2,33)}$ = 15.409, p < .0005). When the four passages with the highest total CT scores were removed, only unfamiliar words at level 12 accounted for a significant amount of variance in total scores. When outlying passage 5 was removed, unfamiliar words at level 12 and sentence length accounted for a statistically significant amount of variance in total CT scores (R^2 = .457, $F_{(2,28)}$ = 11.783, p < .0005; see Table 13).). Unfamiliar words at level 12 combined with sentence length accounted for less variance in total CT scores than unfamiliar words at level 10 combined with sentence length; therefore. The correlations between CT scores and unfamiliar words at level 13 are weaker than the correlation between CT scores and unfamiliar words at level 12; therefore, further regression analyses were not conducted for unfamiliar words at level 13 combined with clause length.

Passages Identified for Removal

Four passages were initially determined to be inappropriate for inclusion in the current study because they were the easiest of the passages according to their corresponding total CT scores (total CT score M = 1004.28, SD = 184.82). These passages were .75 standard deviations above the total CT mean. This cutoff was determined by inspecting the total CT scores. If the cutoff for identifying passages that

were too easy for inclusion were set at 1 standard deviation above the mean, only two passages would have been removed from the analysis. Therefore, the cutoff of .75 standard deviations seemed more appropriate for filtering the appropriate passages.

With a total CT score of 1141, passage 5 was less than 3 points away from meeting the criterion for removal (.75 *SD* cutoff = 1142.89). In addition, high standardized residuals were observed for passage 5 during stepwise multiple regression analyses that included: 1) number of T-units combined with unfamiliar words at levels 4, 8, and 10; 2) T-unit length combined with unfamiliar words at levels 8 and 10; 3) number of clauses combined with unfamiliar words at levels 4, 8, and 10; 4) clause length combined with unfamiliar words at levels 8 and 10; 5) number of sentences combined with unfamiliar words at levels 8 and 10; and 6) sentence length combined with unfamiliar words at levels 10 and 12. Passage 5 was, therefore, examined to determine whether it was appropriate to consider it an outlier and delete it.

The 32 passages that were initially retained for analysis were sorted according to their total CT scores. The mean number of T-units for the 16 most difficult passages (those with the lowest CT scores) was 8.40 and the mean for the easiest 15 passages (not including passage 5) was 12.20. Passage 5 included 9 T-units, which corresponds better with more difficult passages than the total CT score for passage 5 would insinuate. The average for mean T-unit length for the 16 most difficult passages was 18.30 and the average for the easiest passages was 13.02. The mean T-unit length for passage 5 was 16.67, which corresponds better with more difficult passages than the total CT score for passage 5 would insinuate. The range for number of sentences in the entire set of data (all 36 passages) was 5 to 23.84, with the most difficult passages generally including the

fewest number of sentences. Passage 5 included 6 sentences, which does not correspond with its high total CT score. The mean sentence length range for all passages was 6.29 to 30 words, with the most difficult passages generally including longer sentences. Passage 5 had a mean sentences length of 25 words, which does not correspond with its high total CT score.

Passage 5 also had high standardized residuals whenever unfamiliar words at levels 8 or 12 were included in the analysis. Therefore, passage 5 was inspected according to its number of unfamiliar words at levels 8 and 12. Passage 5 included five unfamiliar words at level 8 and level 12. The mean number of unfamiliar words at levels 8 and 12 for the 20 easiest passages were 1.95 and 1.54, respectively. The numbers of unfamiliar words for passage 5 at levels 8 and 12 were not in accordance with the values of the other passages with high total CT scores (easier passages).

Based on the above data, it was determined appropriate to delete passage 5 whenever it showed unreasonably high standardized residuals in analyses that included number of T-units, T-unit length, number of sentences, sentence length, unfamiliar words at level 8, or unfamiliar words at level 12. The data clearly showed that the that number of T-units, T-unit length, number of sentences, sentence length, unfamiliar words at level 8, or unfamiliar words at level 12 values for passage 5 were not in accordance with its total CT score. These values for passage 5 were in accordance with the values for more difficult passages; although, according to its total CT score, passage 5 is the 5th easiest of all 36 passages.

Passage 31 had a total CT score of 810, which indicated it was the eighth most difficult passage. This passage consistently showed high residuals in analyses that

included number of clauses and clause length. Not including passages previously identified for removal or passage 31, the mean number of clauses for the easiest 15 passages was 17.33 and the mean number of clauses for the hardest 15 passages was 12.4. Passage 31 included 15 clauses and that value was nearly equidistant from the mean values for the easiest and hardest passages. The average mean clause length for the easiest 15 passages (not including those previously deleted) was 8.95 and the average mean clause length for the most difficult 15 passages (not including passage 31) was 12.07. The mean clause length for passage 31 was 9.38, which corresponds with the average mean clause length of the easier passages even though passage 31 is the 7th most difficult passage (according to total CT scores).

Although passage 31 did not show in an inordinately high number of clauses, as compared to other difficult passages, its values were higher than the mean for other difficult passages. The number of clauses and mean clause length data for Passage 31 were used to determine that it was appropriate to remove passage 31 when it showed an unreasonably high standardized residual in analyses that included number of clauses or mean clause length. This resulted in the removal of the same passages for both formulas that included clause measures.

Criteria for Selecting New-model Formulas to be Retained for Further Investigation

The stepwise multiple regression analyses that were conducted to determine the variance in the dependent variable (i.e., Total CT scores) accounted for by syntactic and semantic variable combinations resulted in numerous variable combinations that accounted for a statistically significant amount of variance in total CT scores. It was necessary to select equations to be included in the next phase of the investigation. The

following three criteria were, therefore, established to help make these determinations. First, because the passages with the highest total CT scores were previously identified as inappropriate for the current calibration, regression equations that necessitated the inclusion of these four passages were not explored further. Second, it was necessary to establish a cut off for the amount of variance explained. It was determined that 80% of variance explained was a suitable criterion. Several of the regression equations with the four passages with the highest total CT scores removed accounted for a statistically significant amount of variance in total CT scores. The analyses sets for each of the syntactic variables included at least one equation that accounted for more than 80% of variance. Third, when more than one regression equation for the analyses for a syntactic variable included more than one equation that excluded the four passages with the highest total CT scores and accounted for more than 80% of variance in total CT score, the equation with the highest variance explained was selected for further investigation.

New-model formulas selected for retention.

Based on the above criteria, four regression equations were selected for further investigation. The first regression equation (#TU8) included number of T-units as the syntactic variable and unfamiliar words at level 8 as the semantic variable and accounted for 82.8% of variance in total CT scores (R^2 = .828, $F_{(2,28)}$ = 67.296, p < .0001; see Table 14). The #TU8 regression equation was derived with the four passages with the highest total CT scores and outlying passage 5 removed from the analysis. The #TU8 regression equation is as follows: Y′ = 916.646 - (18.506*UFW) + (13.544*#TU).

Table 14

#TU8 regression results

#TU8 Regression equation: Y´ = 916.646 - (18.506*UFW) + (13.544*#TU).								
				b	*b*	β	β	
R^2	Adj R^2	*F*	*a*	UFW	#TU	UFW	#TU	*P*
.828	.815	67.296	916.646	-18.506	13.544	-.764	.257	.0005

The second regression equation (TUL8) included T-unit length as the syntactic variable and unfamiliar words at level 8 as the semantic variable and accounted for 83.1% of variance in total CT scores (R^2 = .831, $F_{(2,28)}$ = 68.691, p < .0005; see Table 15). The TUL8 regression equation was derived with the four passages with the highest total CT scores and outlying passage 5 removed from the analysis. The TUL8 regression equation is as follows: Y´ = 1192.242 - (19.278*UFW) - (8.461*TUL).

Table 15

TUL8 regression results

TUL8 Regression equation: Y´ = 1192.242 - (19.278*UFW) - (8.461*TUL).								
				b	*b*	β	β	
R^2	Adj R^2	*F*	*a*	UFW	TUL	UFW	TUL	*p*
.831	.819	68.691	1192.242	-19.278	-8.461	-.777	-.248	.0005

The third regression equation (#C10) included number of clauses as the syntactic variable and unfamiliar words at level 10 as the semantic variable and accounted for

81.8% of variance in total CT scores (R^2 = .818, $F_{(2,27)}$ = 60.672, p < .0005; see Table 16). The #C10 regression equation was derived with the four passages with the highest total CT scores and outlying passages 5 and 31 removed from the analysis. The #C10 regression equation is as follows: Y′ = 944.244 – (26.154*UFW) + (8.424*#C).

Table 16

#C10 regression results

#C10 Regression equation: Y′ = 944.244 – (26.154*UFW) + (8.424*#C).								
				b	*b*	β	β	
R^2	Adj R^2	*F*	*a*	UFW	#C	UFW	#C	*p*
.818	.805	60.672	944.244	-26.154	8.424	-.779	-.224	.0005

The fourth regression equation (CL8) included clause length as the syntactic variable and unfamiliar words at level 8 as the semantic variable and accounted for 84.9% of variance in total CT scores (R^2 = .849, $F_{(2,27)}$ = 75.934, p < .0001; see Table 17). The CL8 regression equation was derived with the four passages with the highest total CT scores and outlying passages 5 and 31 removed from the analysis. The CL8 regression equation is as follows: Y′ = 1169.09 - (19.92*UFW) - (9.597*CL).

Equations that included number of sentences as the syntactic variable were not further considered in this study for three reasons. First, resulting equations that accounted for more than 80% of variance in total CT scores necessitated the inclusion of all 36 passages. Specifically, although the number of sentences and number of unfamiliar words at levels 4, 6, or 8 accounted for more than 80% of variance in total CT scores (see Table

12); this required the inclusion of all 36 passages in the regression analysis. Four of these passages (those with the highest total CT scores) were previously identified as inappropriate for formula calibration. When those cases were removed, only the number of unfamiliar words at levels 4, 6, or 8 accounted for a significant amount of variance in the dependent variable.

Table 17

CL8 regression results

CL8 Regression equation: Y´=1169.09-(19.92*UFW)-(9.597*CL).								
				b	*b*	β	β	
R^2	Adj R^2	*F*	*a*	UFW	CL	UFW	CL	*p*
.849	.838	75.934	1169.09	-19.920	-9.597	-.818	-.180	.0005

The number of sentences and unfamiliar words at level 10 accounted for 78.1% of variance in total CT scores, but this also required the inclusion of all 36 passages. Second, the only variable combination that included number of sentences and accounted for a statistically significant amount of variance with the four passages with the highest total CT scores removed, accounted for very little variance in the dependent variable as compared to the other equations explored. Specifically, the number of sentences and unfamiliar words at level 12 only accounted for 44.8% of variance in total CT scores. This did not meet the initially established criterion: 80% of variance explained. Third, although some variable combinations that included number of sentences accounted for a significant amount of variance in total CT scores, this calibration was based on passages

of approximately 150 words. The purpose of the present study was to create a formula that is not only appropriate for regular text passages, but would also appropriate for multiple-choice test items that are converted into pseudo-continuous prose. Even after the pseudo-continuous prose conversion, multiple-choice test items tend to include fewer than 100 words. Thus, the number of sentences is not likely the most appropriate measure of syntactic complexity for multiple-choice test items. Instead, measures of smaller syntactic units (T-units and clauses) that allow for more data points are likely more appropriate for the purposes of the present study.

Equations that included sentence length as the syntactic variable were not further considered in this study for two reasons. First, although several of the variable combinations accounted for a significant amount of variance in total CT scores, none of them accounted for more 80% of variance in total CT scores, which was the criterion established for this study. Second, although some variable combinations that included sentence length accounted for a significant amount of variance in total CT scores, this calibration was based on passages of approximately 150 words. The purpose of this study was to create a formula that is not only appropriate for regular text passages, but is also appropriate for multiple-choice test items that are converted into pseudo-continuous prose. Even after the pseudo-continuous prose conversion, multiple-choice test items tend to include fewer than 100 words. Thus, the sentence length is not likely the most appropriate measure of syntactic complexity for multiple-choice test items. Instead, measures of smaller syntactic units (T-units and clauses), which allow for more data points are likely more appropriate for the purposes of the present study.

Existing Formula Recalibration

This subsection includes the details and results of multiple regression analyses that were used to recalibrate the existing readability formulas. These recalibrations were conducted to provide the most consistent comparison possible across the new-model formula and existing formula results during Phase III of the investigation. Dale-Chall (1995), FOG, and Homan-Hewitt readability formulas were recalibrated and the results of these recalibrations are discussed in turn. Multiple regression techniques were used with total CT scores from the Miller and Coleman (1967) passages as the dependent variable and each respective components of each formula as the independent variables. Recalibrating these formulas, while retaining their established variables, provided a consistent comparison of the existing formula and new-model results during the next phase of this investigation. Below are the original readability formulas:

Dale-Chall Cloze (Chall, 1995) = 64 – (.95) (X_1) – (.69) (X_2)

(Where X_1= number of unfamiliar words and X_2=average sentence length.)

Gunning FOG Index = .4 (*sl*) + (*hard words*)

(Where *sl* = sentence length and *hard words* = percentage of hard words.)

Homan-Hewitt = 1.76 + (.15)(*WNUM*) + (.69)(*WUNF*) – (.51)(*WLON*).

(Where *WNUM* = sentence complexity, *WUNF* = number of difficult words, and *WLON* = word length)

Dale-Chall (1995) recalibration.

Stepwise multiple regression was used to recalibrate the Dale-Chall (1995) readability formula. The independent variables were number of unfamiliar words (according to Chall-Dale list) and average sentence length and the dependent variable was total CT

scores. The first analysis included data for all 36 Miller Coleman (1967) passages. The number of unfamiliar words accounted for a statistically significant amount of variance in total CT scores (R^2 = .728, $F_{(1,34)}$ = 91.194, p < .0005). Average sentence length was not included in the solution. In an analysis with the four passages with the highest total CT score removed, unfamiliar words accounted for a statistically significant amount of variance without allowing sentence length to enter the equation (R^2 = .837, $F_{(1, 30)}$ = 154.135, p < .0005).

Because the objective for this portion of the study was to recalibrate the existing Dale-Chall (1995) formula with the Miller and Coleman (1967) passages, both independent variables needed to enter the equation. Therefore, two strategies were used to determine how these independent variables should be weighted. First, standardized residuals were examined to determine whether any passages should be removed. Ideally, all 32 passages would have been included in the equation, but with all passages included, both variables did not enter the equation. Thus, it was deemed appropriate to explore standardized residuals to determine whether there were outliers that should be removed. Second, the 32 passages were included and hierarchical multiple regression was used to force both independent variables into the equation in the order in which Chall and Dale indicated they should enter.

When the four passages with the highest total CT scores were removed, passage 31 showed a high standardized residual. The predictor variable values for passage 31 were, therefore, inspected. The range for number of unfamiliar words in the entire set of data (all 36 passages) was 0 to 51.34, with the most difficult passages generally including the greatest number of unfamiliar words. The average number of unfamiliar words for the 16

most difficult passages (not including passage 31) was 30.43 and the average number of unfamiliar words for the 15 easiest passages (not including previously deleted passages) was 6.60. Passage 31 had 21 unfamiliar words. This was generally in accordance with the more difficult passages, as would be expected because it was the seventh most difficult passage (according to total CT scores). On the other hand, this value was slightly lower than would have been expected. Specifically, passage 31 had fewer unfamiliar words than 12 of the 14 most difficult passages.

Sentence length for passage 31 was then examined. The range for average sentence length in the entire set of data (all 36 passages) was 6.29 to 30 words, with the most difficult passages generally including the sentences with the highest average sentence length. Passage 31 had an average sentence length of 30 words, which was greater than the average sentence length for 5 of the six passages that were more difficult than passage 31. Based on the predictor variable values and standardized residuals for passage 31, it was deemed appropriate to delete it.

When the four passages with the highest total CT scores and outlying passage 31 were removed, number of unfamiliar words and average sentence length accounted for a statistically significant amount of variance in total CT scores (R^2 = .881, $F_{(2,28)}$ = 103.784, p < .0005; see Table 18).

FOG (Gunning, 1952) recalibration.

Unlike the other existing formulas explored in the current study, the FOG formula is a linear equation but it is not a regression equation. The two independent variables, sentence length and percentage of hard words, are added and multiplied by a constant of .4. To recalibrate this formula, the original independent variables were retained and

multiple regression analysis methods were used. Because the original formula involved adding the two independent variables without weighting either of them, two approaches were used. First, the independent variables were entered independently and several multiple regression analyses were conducted with total CT scores as the dependent variable. Second, the independent variables were added together to create a single independent variable and simple linear regression was conducted with total CT scores as the dependent variable. All regression analyses are reported below.

Table 18

Stepwise regression results from Dale-Chall recalibration

Number of Passages included	R^2	Adj R^2	F	β SL	β UFW	Regression Equation
31	.881	.873	103.784	.164	-1.016	Y′=1046.50-(8.849*UFW)+(4.984*SL)

Note. UFW = unfamiliar words, SL = average sentence length.

Stepwise multiple regression was used to recalibrate the FOG (1995) readability formula. The independent variables were percentage of hard words (words with more than two syllables) and average sentence length and the dependent variable was total CT scores. The first analysis included data for all 36 Miller Coleman (1967) passages. The percentage of hard words and average sentence length accounted for a statistically significant amount of variance in total CT scores ($R^2 = .740$, $F_{(2, 33)} = 46.895$, $p < .0005$;

see Table 19). The four passages with the highest total CT score were then removed and the regression was conducted again. Only percentage of hard words accounted for a statistically significant amount of variance without allowing average sentence length to enter the equation (R^2 = .833, $F_{(1, 30)}$ = 149.251, p < .0005). Removing additional outliers did not allow average sentence length to enter the equation.

Table 19

Stepwise regression results for FOG recalibration

N of Passages included	R^2	Adj R^2	*F*	β HW	β SL	Regression Equation
36	.740	.724	46.895	-.699	-.259	Y´=1277.463-(8.849*HW)+(4.984*SL)

Note. HW = percentage of hard words, SL = average sentence length.

Because the objective for this portion of the study was to recalibrate the existing FOG formula with the Miller and Coleman (1967) passages, it was necessary for both independent variables to enter the equation. When all 36 Miller and Coleman passages were included in the regression analysis, both variables entered the equation. In contrast, when the four passages with the highest total CT scores and potential outliers were removed, only one variable (percentage of hard words) entered the equation. All other formula calibrations in this study involved the deletion of the four passages with the highest total CT scores. Therefore, to allow the most consistent comparison of regression

results possible, the four passages were removed and both independent variables were forced into the equation.

The 32 passages (the four with the highest total CT scores were removed) were included and hierarchical multiple regression was conducted to force both independent variables into the equation. Percentage of hard words and average sentence length were the independent variables and total CT scores was the dependent variable. Gunning did not specify the order of entry for the variables; therefore, two orders of entry were explored. From both full models, percentage of hard words and average sentence length accounted for 83.3% of variance in total CT scores (R^2 = .833, $F_{(2,29)}$ = 72.226, p < .0005; see Table 20). When percentage of hard words was entered first in the equation, it explained all 83.3% of variance in total CT scores (p < .0005; see Table 20). Average sentence length did not account for any additional variance in total CT scores (p = .865). When average sentence length was entered first in the equation, it explained 17.6% of variance in total CT scores (p < .017; see Table 20). Percentage of hard words accounted for an additional 65.7% of variance in total CT scores over and above the variance accounted for by percentage of hard words (p = .0005). Both orders of entry resulted in the same regression equation.

For the next set of FOG recalibration regression analyses, the independent variables (sentence length and percentage of hard words) were added together to create a single independent variable. Simple linear regression was conducted with the sum of sentence length and percentage of hard words as the independent variable and total CT scores as the dependent variable. With all 36 passages included, the summed independent variable accounted for a statistically significant amount of variance in total CT scores (b =

-14.042, $t_{(34)}$ = -9.178, R^2 = .712, p < .0005; see Table 21). When the four passages with the highest total CT scores were removed, the summed independent variable accounted for a statistically significant amount of variance in total CT scores (b = -11.375, $t_{(30)}$ = -8.081, R^2 = . 685, p < .0005; see Table 21). When the four passages with the highest total CT scores and outlying passage 5 were removed, the summed independent variable accounted for a statistically significant amount of variance in total CT scores (b = -11.469, $t_{(29)}$ = -8.897, R^2 = . 732, p < .0005; see Table 21).

Table 20

Hierarchical regression results for FOG recalibration

	Variables	R^2 change	F change	p change	Regression equation
HW entered	HW	.833	149.251	.0005	Y′=1109.175-(18.193*HW)-
1st	SL	.000	.029	.865	(.412*SL)
SL entered	SL	.176	6.401	.017	
1st	HW	.657	113.951	.0005	

Note. HW = hard words, SL = average sentence length.

Based on the results of the above regression analyses, it was determined that three formulas would be used for comparisons to the current model. The first regression equation selected was derived via the stepwise multiple regression method including all 36 passages. The second equation was derived via hierarchical multiple regression with the four passages with the highest total CT scores removed and sentence length entered

first. The third equation was derived using simple linear regression with the two independent variables combined into a single independent variable and the four passages with the highest total CT scores and one additional outlying passage removed.

Table 21

Simple regression results for FOG recalibration with independent variables combined

N of Passages	R^2	Adj R^2	*t*	*a*	*b*	β	Regression Equation
36	. 712	.704	-9.178	1347.461	-14.042	-.844	Y′=1347.461-(14.042*(HW+SL))
32	.685	.675	-8.081	1261.026	-11.375	-.828	Y′=1261.026-(11.375*(HW+SL))
31	.732	.723	-8.897	1257.188	-11.469	-.856	Y′=1257.188-(11.469*(HW+SL))

Note. HW = hard words, SL = average sentence length.

Homan and Hewitt recalibration.

Difficulties were encountered in the recalibration of the Homan and Hewitt formula. Validation studies published by Homan et al. (1994) and Hewitt and Homan (2004) indicated that unfamiliar words should be identified at level 4. Using this level of the semantic variable did not allow all of the variables included in the formula to enter the equation. Therefore, several multiple regression approaches were necessary to recalibrate the Homan and Hewitt formula. The analyses conducted are described in detail below.

Stepwise multiple regression analyses were initially conducted to recalibrated the Homan-Hewitt readability formula. The dependent variable was total CT score. The independent, syntactic variable was T-unit length and the independent, semantic variables were number of unfamiliar words (at each level) and number of long words. With the stepwise multiple regression approach, regardless of the level at which unfamiliar words were identified or the removal of outlying passages, not all of the independent variables would enter the equation.

Because the objective for this portion of the study was to recalibrate the existing Homan-Hewitt formula with the Miller and Coleman (1967) passages, it was necessary for all three independent variables to enter the equation. It was also important for the independent variables to enter the equation in the order specified by Homan and Hewitt (2004; 1994) for the recalibrated formula to be as similar to the original formula as possible. Therefore, the initial stepwise multiple regression analysis results of the present study were inspected and several hierarchical multiple regression analyses were conducted in an attempt to force the three independent variables into the equation in order in which Homan and Hewitt specified while retaining acceptable significance levels.

In Homan et al. (1994) and Hewitt and Homan's (2004) validation studies, number of difficult words entered the equation first, followed by sentence complexity, and then word length. The previously conducted analyses for the Homan-Hewitt recalibration in the current study were inspected with special attention to the order of entry for the independent variables and significance levels. It was determined that when unfamiliar words were identified at level 4 and all passages were included, number of difficult words accounted for the most variance in total CT scores, followed by sentence

complexity and word length. Only word length was prevented from entering the equation (p = .413). In addition, when unfamiliar words were identified at level 6, regardless of passages removed, number of difficult words accounted for the most variance in total CT scores, followed by sentence complexity and word length. When the four cases with the highest total CT scores were removed, sentence complexity (p = .249) and word length (p = .299) did not enter the equation. When outlying passage 5 was also removed, sentence complexity (p = .147) and word length (p = .222) did not enter the equation. Regardless of significance values, both of these sets of analysis followed the order of entry indicated by Homan and Hewitt. Therefore, exploratory hierarchical multiple regression techniques were used with unfamiliar words identified at levels 4 (all passages included) and 6 (four passages with highest total CT score and passage 5 removed) to determine how all three independent variables could be forced into the equation in the order in which Homan and Hewitt indicated they should enter. Those analyses are described in detail below.

Hierarchical multiple regression was used to determine the variance accounted for in total CT scores by number of difficult words (level 4), sentence complexity, and word length. The independent variables were entered in the order indicated by Homan and Hewitt (2004, 1994). The full model accounted for 75.7% of variance in total CT scores, (R^2 = .757, $F_{(3,32)}$ = 33.290, p < .0005; see Table 22). Number of difficult words (level 4), which was entered first, explained 70.5% of variance total CT scores (p < .0005). Sentence complexity explained an additional 4.7% of variance of total CT scores over and above that explained by number of difficult words (p < .017). Word length explained an additional .5% of variance in total CT scores beyond the variance explained by the other two independent variables (p = .413).

Table 22

Hierarchical regression change statistics for Homan-Hewitt recalibration with unfamiliar words at level 4

Variables	R^2 change	F change	p change
WUNF (level 4)	.705	81.157	.0005
WNUM	.047	6.307	.017
WLON	.005	.687	.413

Note. WUNF = number of difficult words; WNUM = sentence complexity; WLON = word length.

Hierarchical multiple regression was then used with the number of difficult words (level 6), sentence complexity, and word length as the independent variables and total CT scores as the dependent variable. The independent variables were entered in the order indicated by Homan and Hewitt (2004, 1994). With the four passages with the highest total CT scores removed, the full model explained 84.1% of variance in total CT scores (R^2 = .841, $F_{(3,28)}$ = 49.315, p < .0005; see Table 23). Number of difficult words (level 6), which was entered first, explained 83.3% of variance total CT scores (p < .0005). Sentence complexity explained an additional .8% of variance of total CT scores over and above that explained by number of difficult words (p = .249). Word length did not explain any additional variance in total CT scores beyond the variance explained by the other two independent variables (p = .776).

Hierarchical multiple regression was then used with the same independent variables entered in the same order and the same dependent variable with outlying passage 5 also removed. With the five passages removed, the full model explained 86.3% of variance in

total CT scores (R^2 = .863, $F_{(3,27)}$ = 56.526, p < .0005; see Table 24). Number of difficult words (level 6), which was entered first, explained 85.1% of variance total CT scores (p < .0005). Sentence complexity explained an additional 1.1% of variance of total CT scores over and above that explained by number of difficult words (p = .147). Word length did not explain any additional variance in total CT scores beyond the variance explained by the other two independent variables (p = .808).

Table 23

Hierarchical regression change statistics for Homan-Hewitt recalibration with unfamiliar words at level 6

Variables	R^2 change	F change	p change
WUNF (level 6)	.833	149.392	.0005
WNUM	.008	1.385	.249
WLON	.000	.083	.776

Note. WUNF = number of difficult words; WNUM = sentence complexity; WLON = word length.

Table 24

Hierarchical regression change statistics for Homan-Hewitt recalibration with unfamiliar words at level 6 and passage 5 removed

Variables	R^2 change	F change	p change
WUNF (level 6)	.851	166.167	.0005
WNUM	.011	2.224	.147
WLON	.000	.060	.808

Note. WUNF = number of difficult words; WNUM = sentence complexity; WLON = word length.

Identifying the semantic variable at levels 4 and 6 did not allow all of three of the independent variables to enter the equation. The same hierarchical regression method was, therefore, conducted with difficult words identified at level 8. Number of difficult words (level 8), sentence complexity, and word length were the independent variables and total CT scores was the dependent variable. The independent variables were entered in the order indicated by Homan and Hewitt (2004, 1994). The four passages with the highest total CT scores were removed. The full model explained 82.9% of variance in total CT scores ($R^2 = .829$, $F_{(2,29)} = 45.240$, $p < .0005$; see Table 25). Number of difficult words (level 8), which was entered first, explained 74.5% of variance in total CT scores ($p < .0005$). Sentence complexity explained an additional 4.3% of variance of total CT scores over and above that explained by number of difficult words ($p < .021$). Word length explained an additional 4.1% of variance in total CT scores over and above that explained by number of difficult words, and sentence complexity ($p < .015$). See Tables 25 and 26 for full results.

Table 25

Hierarchical regression change statistics for Homan-Hewitt recalibration with unfamiliar words at level 8 and four passages removed

Variables	R^2 change	F change	p change
WUNF (level 8)	.745	87.417	.0005
WNUM	.043	5.922	.021
WLON	.041	6.737	.015

Note. Four passages with highest total CT scores removed. Difficult words identified at level 8. WUNF = number of difficult words; WNUM = sentence complexity; WLON = word length.

Table 26

Hierarchical regression results for Homan-Hewitt recalibration with unfamiliar words at level 8 and four passages removed

R^2	Adj R^2	F	β WUNF	β WNUM	β WLON	Regression Equation
.829	.811	45.24	-.531	.016	-.453	Y´=1120.253+(.547*WNUM)-(13.497*WUNF)-(27.048*WLON)

Note. Four passages with highest total CT scores removed. Difficult words identified at level 8. WUNF = number of difficult words; WNUM = sentence complexity; WLON = word length.

Hierarchical multiple regression was then conducted with outlying passage 5 also removed, the full model explained 86.3% of variance in total CT scores (R^2 = .863, $F_{(3,27)}$ = 56.925, p < .0005). Number of difficult words (level 8), which was entered first, explained 78% of variance in total CT scores (p < .0005). Sentence complexity explained an additional 5% of variance of total CT scores over and above that explained by number of difficult words (p < .008). Word length explained an additional 3.3% of variance in total CT scores over and above that explained by number of difficult words, and sentence complexity (p < .017). See Tables 27 and 28 for full results.

Based on the results of the stepwise and hierarchical regression analysis conducted for the recalibration of the Homan-Hewitt formula, one formula was selected for comparisons to the current model. The regression equation selected was that which

incorporated the identification of unfamiliar words at level 8 and derived via hierarchical multiple regression with the passages with the highest total CT scores and outlying passage 5 removed.The use of level 8 for the semantic variable, rather than level 4, resulted in a slight deviation from the original Homan-Hewitt formula. Although this departure from the original variables in the existing formula was less than ideal, it was necessary to allow all of the variables to enter the equation.

Table 27

Hierarchical regression change statistics for Homan-Hewitt recalibration with unfamiliar words at level 8 and five passages removed

Variables	R^2 change	*F* change	*p* change
WUNF (level 8)	.780	103.110	.0005
WNUM	.050	8.304	.008
WLON	.033	6.484	.017

Note. Four passages with highest total CT scores and passage 5 removed. Difficult words identified at level 8. WUNF = number of difficult words; WNUM = sentence complexity; WLON = word length.

Table 28

Hierarchical regression results for Homan-Hewitt recalibration with unfamiliar words at level 8 and five passages removed

R^2	Adj R^2	*F*	β WUNF	β WNUM	β WLON	Regression Equation
.863	.848	56.925	-.568	-.026	-.407	Y′=1128.958-(.881*WNUM)-(14.081*WUNF)-(23.722*WLON)

Note. Four passages with highest total CT scores and passage 5 removed. Difficult words identified at level 8. WUNF = number of difficult words; WNUM = sentence complexity; WLON = word length.

Recalibrated formulas selected for retention.

Simple linear, stepwise, and hierarchical multiple regression techniques were used to recalibrate the existing Dale-Chall, FOG, and Homan-Hewitt readability formulas while maintaining the pre-existing predictor variables for each formula. Based on the results of these analyses, one recalibrated equation for the Dale-Chall, three recalibrated equations for the FOG, and one recalibrated equation for the Homan-Hewitt were identified for comparisons to the new model.

The recalibrated Dale-Chall formula was derived via stepwise multiple regression with the four passages with highest total CT scores and outlying passage 31 removed from the analysis. The Dale-Chall recalibrated regression equation accounted for 88.1% of variance in total CT scores with number of unfamiliar words and average sentence length as the independent variables (R^2 = .881, $F_{(2,28)}$ = 103.784, p < .0005; see Table 29). This regression equation accounted for more variance in the total CT scores than the original Dale-Chall formula accounted for in its dependent variable. Chall and Dale (1995) reported that their formula accounted for 80% of variance in text difficulty. When applied to the Miller and Coleman (1967) passages, the results of the original and recalibrated Dale-Chall formulas were significantly correlated: when all 36 passages were included, r = .937, p < .0005 and when only the 31 passages used for the recalibration were included, r = .961, p < .0005.

Table 29

Multiple regression results for selected recalibrated Dale-Chall

Recalibrated Dale-Chall Regression equation: Y´=1046.50-(8.849*UFW)-(4.984*SL).								
				b	*b*	β	β	
R^2	Adj R^2	*F*	*a*	UFW	SL	UFW	SL	*p*
.881	.873	103.784	1046.50	-8.849	4.984	-1.016	.164	.0005

*Note.*UFW = number of unfamiliar words; SL = average sentence length.

The first recalibrated FOG formula (FOG1) was derived via stepwise multiple regression. With all 36 passages included in the analysis, percentage of hard words and average sentence length accounted for 74% variance in total CT scores with the percentage of hard words and average sentence length as the independent variables (R^2 = .740, $F_{(2, 33)}$ = 46.895, p < .0005, see Table 30). When applied to the Miller and Coleman (1967) passages, the results of the original FOG and recalibrated FOG1 formulas were significantly correlated (r = -.982, p < .0005).

The second recalibrated FOG formula (FOG2) was derived via hierarchical multiple regression. The four passages with the highest total CT scores were removed. From the full model, percentage of hard words and average sentence length accounted for 83.3% of variance in total CT scores (R^2 = .833, $F_{(2,29)}$ = 72.226, p < .0005; see Table 31). Average sentence length was entered first in the equation and explained 17.6% of variance in total CT scores (p < .017). Percentage of hard words accounted for an additional 65.7% of variance in total CT scores over and above the variance accounted for by percentage of hard words (p = .0005). When applied to the Miller and Coleman (1967) passages, the

results of the original FOG and recalibrated FOG2 formulas were significantly correlated: when all 36 passages were included, r = -.904, p < .0005 and when only the 32 passages used for the recalibration were included, r = - .907, p < .0005.

Table 30

Stepwise regression results for selected recalibrated FOG1

Stepwise derived recalibrated FOG regression equation 1:

Y′=1277.463-(18.192*HW)-(8.446*SL)

				b	*b*			
R^2	Adj R^2	*F*	*a*	HW	SL	β	β	*p*
.740	.724	46.895	1277.463	8.849	4.984	-.699	-.259	.0005

Note. HW = percentage hard words, SL = average sentence length.

Table 31

Hierarchical regression results for selected recalibrated FOG2

Hierarchical derived recalibrated FOG regression equation 2:

Y′=1109.175-(18.193*HW)-(.412*SL)

				b	*b*	β	β	
R^2	Adj R^2	*F*	*a*	HW	SL	HW	SL	*p*
.833	.821	72.226	1109.175	-18.193	-.412	-.015	-.906	.0005

Note. HW = percentage hard words, SL = average sentence length.

For the third recalibrated FOG formula (FOG3), the independent variables (average sentence length and percentage of hard words) were summed and treated as a single

independent variable. The four passages with the highest total CT scores and outlying passage 5 were removed. Simple linear regression was conducted and the regression equation accounted for 73.2% of variance in total CT scores (R^2 = .732, $F_{(1,29)}$ = 79.164, p < .0005; see Table 32). When applied to the Miller and Coleman (1967) passages, the results of the original FOG and recalibrated FOG3 formulas were significantly correlated: when all 36 passages were included and when only the 31 passages used for the recalibration were included, r = -1.0, p < .0005.

Table 32

Simple linear regression results for selected recalibrated FOG3

Simple linear derived recalibrated FOG regression equation 3: Y′=1257.188-(11.469*(HW+ SL))						
R^2	Adj R^2	t	a	b	β	p
.727	.723	-8.897	1257.188	-11.469	-.856	.0005

Note. HW = percentage hard words, SL = average sentence length.

The recalibrated Homan-Hewitt formula was derived via hierarchical multiple regression with the four passages with highest total CT scores and outlying passage 5 removed from the analysis. The full model accounted for 86.3% of variance in total CT scores with number of difficult words (level 8), sentence complexity, and word length as the independent variables (R^2 = .863, $F_{(3,27)}$ = 56.925, p < .0005; see Table 33). Number of difficult words (level 8), which was entered first, explained 78% of variance in total CT scores (p < .0005). Sentence complexity explained an additional 5% of variance of total CT scores over and above that explained by number of difficult words (p < .008).

Word length explained an additional 3.3% of variance in total CT scores over and above that explained by number of difficult words and sentence complexity ($p < .017$). The recalibrated Homan-Hewitt regression equation accounted for greater variance in total CT scores than the original formula accounted for in its dependent variable. Homan and Hewitt (1994, 2004) reported that during their initial formula calibration, their equation accounted for 49.6% of variance in reading level established by passage sources.

The original Homan-Hewitt used a different level of the same semantic variable than the version that was recalibrated here. Specifically, the original version of the Homan-Hewitt identified difficult words at level 4, whereas the recalibrated version identified difficult words at level 8. Nonetheless, when applied to the Miller and Coleman (1967) passages, the results of the original and recalibrated Homan-Hewitt formulas were significantly correlated: when all 36 passages were included, $r = -.909$, $p < .0005$ and when only the 31 passages used for the recalibration were included, $r = -.902$, $p < .0005$.

Table 33

Hierarchical regression results for selected, recalibrated Homan-Hewitt

Recalibrated Homan-Hewitt Regression equation:

Y′=1128.958-(.881*WNUM)-(14.081*WUNF)-(23.722*WLON)

R^2	Adj R^2	*F*	*a*	*b* IV1	*b* IV2	*b* IV3	β IV1	β IV2	β IV3	*p*
.863	.848	56.93	1128.96	-14.08	-.881	-23.72	-.568	-.026	-.407	.0005

Note. IV1 = WUNF (number of unfamiliar words); IV2 = WNUM (t-unit length); IV3 = WLON (number of long words).

Table 34 includes the recalibrated formulas selected for retention and further consideration during Phase III of the investigation. One recalibrated formula was retained for Dale-Chall formula, three recalibrated formulas were retained for the FOG formula, and one recalibrated formula was retained for the Homan-Hewitt formula.

Table 34

Recalibrated formulas retained for further investigation

Formula name	Formula
Recalibrated Dale-Chall	Y′=1046.50-(8.849*UFW)-(4.984*SL)
Recalibrated FOG1	Y′=1277.463-(18.192*HW)-(8.446*SL)
Recalibrated FOG2	Y′=1109.175-(18.193*HW)-(.412*SL)
Recalibrated FOG3	Y′=1257.188-(11.469*(HW+SL))
Recalibrated Homan-Hewitt	Y′=1128.958-(.881*WNUM)-(14.081*WUNF)-(23.722*WLON)

*Note.*UFW = number of unfamiliar words; SL = average sentence length; HW = percentage hard words; 1SW = number on monosyllabic words; WNUM = T-unit length; WUNF = number of unfamiliar words; and WLON = number of long words.

Phase III: External Validity and Reliability Evidence

The purpose of this phase of the investigation was to collect and analyze external validity and reliability evidence for the new-model formulas by assessing how the new-model and recalibrated formulas performed and how their performance compared when they were applied to the examination items for a credentialing-program. To that end, all of the retained new-model and recalibrated formulas were applied to examination

materials related to a dental licensing program. Correlational analyses, Friedman two-way analysis of ranks tests, Sign tests, and regression analyses were conducted on the results of each formula for the credentialing-program examination materials.

The first set of subsections includes a description of the materials that were used in this phase of the investigation and how the samples were selected and then converted into pseudo-continuous prose. Then, a description is offered for the data collection procedures, comparisons that were made, expected consistencies and differences, and statistical methods that were used to analyze the results. The next set of subsections include the results of the statistical analyses conducted: correlational, dependent comparisons, and regression analyses. Figure 3 offers a graphic representation of the general organization of the Phase III component of the results section.

Materials Used to Collect Validity and Reliability Evidence

This subsection begins with a description of how the examination items were selected and the methods that were used to convert them to pseudo-continuous prose. Next, the readability estimates derived from each formula are outlined. Then the results of the correlational analyses are described and discussed. This is followed by a description and discussion of the results of the Friedman two-way analysis of ranks test and Sign tests used to compare the readability results. The investigation of expected systematic differences is explained last.

Stratified and systematic sampling was used to select examination items from each of the subject areas. Examination items were selected from the two 150-item components (i.e., Book 1 and Book 2) of the knowledge-based portion of the dentistry examination: 24 examination items from Book 1 and 24 from Book 2.

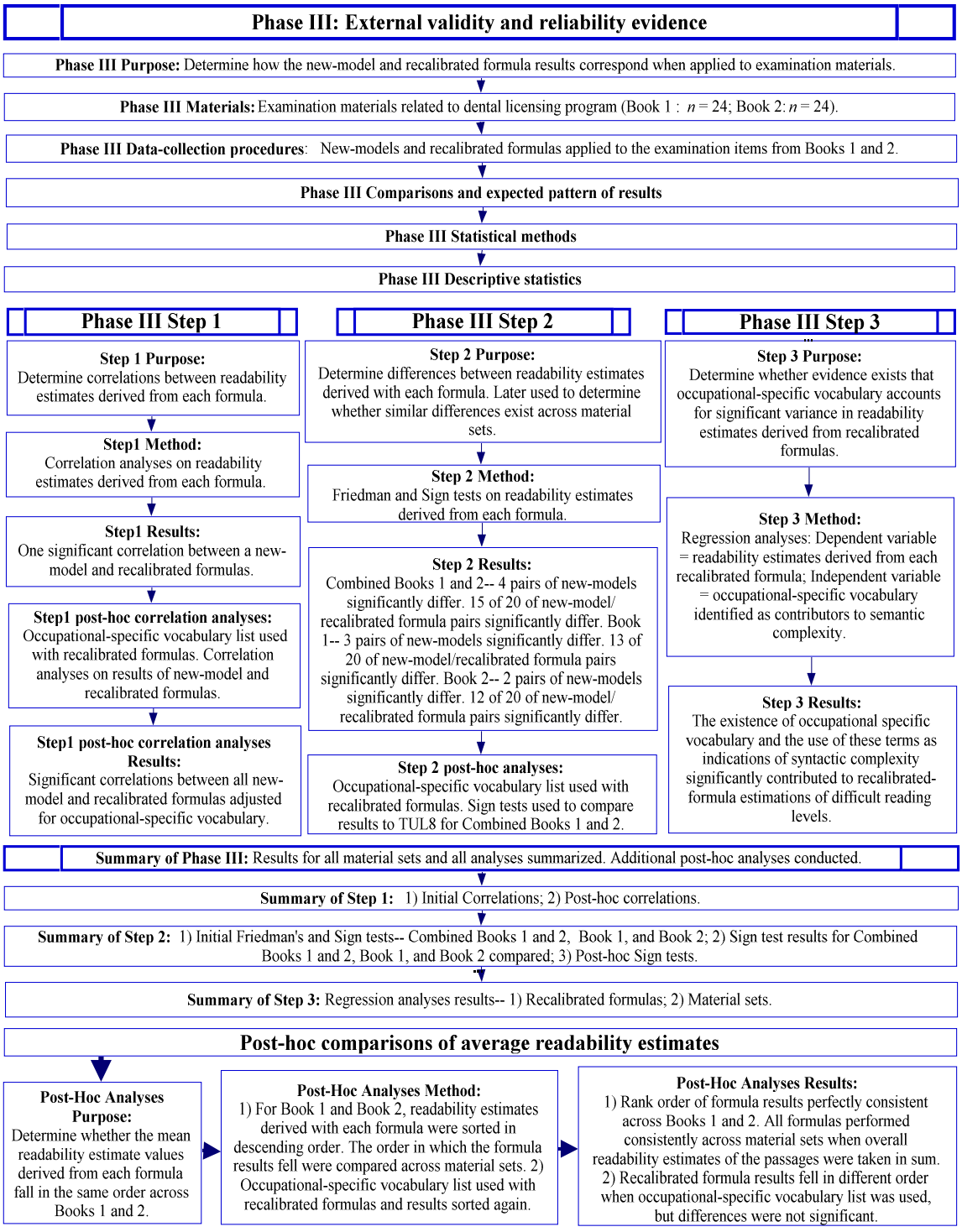


Figure 3. Graphic representation of Phase III results organization.

To select these sample items, each set of 150 items was sorted according to difficulty values and divided into 3 groups of items according to item difficulty (high, middle, low). The 50 items in each of the three groups were then resorted within their respective stratum according to their item identification codes. Staring at the first item in the difficulty stratum, every 6th item was identified for selection. This resulted in the selection of 8 items from each stratum (high, middle, low difficulty) for both Books 1 and 2 (see Table 35 for difficulty values of the selected items).

The 48 multiple-choice examination items were converted into pseudo-continuous prose with a method similar to that used by Plake (1988). Below are the guidelines that were followed to create pseudo-continuous prose from each examination item.

1) If the stem was an incomplete sentence and each of the options completed the sentence, the stem and each option were combined to create individual sentences.
2) If the stem was a complete sentence and the options were not complete sentences, the stem and options were combined to create individual sentences.
3) If the stem and each option were complete sentences, each was considered an individual sentence.
4) If an item included a scenario, the scenario was not combined with the stem or options. The scenario stood alone and each sentence in a scenario was counted once and measured along with the other components of the item.
5) If an item included instructions, such as those indicating that a reference image should be considered, the instructions were counted in the same way as scenarios. If a set of instructions applied to a group of items, the instructions were added to each question and added to their pseudo-continuous prose.

6) Where the stem included options and the options actually referred back to the choices in the stem, the elements were combined to create as many complete sentences as possible.

Table 35

Difficulty values for selected examination items

Book 1			Book 2		
High	Middle	Low	High	Middle	Low
0.581633	0.77551	0.928571	0.357143	0.77551	0.897959
0.602041	0.806122	0.928571	0.612245	0.785714	0.918367
0.632653	0.826531	0.938776	0.673469	0.785714	0.928571
0.642857	0.826531	0.94898	0.683673	0.795918	0.928571
0.642857	0.846939	0.969388	0.693878	0.806122	0.94898
0.683673	0.857143	0.969388	0.693878	0.826531	0.969388
0.72449	0.887755	0.979592	0.704082	0.867347	0.969388
0.734694	0.897959	0.989796	0.72449	0.867347	0.979592

Only one of the 48 selected items had fewer than four options. For the other 47 items, the methods devised for converting the items into pseudo-continuous prose yielded texts of at least four sentences each. The *Methods* section of this study includes examples of how the guidelines were used: for each guideline, a multiple-choice item obtained from websites related to certification and licensure and the pseudo-continuous prose that would be extracted for the respective items are offered.

For the Book 1 sample, 16 items required the method 1 conversion, 5 items required the method 2 conversion, 1 item required the method 3 conversion, and 2 items required the method 6 conversion. In addition, 2 items that required the method 1 conversion also required method 4 and the same was the case for 1 item that required method 6 and 1 item that required method 2. For the Book 2 sample, 13 items required the method 1 conversion, 7 items required the method 2 conversion, 3 items required the method 3 conversion, and 1 item required the method 4 conversion. In addition, 8 of the items that required the method 1 conversion also required method 4, 2 of the items that required the method 3 conversion also required method 4, and 1 item that required the method 3 conversion also required method 6. After the items were converted into pseudo-continuous prose, the mean number of words for items from Book 1 was 83.13 (*SD* = 44.523, range = 41 – 249), and the mean number of words for Book 2 was 93.96 (*SD* = 67.677, range = 44 – 378). An independent t-test revealed that the mean numbers of words were not different for Book 1 and Book 2 ($t_{(46)}$ = -.655, p = .516).

Data Collection Procedures

Variable measures for the Miller and Coleman (1967) passages were adjusted for exactly 150 words in the first phase of this study. Therefore, the same was done for the variable measures for the dental materials. For example, if a passage included 160 words and 7 T-units, the number of T-units was adjusted by dividing the actual number of T-units by the total number of words and multiplying that product by 150 [i.e., (7/160) * 150 = 7.466].

Because identifying T-units and clauses is not as straightforward and simplistic a task as determining the number of sentences in a passage, two raters independently identified

clauses and T-units for each set of passages. The T-unit and clause identification data were then analyzed to determine the inter-rater agreement. The initial T-unit identification agreement for the two sets of examination materials (r = 1.0; r = .989) were acceptable as were the initial clause identification agreement levels (r = .948; r = .988). Where discrepancies existed, the author of the study made the final decision.

The dental-material samples then were analyzed according to all of the syntactic and semantic variables included in the new-model and recalibrated formulas. Specifically, the syntactic analysis for each passage included determining 1) number of T-units; 2) T-unit length (i.e., average number of words per T-unit); 3) number of clauses; 4) clause length (i.e., average number of words per clause); and 5) sentence length (i.e., average number of words per sentence). The semantic analyses for each passage included determining 1) number of unfamiliar words at levels 8 and 10 (accordinig to *The Living Word Vocabulary: A National Vocabulary Inventory*, Dale & O'Rourke, 1981); 2) number of unfamiliar words (according to Chall & Dale word list, 1995); 3) percentage of words comprised of more than two syllables; and 4) number of words comprised of more than six letters. Then, additional tallies of unfamiliar word were created. Words that were identified as unfamiliar according to *The Living Word Vocabulary: A National Vocabulary Inventory* (at levels 8 and 10) but appeared in the occupational specific word list were counted. The new numbers of unfamiliar words included only words that did not appear in *The Living Word Vocabulary: A National Vocabulary Inventory* (at the respective grade levels) or the occupational specific word list. This resulted in two sets of semantic complexity measures: one that involved the use of only *The Living Word Vocabulary: A National Vocabulary Inventory* and one that also involved the use of the

occupational specific word list. To stay true to the nature of the existing formulas, the occupational-specific word list initially was not incorporated in the measures of the respective variables for each formula. However, post-hoc analyses were conducted that included in the occupational-specfic vocabulary list in the identification of semantic complexity with the use of the recalibrated formulas. The recalibrated Dale-Chall, FOG1, FOG2, FOG3, and Homan-Hewitt readability formulas, as well as the four new-model regression equations selected from the first phase of this investigation, were then applied to the examination item materials.

Comparisons and Expected Patterns of Results

Obtaining readability estimates for the materials according to the new model, recalibrated Dale-Chall (1995), recalibrated FOG, and recalibrated Homan-Hewitt readability formulas enabled comparison of the results of the existing formulas to the results of the new model. Comparisons included individual and overall averages of the estimated readability for the examination materials. Relationships among the estimated readabilities derived from each formula were investigated. Finding general consistency would offer some external validity and reliability evidence for the new model and result in some confidence that its use with examination items was supported. Systematic differences in the results determined according to the other readability formulas and the new model were also expected to support the validity and reliability of the new model.

Systematic differences in the estimates of the recalibrated formulas and new-model formula were expected. It was expected that the formulas that incorporate lists of familiar words (i.e., Dale-Chall, Homan-Hewitt) for measures of semantic complexity would yield readability estimates indicating more difficult passages than the new model because

occupational-specific dental terminology would be identified as unfamiliar in the existing formulas and would be considered familiar with the new model. More specifically, it was expected that divergence of the results of the new-model and recalibrated Dale-Chall and Homan-Hewitt formulas would be related to occurrences of occupational-specific terminology in the materials.

Systematic differences between the results of the new model and the recalibrated FOG formula were also anticipated. The FOG involves the use of number of syllables as a measure of semantic complexity. Specifically, it requires counting the number multisyllabic words in a sample. The dentistry occupational-specific terms tend to be comprised of many multisyllabic words; but those words should be considered familiar to the audience. Therefore, the greatest divergence between results of the new-model and those of the FOG was predicted to occur for samples that included large numbers of multisyllabic, occupational-specific terms. Specifically, it was expected that divergence of the results of the new-model and recalibrated FOG formulas would be related to occurrences of occupational-specific terminology in the materials.

Statistical Methods

Correlations between the predicted values derived with each formula were calculated to determine how the results of the formulas related. The results of the correlation analyses were then inspected in detail and are discussed in turn in the following sections. Post-hoc correlational analyses were conducted to determine how the new-model results would correlate with the results of the recalibrated formulas if occupational-specific vocabulary were not considered to contribute to semantic complexity with the use of the recalibrated formulas. More specifically, the recalibrated formulas were adjusted so that

occupational-specific vocabulary no longer contributed to increases in semantic complexity and correlational analyses were conducted for the results of the adjusted recalibrated formula and the new-model formulas.

To determine whether the formulas resulted in differential readability estimates, Freidman two-way analysis of ranks tests and Sign tests were used to compare the results for combined Books 1 and 2, Book 1, and Book 2. The results of the Friedman two-way analyses of ranks test and Sign tests were then inspected in detail and are discussed in turn in the following sections. The results of the analyses conducted within-material-set were used for informative purposes only. The within-material-set results were not considered to provide support of, or evidence against, the utility of each model. Instead, the results were meant to provide information about how the results of each formula corresponded. However, the results of dependent tests for each material set were compared across material sets later in the investigation to assess whether they were consistent for the different sets of examination items. Post-hoc Sign tests were conducted to determine how the new-model TUL8 results would compare with the results of the recalibrated formulas if occupational-specific vocabulary were not considered to contribute to semantic complexity with the use of the recalibrated formulas.

Regression techniques were used to determine whether differences among the results of the new-model and recalibrated formula readability estimates were related to the unfamiliar and multisyllabic occupational-specific terms in the passages. Specifically, the recalibrated formula results were regressed against the frequency of occupational-specific vocabulary that each respective model identified as contributors to syntactic complexity.

Readability Estimates

The four new readability model formulas and the recalibrated existing readability formulas were used to acquire readability estimates for each examination item and an average readability level across the compilation of examination items (Book 1 and 2 items) and Books 1 and 2 separately (see Table 36). For the new-model and recalibrated formulas, low mean readability values indicate harder-to-read text and high mean readability values indicate easier to read texts. The counterintuitive nature of these values is due to the nature of the cloze scores for the calibration passages. Specifically, the total cloze test scores tended to be higher for easier to read passages and lower for harder to read passages and these cloze scores served as the dependent variable upon which the formulas were calibrated or recalibrated.

Step I: relationships between formula results

The relationships between predicted values derived from each of the four new readability model formulas and five recalibrated existing formulas were analyzed. Three separate correlation analysis were conducted and are discussed in turn: 1) combined Books 1 and 2; 2) Book 1; and 3) Book 2. The combined Book 1 and 2 examination item correlation matrix shows that all four sets of results for the new-model formulas were significantly correlated with one another with a range of correlation values from $r = .915$ to $r = .986$ ($p < .01$). The results from the new-model TUL8 were significantly correlated with the results of the recalibrated FOG3 ($r = .244$, $p < .05$; see Table 37). No other recalibrated-formula results were significantly correlated with the results of the four new models.

Table 36

Descriptive statistics for all formulas

	Formula	Range	Mean	*SEM*	*SD*	Skewness	Kurtosis
Books 1 & 2	#TU8	583.46	882.56	19.16	132.76	-.919	.806
	TUL8	637.57	869.72	20.21	140.03	-.827	.776
	#C10	872.33	810.82	29.04	201.18	-.937	.627
	CL8	706.33	837.28	21.26	147.32	-.885	1.166
	DC	626.89	588.86	18.23	126.32	-.291	.193
	FOG1	631.87	659.37	25.6	177.93	-.232	-.968
	FOG2	636.80	616.91	25.05	173.54	-.405	-.770
	FOG3	500.63	771.70	18.31	126.84	-.199	-.776
	HH	2100.43	-688.71	61.57	426.55	-.455	.969
	Formula	Range	Mean	*SEM*	*SD*	Skewness	Kurtosis
Book 1	#TU8	583.46	866.08	31.36	153.62	-.794	.148
	TUL8	637.57	852.02	33.21	162.68	-.685	.273
	#C10	872.33	776.27	47.32	231.84	-.710	.047
	CL8	677.91	816.46	35.43	170.79	-.520	.164
	DC	626.89	546.88	28.39	139.07	.204	.452
	FOG1	539.89	626.94	35.27	172.81	-.232	-1.375
	FOG2	581.88	582.08	35.76	175.18	-.221	-.984
	FOG3	426.70	753.08	24.11	118.13	-.360	-1.030
	HH	2096.86	-794.29	89.04	436.19	-.422	-.442
	Formula	Range	Mean	*SEM*	*SD*	Skewness	Kurtosis
Book 2	#TU8	483.52	899.05	22.22	108.86	-.840	1.521

	Formula	Range	Mean	*SEM*	*SD*	Skewness	Kurtosis
	TUL8	459.75	887.42	23.23	113.79	-.760	.883
	#C10	661.30	845.36	33.21	162.69	-1.015	.995
	CL8	492.73	858.93	23.50	115.27	-1.060	1.943
	DC	429.22	630.84	19.99	97.91	-.495	.497
	FOG1	628.88	691.80	36.88	180.65	-.319	-.635
	FOG2	620.97	651.74	34.35	168.28	-.644	-.203
	FOG3	464.09	790.33	27.53	134.89	-.216	-.702
	HH	1476.98	-583.12	81.21	397.83	.053	-.012

Note. Combined Books 1 and 2 standard error for skewness = .343; standard error for kurtosis = .647. Individual Books 1 and 2 standard error for Skewness = .472; standard error for kurtosis = .918. SD = standard deviation. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt.

All of the recalibrated-formula results were significantly ($p < .01$) correlated with one another. The recalibrated Dale-Chall results were most strongly correlated with the results of the recalibrated Homan-Hewitt (FOG1: $r = .484$; FOG2: $r = .569$; FOG3: $r = .351$; Homan-Hewitt: $r = .710$). The results of the recalibrated FOG1 were most strongly correlated with the results of the recalibrated FOG2 (Dale-Chall: $r = .484$; FOG2: $r = .958$; FOG3: $r = .954$; Homan-Hewitt: $r = .718$). In turn, the results of the recalibrated FOG2 were most strongly correlated with the results of the FOG1 formula (Dale-Chall:

r = .569; FOG1: *r* = .958; FOG3: *r* = .829; Homan-Hewitt: *r* = .733). The results of the recalibrated FOG3 were also most strongly correlated with the results of the recalibrated FOG1 formula (Dale-Chall: *r* = .351; FOG1: *r* = .954; FOG2: *r* = .829; Homan-Hewitt: *r* = .596). The recalibrated Homan-Hewitt formulas were most strongly correlated with the results of the recalibrated FOG2 (Dale-Chall: *r* = .710; FOG1: *r* = .718; FOG2: *r* = .773; FOG3: *r* = .596).

Table 37

Combined Books 1 and 2—correlations between formulas

	TUL8	#C10	CL8	DC	FOG1	FOG2	FOG3	HH
#TU8	.986**	.939**	.969**	.029	.222	.189	.238	.168
TUL8	--	.915**	.965**	.024	.211	.161	.244*	.159
#C10	--	--	.956**	.074	.148	.187	.095	.165
CL8	--	--	--	.031	.194	.179	.192	.134
DC	--	--	--	--	.484**	.569**	.351**	.710**
FOG1	--	--	--	--	--	.958**	.954**	.718**
FOG2	--	--	--	--	--	--	.829**	.773**
FOG3	--	--	--	--	--	--	--	.596**

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

The combined Books 1 and 2 correlations between the four new-model formula results were inspected. The results of the two new-model formulas that involved measures of T-units (#TU8 and TUL8) were more strongly correlated with one another than they were with the results obtained with new-model formulas that involved measure of clauses (#C10 and CL8), but the correlation between #TU8 and TUL8 (r = .986) was only marginally stronger than the correlation between #TU8 and CL8 (r = .969). In addition, the results of the new-model formula that involved number of clauses (#CL10) were more strongly correlated with the results of the new-model formula that involved a measure of clause length (CL8) than they were with the results of the new-model formulas that involved measures of T-units. The results of the new-model formula that involved the measure of clause length (CL8) were marginally more strongly correlated with the results of the formula that involved the measure of T-unit length (TUL8; r = .965) than they were with the results of the new-model formula that involved number of clauses (#C10; r = .956).

The combined Books 1 and 2 correlations between the new-model and recalibrated-formula results were then inspected. The new models that included T-unit length (TUL8) and number of T-units (#TU8) were significantly correlated with the results of the recalibrated FOG3. There were no other significant correlations between the results of the new-models and recalibrated formulas.

The correlations between the results of the recalibrated formulas for combined Books 1 and 2 were then inspected. All of the recalibrated-formula results were significantly positively correlated with one other. Among the recalibrated formula results, no single recalibrated formula had results that better correlated with the results of all other

recalibrated formulas; although, the highest correlations between the recalibrated-formula results were among the three FOG formulas with a range of correlations values from r = .829 to r = .958. Of the three recalibrated FOG formula results, the FOG2 results showed the highest correlations with the other recalibrated formulas (not including FOG1 and FOG3) with a range of correlation values from r = .517 to r = .813 ($p < .01$). Of all recalibrated-formula results, the Homan-Hewitt had the strongest correlations with the Dale-Chall. The recalibrated Homan-Hewitt results were also more strongly correlated with the results of all recalibrated FOG results than were the results of the recalibrated Dale-Chall.

The relationships between predicted values derived from each of the four new readability model formulas and five recalibrated existing formulas were then analyzed for Book 1 (see Table 38). The Book 1 correlation matrix shows that all four sets of results for the new-model formulas were significantly correlated with one another with a range of correlation values from r = .970 to r =.991 ($p < .01$). Results from the recalibrated formulas were not significantly correlated with the results from the four new models.

The recalibrated Dale-Chall results were significantly correlated with the results of the recalibrated FOG1 (r = .369, $p < .05$), FOG2 (r = .470, $p < .05$), and Homan-Hewitt (r = .680, $p < .01$). The results of the recalibrated Dale-Chall were not significantly correlated with the results of the recalibrated FOG3. The results of the recalibrated FOG1 were significantly correlated with the results of all other recalibrated formulas (Dale-Chall: r = .369, $p < .05$; FOG2: r = .961, $p < .01$; FOG3: r = .951, $p < .01$; Homan-Hewitt: r = .705, $p < .01$). The results of the recalibrated FOG2 were also significantly correlated with the results of all other recalibrated formulas (Dale-Chall: r = .470,

$p < .05$; FOG3: $r = .828$, $p < .01$; Homan-Hewitt: $r = .813$, $p < .01$). The results of the recalibrated FOG3 formula were significantly correlated with the results of the other two recalibrated FOG formulas and recalibrated Homan-Hewitt ($r = .586$, $p < .01$), but were not significantly correlated with the results of the Dale-Chall. The recalibrated Homan-Hewitt formula results were significantly correlated with the results of all other recalibrated formulas with a range of correlation values from $r = .586$ to $r = .815$ ($p < .01$).

Table 38

Book 1—correlations between formulas

	TUL8	#C10	CL8	DC	FOG1	FOG2	FOG3	HH
#TU8	.991**	.973**	.970**	-.017	.145	.107	.174	.108
TUL8	--	.970**	.982**	-.022	.138	.097	.171	.101
#C10	--	--	.976**	-.006	.097	.100	.084	.093
CL8	--	--	--	-.039	.116	.078	.148	.056
DC	--	--	--	--	.369*	.470*	.221	.680**
FOG1	--	--	--	--	--	.961**	.951**	.709**
FOG2	--	--	--	--	--	--	.828**	.813**
FOG3	--	--	--	--	--	--	--	.586**

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH =

recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

The Book 1 correlations between the four new-model formula results were inspected. All of the correlations between the new-model results were very high and only marginal differences existed. With a correlation value of $r = .991$, the strongest correlation was between the two formulas that involved measures of T-units (#TU8 and TUL8). The weakest correlation value among the new-model results was $r = .970$. Two pairs of new-model results had correlations with this value: the formula that involved number of T-units and the formula that involved a measure of clause length (#TU8 and CL8) and the formula that involved a measure of T-unit length. The Book 1 correlations between the new-model and recalibrated-formula results were then inspected. None of the recalibrated formulas results had significant correlations with any of the new-model results.

The correlations between the results of the recalibrated formulas for Book 1 were then inspected. The recalibrated Dale-Chall had the only results that were not significantly correlated with the results of all other recalibrated formula. Specifically, the recalibrated Dale-Chall results were not correlated with those of the recalibrated FOG3. No single recalibrated formula had results that better correlated with the results of all other recalibrated formulas; although the highest correlations were among the three FOG formulas with a range of correlations values from $r = .828$ to $r = .961$. Of the three recalibrated FOG formula results, the FOG2 results showed the highest correlations with the other recalibrated formulas with a range of correlation values from $r = .470$ to $r = .813$ ($p < .05$).

The relationships between predicted values derived from each of the four new readability model formulas and six recalibrated existing formulas were then analyzed for Book 2 (see Table 39). The Book 2 correlation matrix shows that all four sets of results for the new-model formulas were significantly correlated with one another with a range of correlation values from r = .871 to r =.977 (p < .01). None of the recalibrated formula results were significantly correlated with those of the new models.

All of the recalibrated-formula results were significantly (p < .01) correlated with one another. The recalibrated Dale-Chall results were significantly correlated with the results of all other recalibrated formulas and were most strongly correlated with the results of the recalibrated Homan-Hewitt (FOG1: r = .594; FOG2: r = .667; FOG3: r = .473; Homan-Hewitt: r = .711).The results of the recalibrated FOG1 were significantly correlated with the results of all other recalibrated formulas and were most strongly correlated with the results of the recalibrated FOG3 (Dale-Chall: r = .594; FOG2: r = .954; FOG3: r = .958, Homan-Hewitt: r = .705). The results of the recalibrated FOG2 were also significantly correlated with the results of all other recalibrated formulas (Dale-Chall: r = .667; FOG3: r = .828; Homan-Hewitt: r = .704). The results of the recalibrated FOG3 were also significantly correlated with the results of all other recalibrated formulas (Dale-Chall: r = .473; FOG1: r = .958; FOG2: r = .828; Homan-Hewitt: r = .646).The recalibrated Homan-Hewitt formula results were also significantly correlated with the results of all other recalibrated formulas with a range of correlation values from r = .578 to r =.711 and were most strongly correlated with the results of the recalibrated Dale-Chall.

Table 39

Book 2—correlations between formula

	TUL8	#C10	CL8	DC	FOG1	FOG2	FOG3	HH
#TU8	.977**	.871**	.970**	-.008	.294	.262	.298	.198
TUL8	--	.800**	.931**	-.017	.277	.207	.319	.182
#C10	--	--	.913**	.061	.158	.248	.059	.184
CL8	--	--	--	.011	.248	.264	.211	.173
DC	--	--	--	--	.594**	.667**	.473**	.711**
FOG1	--	--	--	--	--	.954**	.958**	.705**
FOG2	--	--	--	--	--	--	.828**	.704**
FOG3	--	--	--	--	--	--	--	.646**

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

The Book 2 correlations between the new-model formula results were then inspected. The results of the two new-model formulas that involved measures of T-units (#TU8 and TUL8) were more strongly correlated with one another than they were with the results obtained with new-model formulas that involved measure of clauses (#C10 and CL8), but the correlation between #TU8 and TUL8 (r = .977) was only marginally stronger than the correlation between #TU8 and CL8 (r = .970). In addition, the results of the new-model

formula that involved number of clauses (#CL10) were more strongly correlated with the results of the new-model formula that involved a measure of clause length (CL8) than they were with the results of the new-model formulas that involved measures of T-units. The results of the new-model formula that involved the measure of clause length (CL8) were more strongly correlated with the results of the formula that involve the measure of T-unit length (TUL8; r = .931) than they were with the results of the new-model formula that involved number of clauses (#C10; r = .913).

The correlations between the results of the recalibrated formulas for Book 2 were then inspected. All of the recalibrated-formula results were significantly, positively correlated with one other. No single recalibrated formula had results that better correlated with the results of all other recalibrated formulas; although, the highest correlations between the recalibrated-formula results were among the three FOG formulas with a range of correlations values from r = .828 to r = .958. Of the three recalibrated FOG formula results, the FOG2 results showed the highest correlations with the other recalibrated formulas with a range of correlation values from r = .470 to r = .813 ($p < .05$). Of all recalibrated-formula results, those of the Homan-Hewitt had the strongest correlations with the Dale-Chall. The recalibrated Homan-Hewitt results were also more strongly correlated with the results of all recalibrated FOG results than were the results of the recalibrated Dale-Chall.

Post-hoc correlation analyses.

The previously discovered weak and non-significant correlations between the new-model and recalibrated-formula results were further investigated. The weak and non-significant correlations were assumed to be due to the fact that the recalibrated formulas

did not account for occupational-specific vocabulary in the identification of multisyllabic and unfamiliar words. Therefore, the correlations between the new-model results and those of the recalibrated formulas were reanalyzed for the examination materials, but the occupational-specific vocabulary words that were identified as unfamiliar or multisyllabic in the recalibrated formulas were removed from the totals. In other words, the results of the recalibrated formulas were adjusted to account for occupational-specific vocabulary that should be familiar to the respective audience of readers.

New correlations were calculated for combined Books 1 and 2 (N = 48; see Table 40). As expected, the correlations between all new-model and recalibrated formula results strengthened. The correlations between the recalibrated-formula results decreased, compared to the original correlation analysis, but some of the relationships remained significant.

Summary of correlational analyses of the examination materials.

Two conclusions may be reached based on the results of the initial and post-hoc correlational analyses conducted for the examination materials. First, there was a significant correlation between the results of the new-model TUL8 and recalibrated FOG3 formulas; but, no other significant correlations between the results of new-model and recalibrated formulas were observed. Second, when the results derived from the recalibrated formulas were adjusted to account for occupational specific vocabulary (i.e., to not identify the words as contributors to semantic complexity), significant correlations were observed between the results of all new-model and recalibrated formulas.

Table 40

Combined Books 1 and 2—correlations between formulas with occupational vocabulary considered

	TUL8	#C10	CL8	DC	FOG1	FOG2	FOG3	HH
#TU8	.986**	.939**	.969**	.335**	.636**	.642**	.500**	.694**
TUL8	--	.915**	.965**	.359**	.651**	.625**	.535**	.714**
#C10	--	--	.956**	.384**	.486**	.631**	.280*	.654**
CL8	--	--	--	.334*	.561**	.598**	.418**	.654**
DC	--	--	--	--	.097	.389**	-.135	.582**
FOG1	--	--	--	--	--	.833**	.915**	.514**
FOG2	--	--	--	--	--	--	.540**	.646**
FOG3	--	--	--	--	--	--	--	.312*

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

Step II: differences between formula results.

To determine whether the formulas resulted in differential readability estimates for the examination items, the Friedman two-way analysis of ranks test, which is a nonparametric version of repeated-measures analysis of variance, was used to compare the results. Dependent t-test were not used for the examination items because all four

new-model results were severely negatively skewed and therefore violated normality (see Table 36). Comparisons were made between all new-model formula results (#TU8, TUL8, #C10, and CL8). The results of each new model were then compared to the results of each recalibrated formula and the results of each recalibrated formula were compared to the results of the other recalibrated formulas. The three Friedman's test revealed that significant differences existed among the rankings of the examination items (Combined Books 1 and 2: $\chi^2_{(8)} = 263.03$, $p < .0005$; Book 1: $\chi^2_{(8)} = 128.68$, $p < .0005$; Book 2: $\chi^2_{(8)} = 134.52$, $p < .0005$). Table 41 shows the mean ranks chi square values for the data sets.

Table 41

Friedman test statistics

	Combined Books 1 and 2			Book 1			Book 2		
	Mean Rank	χ^2	*p*	Mean Rank	χ^2	*p*	Mean Rank	χ^2	*p*
#TU8	7.98	263.03	3.0E-52	8.00	128.68	5.3E-24	7.96	134.52	3.3E-25
TUL8	7.44			7.38			7.50		
#C10	5.60			5.54			5.67		
CL8	6.13			6.04			6.21		
DC	3.13			3.21			3.04		
FOG1	4.25			4.29			4.21		
FOG2	3.25			3.29			3.21		
FOG3	6.23			6.25			6.21		
HH	1.00			1.00			1.00		

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt.

The Sign test, which is a nonparametric version of dependent t-tests, was used to follow up the significant differences identified with the Friedman's test. The Wilcoxen Signed Ranks test is a more powerful follow-up test for significant differences observed with a Friedman's test, but the Wilcoxen Signed Ranks test is not robust to normality violations. Therefore, the less powerful Sign test was used. Because significant correlations existed between many of the results, a Bonferroni correction for familywise error was used to adjust alpha for each Sign test (36 comparisons total; α per comparison = .00143).

Table 42 shows the Book 1 and Book 2 combined results for the 36 comparisons. The Sign tests results revealed 27 of the 36 comparisons were significant ($p < .00143$). No significant differences existed between the results derived with new-models #TU8 and TUL8 or new-models #C10 and CL8. Significant differences existed between the results derived with new-models #TU8 and #C10 ($Z = -4.62$, $p < .00143$), #TU8 and CL8 ($Z = -5.48$, $p < .00143$), TUL8 and #C10 ($Z = -4.04$, $p < .00143$), and TUL8 and CL8 ($Z = -5.48$, $p < .00143$).

Table 42

Combined Books 1 and 2: Sign test statistics for 36 comparisons

Formula 1	Formula 2	Negative Difference	Positive Difference	*Z*	*p*
#TU8	TUL8	34	14	-2.89	0.0061
#TU8	#C10	40	8	-4.62*	7.7E-06
#TU8	CL8	43	5	-5.48*	9.3E-08
TUL8	#C10	38	10	-4.04*	9.7E-05
TUL8	CL8	43	5	-5.48*	9.3E-08
#C10	CL8	15	33	2.60	0.0142
#TU8	DC	46	2	-6.35*	5.4E-10
#TU8	FOG1	44	4	-5.77*	1.8E-08
#TU8	FOG2	45	3	-6.06*	3.3E-09
#TU8	FOG3	35	13	-3.18	0.0024
#TU8	HH	48	0	-6.93*	1.2E-11
TUL8	DC	46	2	-6.35*	5.4E-10
TUL8	FOG1	42	6	-5.20*	4.4E-07
TUL8	FOG2	44	4	-5.77*	1.8E-08
TUL8	FOG3	34	14	-2.89	0.0061
TUL8	HH	48	0	-6.93*	1.2E-11
#C10	DC	39	9	-4.33*	2.8E-05
#C10	FOG1	34	14	-2.89	0.0061
#C10	FOG2	41	7	-4.91*	1.9E-06

Formula 1	Formula 2	Negative Difference	Positive Difference	*Z*	*p*
#C10	FOG3	26	22	-0.58	0.6650
#C10	HH	48	0	-6.93*	1.2E-11
CL8	DC	44	4	-5.77*	1.8E-08
CL8	FOG1	39	9	-4.33*	2.8E-05
CL8	FOG2	43	5	-5.48*	9.3E-08
CL8	FOG3	29	19	-1.44	0.1939
CL8	HH	48	0	-6.93*	1.2E-11
DC	FOG1	14	34	2.89	0.0061
DC	FOG2	19	29	1.44	0.1939
DC	FOG3	4	44	5.77*	1.8E-08
DC	HH	48	0	-6.93*	1.2E-11
FOG1	FOG2	40	8	-4.62*	7.7E-06
FOG1	FOG3	1	47	6.64*	8.3E-11
FOG1	HH	48	0	-6.93*	1.2E-11
FOG2	FOG3	4	44	5.77*	1.8E-08
FOG2	HH	48	0	-6.93*	1.2E-11
FOG3	HH	48	0	-6.93*	1.2E-11

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH =

recalibrated Homan-Hewitt.. **Z value significant at .00143 level. Where necessary, significance values are reported in exponential format.

Significant differences were not found between formula pairs #TU8 and recalibrated FOG3, TUL8 and recalibrated FOG3, #C10 and recalibrated FOG1, #C10 and recalibrated FOG3, CL8 and recalibrated FOG3, recalibrated Dale-Chall and FOG1, or recalibrated Dale-Chall and FOG2. The Sign tests results were significant for all other formula pairings.

Failure to find differences between the results of the above formula pairs might be thought to be a product of the extremely conservative *p* criterion (α per comparison = . 00143) established with the Bonferroni strategy that was implemented to control for familywise error. Inspection of the significance values for these pairs showed that if a criterion of $p = .01$ had been used, the results of new-models #TU8 and TUL8 would have significantly differed and if a criterion of $p = .05$ would have been implemented; the results of new-models #C10 and CL8 would have significantly differed. In addition, if a criterion of $p = .01$ were used, the differences between the results of the following formula pairs would have significantly differed: #C10 compared to recalibrated FOG1; and recalibrated Dale-Chall compared to recalibrated FOG1. In contrast, if a criterion of $p = .05$ would have been implemented, the results of the comparisons for the following formula pairs would have remained non-significant: #C10 compared to recalibrated FOG3; CL8 compared to recalibrated FOG3; and recalibrated Dale-Chall compared to recalibrated FOG2.

Inspection of mean rankings and positive differences for the pair-wise comparisons that resulted in significant differences revealed a pattern of results among the formulas. The recalibrated Homan-Hewitt formula consistently resulted in readability estimates indicating significantly greater reading difficulty levels for the examination materials than the other formulas included in the analysis (lower readability estimation values indicate greater reading difficulty and greater readability estimation values indicate less reading difficulty). The four new-model formulas resulted in readability estimates that indicated significantly lower reading difficulty (easier-to-read text) than nearly all other formulas included in the analysis. Exceptions were found for the following formula pairs, which did not result in significant results: #TU8 and recalibrated FOG3, TUL8 and recalibrated FOG3, #C10 and recalibrated FOG1, #C10 and recalibrated FOG3, and CL8 and recalibrated FOG3. Inspection of these results shows that new-model formula #C10 results tended to correspond with the results of a greater number of recalibrated formulas (FOG1 and FOG3) than did the other new-model formula results. In addition, the recalibrated FOG3 was the only recalibrated formula with results that did not significantly differ from the results of any of the new-model formula results.

The recalibrated Dale-Chall formula readability estimation rankings were not significantly different from those of the FOG1 or FOG2 formula readability estimate rankings. According to the mean ranks established with the Friedman's test and the positive differences revealed with the Sign tests of the recalibrated formulas the FOG3 returned readability estimates indicating the lowest reading difficulty for the examination materials (mean rank = 6.23) and those results were significantly different from all other recalibrated formula results.

Not only did the recalibrated Homan-Hewitt return readability estimates indicating the most difficult reading levels for the examination items, the predicted values for 45 of the 48 pseudo-continuous prose were negative. There are two primary, interrelated reasons for these estimates of great difficulty and even negative estimated readability values. First, the Homan-Hewitt is the only formula that includes three variables (average T-unit length, number of words with seven or more letters, and number of unfamiliar words). Second, the number of words with seven or more letters was markedly lower for the passages upon which the formula was calibrated (Miller and Coleman, 1967; M = 9.40, SD = 3.18) than the dental examination items or pseudo-continuous prose (M = 56.86, SD = 13.86). The same was true for the average number of unfamiliar words (calibration passage M = 19.87, SD = 4.64; examination items M = 32.32, SD = 13.06).

Table 43 shows the Book 1 results for the 36 comparisons. Because significant correlations existed between many of the results, a Bonferroni correction for familywise error was used to adjust alpha for each Sign test (36 comparisons total; α per comparison = . 00143). The sign test results revealed 24 of the 36 comparisons were significant (p < .00143). Many of the comparison results were similar to those found for Books 1 and 2 combined. No significant differences existed between the results derived with #TU8 and TUL8, TUL8 and #C10, or #C10 and CL8. Significant differences existed between the results derived with #TU8 and C10 (Z = -3.67, p < . 00143), #TU8 and CL8 (Z = -4.08, p < . 00143), and TUL8 and CL8 (Z = -3.67, p < . 00143).

Table 43

Book 1: Sign test statistics for 36 comparisons

Formula 1	Formula 2	Negative Difference	Positive Difference	*Z*	*P*
#TU8	TUL8	17	7	-2.04	0.0639
#TU8	#C10	21	3	-3.67*	0.0003
#TU8	CL8	22	2	-4.08*	3.6E-05
TUL8	#C10	19	5	-2.86	0.0066
TUL8	CL8	21	3	-3.67*	0.0003
#C10	CL8	8	16	1.63	0.1516
#TU8	DC	22	2	-4.08*	3.6E-05
#TU8	FOG1	22	2	-4.08**	3.6E-05
#TU8	FOG2	22	2	-4.08*	3.6E-05
#TU8	FOG3	18	6	-2.45	0.0227
#TU8	HH	24	0	-4.90*	1.2E-07
TUL8	DC	22	2	-4.08*	3.6E-05
TUL8	FOG1	21	3	-3.67*	0.0003
TUL8	FOG2	22	2	-4.08*	3.6E-05
TUL8	FOG3	17	7	-2.04	0.0639
TUL8	HH	24	0	-4.90*	1.2E-07
#C10	DC	19	5	-2.86	0.0066
#C10	FOG1	17	7	-2.04	0.0639
#C10	FOG2	21	3	-3.67*	0.0003

Formula 1	Formula 2	Negative Difference	Positive Difference	*Z*	*P*
#C10	FOG3	12	12	0.00	1
#C10	HH	24	0	-4.90*	1.2E-07
CL8	DC	21	3	-3.67*	0.0003
CL8	FOG1	20	4	-3.27	0.0015
CL8	FOG2	21	3	-3.67*	0.0003
CL8	FOG3	14	10	-0.82	0.5413
CL8	HH	24	0	-4.90*	1.2E-07
DC	FOG1	6	18	2.45	0.0227
DC	FOG2	8	16	1.63	0.1516
DC	FOG3	3	21	3.67*	0.0003
DC	HH	24	0	-4.90*	1.2E-07
FOG1	FOG2	21	3	-3.67*	0.000277
FOG1	FOG3	0	24	4.90**	1.2E-07
FOG1	HH	24	0	-4.90*	1.2E-07
FOG2	FOG3	2	22	4.08*	3.6E-05
FOG2	HH	24	0	-4.90*	1.2E-07
FOG3	HH	24	0	-4.90*	1.2E-07

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH =

recalibrated Homan-Hewitt. * Z value significant at .00143 level. Where necessary, significance values are reported in exponential format.

Similar to the results of the comparisons made that included both books, the Book 1 results showed no significant differences between formula pairs #TU8 and recalibrated FOG3, TUL8 and recalibrated FOG3, #C10 and recalibrated FOG1, #C10 and recalibrated FOG3, CL8 and recalibrated FOG3, recalibrated Dale-Chall and FOG1, and recalibrated Dale-Chall and FOG2. When Book 1 examination materials were analyzed, two other formula combinations showed no significant results: 1) #C10 and recalibrated Dale-Chall and 2) CL8 and recalibrated FOG1. The Sign tests results were significant for all other Book 1 formula pairings.

Inspection of mean rankings and positive differences for the pair-wise comparisons that resulted in significant differences revealed a pattern of results among the formulas. The recalibrated Homan-Hewitt formula consistently resulted in readability estimates indicating significantly greater readability level for the examination materials than the other formulas included in the analysis (lower readability estimation values indicate greater reading difficulty and higher readability estimation values indicate less reading difficulty). The four new-model formulas resulted in readability estimates indicating significantly lower reading difficulty than nearly all other formulas included in the analysis. Exceptions were found for the following formula pairs, which did not result in significant results: #TU8 and recalibrated FOG3, TUL8 and recalibrated FOG3, #C10 and recalibrated Dale-Chall, #C10 and recalibrated FOG1, #C10 and recalibrated FOG3, CL8 and recalibrated FOG1, and CL8 and recalibrated FOG3. Inspection of these results

shows the same pattern found with the analysis that included both Books 1 and 2: formula #C10 results tended to correspond with the results of a greater number of recalibrated formulas (Dale-Chall, FOG1, and FOG3) than did the other new-model formula results. Another similar pattern was found in that the recalibrated FOG3 formula did not significantly differ from the results of any of the new-model formula results.

The Sign test recalibrated Dale-Chall results for Book 1 were identical to those found with both Books 1 and 2 combined: the recalibrated Dale-Chall formula readability estimation rankings were not significantly different from those of the recalibrated FOG1 or FOG2 formula readability estimate rankings. According to the mean ranks established with the Friedman's test and the positive differences revealed with the Sign tests, the FOG3 resulted in readability estimates indicating a lower difficulty level (mean rank = 6.25) for the materials than any of the other recalibrated formulas. The recalibrated FOG3 results were significantly different from the results of all other recalibrated formulas.

Not only did the recalibrated Homan-Hewitt result in readability estimates indicating the most difficult level for the examination materials, the predicted values for 23 of the 24 pseudo-continuous prose items were negative. There are two primary, interrelated reasons for these estimates of great difficulty and even negative estimated readability values. First, the Homan-Hewitt is the only formula that includes measures of three variables (average T-unit length, number of words with seven or more letters, and number of unfamiliar words). Second, the number of words with seven or more letters was markedly lower for the passages upon which the formula was calibrated (Miller and Coleman, 1967; M = 9.40, SD = 3.18) than the pseudo-continuous prose or examination items (M = 58.99, SD = 13.31). The same was true for the average number of unfamiliar

words (calibration passage M = 19.87, SD = 4.64; examination items M = 36.25, SD = 14.80).

Table 44 shows the Book 2 results for the 36 comparisons with the initial and adjusted significance values. Because significant correlations existed between many of the formula results, a Bonferroni correction for familywise error was used to adjust alpha for each Sign test (36 comparisons total; α per comparison = . 00143).

The sign test results revealed 21 of the 36 comparisons were significant (p < . 00143). Many of the comparison results were similar to those found for Books 1 and 2 combined and Book 1. No significant differences existed between the results derived with #TU8 and TUL8, #TU8 and C10, TUL8 and #C10, or #C10 and CL8. In contrast, significant differences existed between the results derived with #TU8 and CL8 (Z = -3.67, p < . 00143) and TUL8 and CL8 (Z = -4.08, p < . 00143).

Similar to the results of the comparisons made that included both books combined and Book 1 individually, Book 2 results showed no significant differences between formula pairs #TU8 and recalibrated FOG3, TUL8 and recalibrated FOG3, #C10 and recalibrated FOG1, #C10 and recalibrated FOG3, CL8 and recalibrated FOG3, recalibrated Dale-Chall and FOG1, and recalibrated Dale-Chall and FOG2. When Book 2 examination materials were analyzed, several other formula combinations showed no significant differences: #C10 and recalibrated Dale-Chall, #C10 and recalibrated FOG2, CL8 and recalibrated FOG1, and recalibrated FOG1 and FOG2. The Sign tests results were significant for all other Book 2 formula pairings.

Table 44

Book 2: Sign test statistics for 36 comparisons

Formula 1	Formula 2	Negative Difference	Positive Difference	*Z*	*P*
#TU8	TUL8	17	7	-2.04	0.0639
#TU8	#C10	19	5	-2.86	0.0066
#TU8	CL8	21	3	-3.67*	0.0003
TUL8	#C10	19	5	-2.86	0.0066
TUL8	CL8	22	2	-4.08*	3.6E-05
#C10	CL8	7	17	2.04	0.0639
#TU8	DC	24	0	-4.90*	1.2E-07
#TU8	FOG1	22	2	-4.08*	3.6E-05
#TU8	FOG2	23	1	-4.49*	2.9E-06
#TU8	FOG3	17	7	-2.04	0.0639
#TU8	HH	24	0	-4.90*	1.2E-07
TUL8	DC	24	0	-4.90*	1.2E-07
TUL8	FOG1	21	3	-3.67*	0.0003
TUL8	FOG2	22	2	-4.08*	3.6E-05
TUL8	FOG3	17	7	-2.04	0.0639
TUL8	HH	24	0	-4.90*	1.2E-07
#C10	DC	20	4	-3.27	0.0015
#C10	FOG1	17	7	-2.04	0.0639
#C10	FOG2	20	4	-3.27	0.0015

Formula 1	Formula 2	Negative Difference	Positive Difference	*Z*	*P*
#C10	FOG3	14	10	-0.82	0.5413
#C10	HH	24	0	-4.90*	1.2E-07
CL8	DC	23	1	-4.49*	2.9E-06
CL8	FOG1	19	5	-2.86	0.0066
CL8	FOG2	22	2	-4.08*	3.6E-05
CL8	FOG3	15	9	-1.22	0.3075
CL8	HH	24	0	-4.90*	1.2E-07
DC	FOG1	8	16	1.63	0.1516
DC	FOG2	11	13	0.41	0.8388
DC	FOG3	1	23	4.49*	2.9E-06
DC	HH	24	0	-4.90*	1.2E-07
FOG1	FOG2	19	5	-2.86	0.0066
FOG1	FOG3	1	23	4.49*	2.9E-06
FOG1	HH	24	0	-4.90*	1.2E-07
FOG2	FOG3	2	22	4.08*	3.6E-05
FOG2	HH	24	0	-4.90*	1.2E-07
FOG3	HH	24	0	-4.90*	1.2E-07

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH =

recalibrated Homan-Hewitt. * Z value significant at .00143 level. Where necessary, significance values are reported in exponential format.

Inspection of mean rankings and positive differences for the pair-wise comparisons that resulted in significant differences revealed a pattern of results among the formulas. The recalibrated Homan-Hewitt formula consistently resulted in readability estimates indicating significantly greater difficulty levels (harder-to-read text) for examination materials than the other formulas included in the analysis. The four new-model formulas resulted in readability estimates indicating significantly lower levels of reading difficulty (easier-to-read text) than nearly all other formulas included in the analysis. Exceptions were found for the following formula pairs, which did not result in significant results: #TU8 and recalibrated FOG3, TUL8 and recalibrated FOG3, #C10 and recalibrated Dale-Chall, #C10 and recalibrated FOG1, #C10 and recalibrated FOG2, #C10 and recalibrated FOG3, CL8 and recalibrated FOG1, and CL8 and recalibrated FOG3. Inspection of these results shows a pattern similar to that found with the analysis that included both Books 1 and 2 and Book 1 independently: new-model formula #C10 results tended to correspond with the results of a greater number of recalibrated formulas (Dale-Chall, FOG1, FOG2, and FOG3) than did the other new-model formula results. Another similar pattern was found in that the recalibrated FOG3 recalibrated formula did not significantly differ from the results of any new-model formula results.

The recalibrated Dale-Chall Sign test results for Book 2 were identical to the results found with both Books 1 and 2 combined and Book 1 independently: the recalibrated Dale-Chall formula readability estimation rankings were not significantly different from

those of the recalibrated FOG1 or FOG2 formula readability estimate rankings. According to the mean ranks established with the Friedman's test and the positive differences revealed with the Sign tests, the FOG3 resulted in readability estimates indicating a lower difficulty level (mean rank = 6.21) for the materials than any of the other recalibrated formulas..

Not only did the recalibrated Homan-Hewitt result in readability estimates indicating the most difficult reading levels for the examination materials, the predicted values for 23 of the 24 pseudo-continuous prose items were negative. There are two primary, interrelated reasons for these estimates of great difficulty and even negative estimated readability values. First, the Homan-Hewitt is the only formula that includes measures of three variables (average T-unit length, number of words with seven or more letters, and number of unfamiliar words). Second, the number of words with seven or more letters was markedly lower for the passages upon which the formula was calibrated (Miller and Coleman, 1967; *M* = 9.40, *SD* = 3.18) than the pseudo-continuous prose or examination items (*M* = 54.74, *SD* = 14.35). The same was true for the average number of unfamiliar words (calibration passage *M* = 19.87, *SD* = 4.64; examination items *M* = 28.39, *SD* = 9.88).

Post-hoc Sign tests of readability estimates: occupational-specific vocabulary list used with recalibrated formulas.

As explained in the Phase III, Step I portion of the results section, occupational-specific vocabulary was addressed differently by new-model and recalibrated formulas. The new-models incorporate the use of the occupational-specific vocabulary list and do not identify as unfamiliar any words from that list. The recalibrated formulas, however,

do not incorporate the use of the occupational-specific word list and, therefore, tend to identify occupational-specific vocabulary as a contributor to semantic-complexity.

To determine whether addressing occupational-specific vocabulary in the same manner would result in a different pattern of significant differences between new-model and recalibrated formulas, additional analyses were conducted. The occupational-specific vocabulary list was used with the recalibrated formulas. This resulted in adjustments to the totals for number of unfamiliar words (Dale-Chall and Homan-Hewitt), percentage of multisyllabic words (FOGs), and number of unfamiliar words plus number of long words (Homan-Hewitt). Specifically, any words that existed in the list of occupational-specific vocabulary were removed from the totals.

Table 45 offers a side-by-side comparison of the results of the recalibrated formulas and the recalibrated formulas with consideration of occupational-specific vocabulary for combined Books 1 and 2. Once occupational-specific vocabulary words were not considered unfamiliar, long, or multisyllabic, the readability estimate ranges for all recalibrated formulas narrowed and the standard deviations and standard error of the means decreased. In addition, the mean readability-estimate values for all of the formulas increased substantially when occupational-specific vocabulary words were considered. This increase indicated that when the occupational-specific vocabulary words were no longer identified as unfamiliar, long, or multisyllabic, the readability estimates indicated that passages were much easier to read. Sign tests were then conducted for combined Books 1 and 2 to compare the readability estimates derived with the new-model TUL8 and the recalibrated formulas. The TUL8 was the only model investigated here because it

has shown the most stable results across materials, as discussed in the summary of Phase III, Step I.

Table 45

Juxtaposition of combined Book 1 and Book 2 results for recalibrated formulas and recalibrated formulas with consideration of occupational-specific vocabulary words

Formula	Range	Mean	SEM	SD
DC	626.89	588.86	18.23	126.32
DC-O	332.28	996.70	12.86	89.08
HH	2100.43	-688.71	61.57	426.55
HH-O	1218.52	542.29	43.42	300.84
FOG1	631.87	659.37	25.6	177.93
FOG1-O	422.86	997.69	13.29	92.08
FOG2	636.80	616.91	25.05	173.54
FOG2-O	405.51	955.25	11.62	80.52
FOG3	500.63	771.70	18.31	126.84
FOG3-O	381.99	984.99	12.15	84.20

Note. DC = recalibrated Dale-Chall; DC-O = recalibrated Dale-Chall identifying occupational-specific vocabulary as familiar; HH = recalibrated Homan-Hewitt; HH-O = recalibrated Homan-Hewitt identifying occupational-specific vocabulary as familiar; FOG1 = stepwise derived recalibrated FOG; FOG1-O= stepwise derived recalibrated FOG identifying occupational-specific vocabulary as not multisyllabic; FOG2 = hierarchically derived recalibrated FOG; FOG2-O = hierarchically derived recalibrated FOG identifying occupational-specific vocabulary as not multisyllabic; FOG3 = simple derived recalibrated FOG; and FOG3-O simple derived recalibrated FOG identifying occupational-specific vocabulary as not multisyllabic.

Bonferroni adjustments were made to control for familywise error (five comparisons; α per comparison = .0102). The Sign test results for the examination materials revealed that significant differences still existed between the readability estimates of the TUL8 and recalibrated formulas (see Table 46). The direction of the differences, however, shifted for the recalibrated Dale-Chall, FOG1, and FOG2. Specifically, when the occupational-specific vocabulary list was not used with the recalibrated formulas, the recalibrated Dale-Chall, FOG1, and FOG2 formulas returned readability estimates reflecting significantly harder-to-read texts than the readability estimates derived with the TUL8.

Table 46

Sign test results: readability estimates of TUL8 compared to those of recalibrated formulas with use of occupational-specific vocabulary list

Formula 1	Formula 2	Negative Difference	Positive Difference	*Z*	*p*
TUL8	DC	6	42	-5.052*	4.4E-07
TUL8	FOG1	5	43	-5.340*	9.3E-08
TUL8	FOG2	10	38	-3.897*	9.7E-05
TUL8	FOG3	10	38	-3.897*	9.7E-05
TUL8	HH	43	5	-5.340*	9.3E-08

Note. Results of all recalibrated formulas derived with the use of the occupational-specific vocabulary list. TUL8 = new model incorporating T-unit length and unfamiliar words at level 8 DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt. *Z value significant at

.0102 level. Significance values are reported in exponential format. Determining whether differences were systematic

When the occupational-specific vocabulary list was used with the recalibrated formulas, the recalibrated Dale-Chall, FOG1, and FOG2 formulas returned readability estimates reflecting significantly easier-to-read texts than the readability estimates derived with the TUL8. The original Sign tests showed no significant difference between the readability estimates derived with the TUL8 and recalibrated FOG3. However, the new sign tests revealed that the readability estimates of the FOG3 formula resulted in readability estimates reflecting significantly easier-to-read texts than the readability estimates derived with the TUL8. When the occupational-specific vocabulary list was used with the recalibrated formulas, the readability estimates reflected significantly easier-to-read texts than the readability estimates derived with the TUL8.

Step III: determining whether differences were systematic.

The results were further examined according to the number of unfamiliar occupational-specific vocabulary terms in the passages as well as the number of multisyllabic occupational-specific vocabulary terms in the passages. Simple linear and stepwise multiple regression techniques were used to determine whether relationships existed between the results determined according to the formulas that require the use of lists of familiar words (i.e., recalibrated Dale-Chall and Homan-Hewitt) and the number of unfamiliar occupational-specific vocabulary terms that appear in the passages. Simple linear regression was used to determine whether relationships existed between the results to the formulas that required the identification of multisyllabic words (i.e., FOG1, FOG2,

and FOG3) and the number of multisyllabic occupational-specific vocabulary terms that appear in the passages. Nonparametric methods were not necessary for these analyses because the variables of interest did not violate normality.

It was expected that a significant amount of variance in the recalibrated Dale-Chall and Homan-Hewitt formula readability estimates would be accounted for by the respective numbers of unfamiliar words that were included in the occupational-specific vocabulary list. It was also expected that a significant amount of variance in recalibrated FOG1, FOG2, and FOG3 formula readability estimates would be accounted for by the number of occupational-specific vocabulary comprised of three or more syllables.

Dale-Chall regressions.

Three simple linear regressions were conducted for the Dale-Chall formula: Books 1 and 2 combined, Book 1 individually, and Book 2 individually. For each analysis, the estimated readability derived with the use of the recalibrated Dale-Chall formula was the dependent variable and the number of unfamiliar words that were included in the occupational-specific vocabulary list was the independent variable. This independent variable was calculated as the difference between the total number of unfamiliar words according to the Dale-Chall (1995) word list and the number of those unfamiliar words that were not included in the occupational-specific vocabulary list. The independent variable accounted for a significant amount of variance in the dependent variable for all sets of data (see Table 47): Books 1 and 2 combined (b = -6.393, $t_{(46)}$ = -8.132, R^2 = .590, p < .0005), Book1 individually (b = -7.378, $t_{(22)}$ = -6.390, R^2 = .650, p < .0005), and Book 2 individually (b = -4.669, $t_{(22)}$ = -4.594, R^2 = .490 , p < .0005).

Table 47

Simple linear regression results for recalibrated Dale-Chall formula number of unfamiliar words with consideration of occupational vocabulary

Data	R	R^2	Adj R^2	Standard Error of the Estimate	R^2 Change	F Change	df_1	df_2	Sig. F Change
Both Books	-.768	.590	.581	81.782	.590	66.125	1	46	.0005
Book 1	-.806	.650	.634	84.137	.650	40.838	1	22	.0005
Book 2	-.700	.490	.466	71.518	.490	21.108	1	22	.0005

The relationship between number of occupational-specific vocabulary initially identified as unfamiliar and readability-estimate values was negative (Book 1 and 2 combined: r = -.768; Book1: r = -.806; Book 2: r = -.700). This indicated that fewer instances of occupational-specific vocabulary terms that were initially identified as unfamiliar were related to higher readability-estimate values or easier-to-read text.

Homan-Hewitt regressions.

The Homan-Hewitt formula includes two semantic variables (number of difficult words and number of long words). Therefore, three linear regression analyses were conducted for the Homan-Hewitt formula on the three sets of data: Books 1 and 2 combined, Book 1 individually, and Book 2 individually. Simple linear regression was conducted for the first two analyses. The first analyses was conducted to investigate number of difficult words, and the second was conducted to investigate number of long

words. Stepwise multiple regression was used for the third analysis to investigate both semantic variables together.

Simple linear regression was used for the first analysis. For all three sets of data, the estimated readability derived with the use of the recalibrated Homan-Hewitt formula was the dependent variable and the number of unfamiliar words that were included in the occupational-specific vocabulary list was the independent variable. This independent variable was calculated as the difference between the total number of unfamiliar words and the number of those words that were not included in the occupational-specific vocabulary list but not *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981). The independent variable accounted for a significant amount of variance in the dependent variable for all sets of data (see Table 48):

Table 48

Simple linear regression results for recalibrated Homan-Hewitt formula number of unfamiliar words with consideration of occupational vocabulary

Data	R	R^2	Adj R^2	Standard Error of the Estimate	R^2 Change	F Change	df_1	df_2	Sig. F Change
Both Books	-.687	.471	.460	313.506	.471	41.005	1	46	.000
Book 1	-.766	.587	.569	286.462	.587	31.327	1	22	.000
Book 2	-.525	.275	.242	346.260	.275	8.362	1	22	.008

Books 1 and 2 combined (b = -26.787, $t_{(46)}$ = -6.403, R^2 = .471, p < .0005), Book1 individually (b = -28.213, $t_{(22)}$ = -5.597, R^2 = .587, p < .0005), and Book 2 individually (b = -22.521, $t_{(22)}$ = -2.892, R^2 = .275, p < .008). The relationship between number of occupational-specific vocabulary initially identified as unfamiliar and readability-estimate values was negative (Book 1 and 2 combined: r = -.687; Book1: r = -.766; Book 2: r = -.525). This indicated that fewer instances of occupational-specific vocabulary terms that were initially identified as unfamiliar were related to higher readability-estimate values or easier-to-read text.

Simple linear regression was also used for the second Homan-Hewitt analysis. For all three sets of data, the estimated readability derived with the use of the recalibrated Homan-Hewitt formula was the dependent variable and the number of words comprised of seven or more letters that were included in the occupational-specific vocabulary list was the independent variable. This independent variable was calculated as the difference between the number of words comprised of seven or more letters minus the number of those words that were not included in the occupational-specific vocabulary list. The independent variable accounted for a significant amount of variance in the dependent variable for all sets of data (see Table 49): Books 1 and 2 combined (b = -20.262, $t_{(46)}$ = -6.253, R^2 = .459, p < .0005), Book1 individually (b = -17.271, $t_{(22)}$ = -3.531, R^2 = .362, p < .002), and Book 2 individually (b = -22.781, $t_{(22)}$ = -5.658, R^2 = .593 , p < .0005). The relationship between number of occupational-specific vocabulary initially identified as long (seven or more letters) and readability-estimate values was negative (Book 1 and 2 combined: r = -.678; Book1: r = -.601; Book 2: r = -.770). This indicated

that fewer instances of occupational-specific vocabulary terms that were initially identified as long were related to higher readability-estimate values or easier-to-read text.

Table 49

Simple linear regression results for recalibrated Homan-Hewitt formula number of long words with consideration of occupational vocabulary

Data	R	R^2	Adj R^2	Standard Error of the Estimate	R^2 Change	F Change	df_1	df_2	Sig. F Change
Both Books	-.678	.459	.448	316.992	.459	39.102	1	46	.000
Book 1	-.601	.362	.333	356.326	.362	12.466	1	22	.002
Book 2	-.770	.593	.574	259.615	.593	32.010	1	22	.000

Stepwise multiple linear regression was used for the third Homan-Hewitt analysis. For all three sets of data, the estimated readability derived with the use of the recalibrated Homan-Hewitt formula was the dependent variable and the number of words comprised of seven or more letters that were included in the occupational-specific vocabulary list and the number of unfamiliar words that were included in the occupational-specific vocabulary list were the independent variables. When combined Books 1 and 2 were analyzed, both independent variables accounted for a significant amount of variance in the dependent variable ($R^2 = .603$, $F_{(2,45)} = 34.127$, $p < .0005$; see Table 50).

Table 50

Stepwise regression results for recalibrated Homan-Hewitt formula: Combined Books 1 and 2

Independent Variable	R	R^2	Adj R^2	Standard Error of the Estimate	R^2 Change	F Change	df_1	df_2	Sig. F Change
1	-.687	.471	.460	313.506	.471	41.005	1	46	.0005
2	-.776	.603	.585	274.784	.131	14.878	1	45	.0005

Note. Independent variable 1 = unfamiliar occupational-specific words; Independent variable 2 = number of occupational-specific words comprised of seven or more letters.

When Book 1 was analyzed independently, only the number of unfamiliar words that were included in the occupational-specific vocabulary list accounted for a statistically significant amount of variance in the dependent variable. These results were, therefore, the same as those for the simple linear regression analysis that included the number of words comprised of seven or more letters that were included in the occupational-specific vocabulary list (b = -28.213, $t_{(22)}$ = -5.597, R^2 = .587, p < .0005). The number of words comprised of seven or more letters that were included in the occupational-specific vocabulary list did not enter the equation (p = .071). When Book 2 was analyzed independently, only number of words comprised of seven or more letters accounted for a statistically significant amount of variance in the dependent variable. These results were, therefore, the same as those for the simple linear regression analysis that included number of words comprised of seven or more letters as the independent variable (b = -22.781, $t_{(22)}$

= -5.658, R^2 = .593, p < .0005). Number of unfamiliar words did not enter the equation (p = .517).

FOG regressions.

Simple linear regression was conducted for the FOG1, FOG2 and FOG3 formula results on all three sets of data: Books 1 and 2 combined, Book 1 individually, and Book 2 individually. For all three sets of data, the first FOG analysis included the estimated readability derived with the use of the recalibrated FOG1 formula as the dependent variable and the percentage of words comprised of three or more syllables (multisyllabic) that were included in the occupational-specific vocabulary list as the independent variable. This independent variable was calculated as the difference between the total number of multisyllabic words minus the number of those words that were not on the occupational-specific vocabulary list. The independent variable accounted for a significant amount of variance in the dependent variable for all sets of data (see Table 51): Books 1 and 2 combined (b = -16.793, $t_{(46)}$ = -11.362, R^2 = .737, p < .0005), Book1 individually (b = -16.249, $t_{(22)}$ = -7.627, R^2 = .852, p < .0005), and Book 2 individually (b = -17.357, $t_{(22)}$ = -7.749, R^2 = .732 , p < .0005). The relationship between percentage of occupational-specific vocabulary initially identified as multisyllabic and readability-estimate values was negative (Book 1 and 2 combined: r = -.859; Book1: r = -.852; Book 2: r = -.855). This indicated that fewer instances of occupational-specific vocabulary terms that were initially identified as multisyllabic were related to higher readability-estimate values or easier-to-read text.

Table 51

Simple linear regression results for recalibrated FOG1 formula percentage of multisyllabic words with consideration of occupational vocabulary

Data	R	R^2	Adj R^2	Standard Error of the Estimate	R^2 Change	F Change	df_1	df_2	Sig. F Change
Both Books	-.859	.737	.732	92.184	.737	129.094	1	46	.0005
Book 1	-.852	.726	.713	92.562	.726	58.166	1	22	.0005
Book 2	-.855	.732	.720	95.650	.732	60.044	1	22	.0005

For all three sets of data, the second FOG analysis included the estimated readability derived with the use of the recalibrated FOG2 formula as the dependent variable and the percentage of words comprised of three or more syllables (multisyllabic) that were included in the occupational-specific vocabulary list as the independent variable. The independent variable accounted for a significant amount of variance in the dependent variable for all sets of data (see Table 52): Books 1 and 2 combined (b = -16.943, $t_{(46)}$ = -13.114, R^2 = .789, p < .0005), Book1 individually (b = -17.269, $t_{(22)}$ = -9.305, R^2 = .797, p < .0005), and Book 2 individually (b = -16.500, $t_{(22)}$ = -8.396, R^2 = .762 , p < .0005). The relationship between percentage of occupational-specific vocabulary initially identified as multisyllabic and readability-estimate values was negative (Book 1 and 2 combined: r = -.888; Book1: r = -.893; Book 2: r = -.873). This indicated that

fewer instances of occupational-specific vocabulary terms that were initially identified as multisyllabic were related to higher readability-estimate values or easier-to-read text.

Table 52

Simple linear regression results for recalibrated FOG2 formula percentage of multisyllabic words with consideration of occupational vocabulary

Data	R	R^2	Adj R^2	Standard Error of the Estimate	R^2 Change	F Change	df_1	df_2	Sig. F Change
Both Books	-.888	.789	.784	80.57906	.789	171.986	1	46	.0005
Book 1	-.893	.797	.788	80.62573	.797	86.584	1	22	.0005
Book 2	-.873	.762	.751	83.91103	.762	70.501	1	22	.0005

For all three sets of data, the third FOG analysis included the estimated readability derived with the use of the recalibrated FOG3 formula as the dependent variable. Initially, the recalibrated FOG3 formula included one independent variable that was created by adding the average sentence length and the percentage of multisyllabic words. Therefore, for these analyses, the independent variable was created by calculating the percentage of words comprised of three or more syllables (multisyllabic) that were included in the occupational-specific vocabulary list and adding that value to sentence length. The independent variable accounted for a significant amount of variance in the dependent variable for all sets of data (see Table 53): Books 1 and 2 combined (b = -10.642, $t_{(46)}$ = -15.774, R^2 = .844, p < .0005), Book1 individually (b = -10.860,

$t_{(22)}$ = -10.120, R^2 = .823, p < .0005), and Book 2 individually (b = -10.463, $t_{(22)}$ = -11.362, R^2 = .854 , p < .0005). The percentage of occupational-specific vocabulary initially identified as multisyllabic combined with sentence length was negatively related to readability-estimate values (Book 1 and 2 combined: r = -.919; Book1: r = -.907; Book 2: r = -.924).This indicated that fewer instances of occupational-specific vocabulary terms that were initially identified as multisyllabic were related to higher readability-estimate values or easier-to-read text.

Table 53

Simple linear regression results for recalibrated FOG3 formula combined percentage of multisyllabic words and sentence length with consideration of occupational vocabulary

Data	R	R^2	Adj R^2	Standard Error of the Estimate	R^2 Change	F Change	df_1	df_2	Sig. F Change
Both Books	-.919	.844	.841	50.641	.844	248.833	1	46	.0005
Book 1	-.907	.823	.815	50.793	.823	102.418	1	22	.0005
Book 2	-.924	.854	.848	52.627	.854	129.100	1	22	.0005

Summary of regression results.

The results for the regression analyses indicated that for the recalibrated Dale-Chall, Homan-Hewitt, FOG1, FOG2, and FOG3 formulas, an extraordinary amount of variance in readability-estimates could be attributed to the frequency with which occupational-

specific vocabulary words occurred in the passages and were identified as unfamiliar, long, or multisyllabic. Specifically, as instances of occupational-specific vocabulary words increased, reading difficulty increased. These occupational-specific vocabulary terms, though, should not be considered unfamiliar or difficult to the respective audience. Therefore, these recalibrated formulas likely resulted in readability estimates that indicated unduly high difficulty levels.

Results from External Validity Analyses

This section includes a summary of Phase III results obtained from the external validity analysis. The first subsection includes a summary of the correlation analyses (Phase IIII, Step I) that were used to establish how the results of the formulas were related. It also includes a summary of the results for the post-hoc correlational analyses that were conducted during Phase III. The second subsection includes a results summary for the Sign tests (Phase III, Step II) that were used to determine whether the formulas resulted in significantly different readability estimates for the dental materials and the post-hoc Sign test results. The next subsection includes a results summary of the regression analyses (Phase III, Step III) that were used to determine whether the differences found in the formula results were systematic. The last subsection includes descriptions and results for additional post-hoc analyses that were conducted to compare mean readability levels derived with the formulas.

Results from Phase III, Step I: correlation analyses.

This subsection includes a summary of the correlation analyses conducted for the examination materials. It begins with a discussion of the initial correlational analyses that were conducted between the new-model and recalibrated formula results. Then, the

results of the post-hoc correlation analyses, which were conducted between the results of the recalibrated formulas with adjustments made for the existence of occupational-specific vocabulary and the new-models, are discussed.

Initial correlation analyses.

The initially conducted correlation analyses between the results of the new-model and recalibrated formulas revealed very weak relationships. When Book 1 and Book 2 examination materials were analyzed together, only a single significant relationship existed between a new-model and recalibrated formula: TUL8 and recalibrated FOG3 ($p < .05$). When Books 1 and 2 were analyzed independently, none of the new-model results were significantly correlated with the results of the any of the recalibrated formulas results.

Post-hoc correlational analyses.

Post-hoc correlation analyses were conducted to investigate the weak and non-significant correlations between the new-model and recalibrated formula results. The weak and non-significant correlations were assumed to have occurred because the recalibrated formulas considered occupational-specific vocabulary as multisyllabic or unfamiliar and, thereby, contributors to semantic complexity. Therefore, the correlations between the new-model results and those of the recalibrated Dale-Chall, FOG, and Homan-Hewitt formulas were reanalyzed, but the occupational-specific vocabulary words that were identified as unfamiliar or multisyllabic in the recalibrated formulas were removed from the totals. In other words, the results of the recalibrated formulas were adjusted to account for occupational specific vocabulary that should be familiar to the respective audience of readers.

New correlations were calculated for combined Books 1 and 2 (N = 48). Table 54 offers a juxtaposition of the results from the initial and post-hoc correlation results.

Table 54

Combined Books 1 & 2—juxtaposition of the correlations between results of initial and post-hoc correlation analyses of new-model and recalibrated formulas

Recalibrated Formulas	#TU8	TUL8	#C10	CL8
DC	.029	.024	.074	.031
Occupational DC	.335**	.359**	.384**	.334*
FOG1	.222	.211	.148	.194
Occupational FOG1	.636**	.651**	.486**	.561**
FOG2	.189	.161	.187	.179
Occupational FOG2	.642**	.625**	.631**	.598**
FOG3	.238	.244*	.095	.192
Occupational FOG3	.500**	.535**	.280*	.418**
HH	.168	.159	.165	.134
Occupational HH	.694**	.714**	.654**	.654**

Note #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

As expected, the correlations between the results of the new-model and all recalibrated formulas strengthened, as compared to the initial correlation results. When the occupational-specific vocabulary list was used with the recalibrated formulas, all relationships reached significance.

Taken together, the results of the initial correlation analyses for the new-model and recalibrated formulas and the post-hoc correlation analyses that involved adjusting the unfamiliar and monosyllabic numbers for the recalibrated formulas according to the occupational-specific vocabulary list indicate that of the four new-model formulas, the TUL8 formula appeared to offer slightly more stable results. Table 55 offers the frequency of significant correlations between the results of each new-model and two versions of the existing formulas. The data show that when the results of both versions of the existing formulas were considered in sum, a marginally greater number of significant relationships were observed for the TUL8. Data in this table also reveal that when the total number of significant relationships are compared across both versions of the existing formulas, the greatest number of significant relationships were observed between the results of the new-models and recalibrated formulas that were adjusted for the existence of occupational-specific vocabulary.

More specifically, for the initially conducted correlation analyses of the examination materials, only one significant relationship was observed between new-model and recalibrated-formula results: the TUL8 results were significantly correlated with the results of the recalibrated FOG3 ($r = .244$, $p < .05$). For the new correlation analyses conducted with attention to occupational-specific vocabulary in the recalibrated formulas (post-hoc correlation analyses), the TUL8 results were significantly correlated with the

results of the recalibrated Dale-Chall (r = .359, p < .01), FOG1 (r = .651, p < .01), FOG2 (r = .625, p < .01), FOG3 (r = .535, p < .01), and Homan-Hewitt formulas (r = .714, p < .01). Of the new-model formulas, the results of the TUL8 were strongest for the recalibrated FOG1, FOG3, and Homan-Hewitt.

Table 55

Combined Books 1 and 2: Frequency of significant correlations at p < .05 and p < .01 between the results of the new-models and all versions of existing formulas

	Recalibrated formulas		Occupational Recalibrated formulas		Totals p <.05	Totals p <.01	Grand totals of significant relationships
	p <.05	p <.01	p <.05	p <.01	Total p <.05	Total p <.01	Grand totals
#TU8	0	0	0	5	0	5	5
TUL8	1	0	0	5	1	5	6
#C10	0	0	1	4	1	4	5
CL8	0	0	1	4	1	4	5
Total	1	0	2	18			

Note. Original, existing formulas examined: Dale-Chall, FOG, Homan-Hewitt and revised (sign changed) Homan-Hewitt. Recalibrated and Occupational recalibrated formulas examined: Dale-Chall, FOG1, FOG2, FOG3, and Homan-Hewitt. Occupational recalibrated formula results are those for which totals were adjusted to remove instances of occupational-specific vocabulary. Cells in the columns labeled as *Totals for p < .05* and *Totals for p < .01* include the total number of significant relationships observed between each respective new-model formula and the three versions of the existing models. Cells in the column labeled

Grand totals of significant relationships include the number of significant relationships ($p < .05$ and $p < .01$) observed across all sets of analyses for each respective new-model formula. The cells in the bottom row, which is labeled *Total*, includes the total number of significant relationships observed between the results of the new-models and each version of the existing formulas across each material sets. Lrng = learning materials; Occ = occupational materials; and Exam = examination materials. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8.

Results from Phase III, Step II: comparisons of readability estimates.

This subsection includes a summary of the Friedman's two-way analyses of ranks tests and Sign test results. First, the results of the planned comparisons are described and compared across Books 1 and 2, Book 1, and Book 2. Then, the results of the post-hoc sign tests are outlined.

Summary of planned Friedman's and Sign test results.

The readability estimates of all models were compared. Specifically, Friedman's tests and Sign tests were used to compare the readability estimates of the examination materials for all models. For all three sets of data (combined Book 1 and 2, Book 1, and Book 2), Friedman's tests revealed significant differences between the results of the different readability models.

The Sign test results for combined Books 1 and 2 showed that there were significant differences between the results of 15 of the 20 new-model and recalibrated formula pairings and 4 of the 6 new-model comparisons. For Book 1, significant differences existed between 13 or the 20 new-model and recalibrated formula pairings and 3 of the 6

new-model comparisons. For Book 2, significant differences existed between 12 or the 20 new-model and recalibrated formula pairings and 2 of the 6 new-model comparisons. Table 56 allows a side-by-side comparison of where significant differences were observed for combined Books 1 and 2, Book 1. and Book 2.

The results were consistent across combined Books 1 and 2, Book 1, and Book 2 for four of the six comparisons that were made between the results of the new-models. Specifically, no significant differences were observed between the results of #TU8 and TUL8 or #C10 and CL8; whereas, significant differences were observed between the results of #TU8 and CL8 as well as TUL8 and CL8. When the results of the new-model and recalibrated formulas were compared, many of the results concurred across combined Books 1 and 2, Book 1, and Book 2. When the #TU8 and TUL8 were compared to the recalibrated formulas, the results were consistent across all material sets. Specifically, the results of #TU8 and TUL8 were significantly different from the results of the recalibrated Dale-Chall, FOG1, FOG2, and Homan-Hewitt, but were not significantly different from the results of the recalibrated FOG3. When the results of the #C10 and recalibrated formulas were compared, the results were consistent for 3 of the 5 comparisons. For all material sets, the results of #C10 were significantly different from the results of the recalibrated Homan-Hewitt, but were not significantly different from the results of recalibrated FOG1 or FOG3. When the results of the CL8 and recalibrated formulas were compared, the results were consistent for 4 of the 5 comparisons. For all material sets, the results of CL8 were significantly different from the results of the recalibrated Dale-Chall, FOG2, and Homan-Hewitt, but were not significantly different from the results of recalibrated FOG3.

Table 56

Significant differences between formula results according to Sign tests

		Significant differences between formula results for material sets		
Formula 1	Formula 2	Books 1 & 2	Book 1	Book 2
#TU8	TUL8			
#TU8	#C10	X	X	
#TU8	CL8	X	X	X
TUL8	#C10	X		
TUL8	CL8	X	X	X
#C10	CL8			
#TU8	DC	X	X	X
#TU8	FOG1	X	X	X
#TU8	FOG2	X	X	X
#TU8	FOG3			
#TU8	HH	X	X	X
TUL8	DC	X	X	X
TUL8	FOG1	X	X	X
TUL8	FOG2	X	X	X
TUL8	FOG3			
TUL8	HH	X	X	X
#C10	DC	X		
#C10	FOG1			
#C10	FOG2	X	X	
#C10	FOG3			

		Significant differences between formula results for material sets		
Formula 1	Formula 2	Books 1 & 2	Book 1	Book 2
#C10	HH	X	X	X
CL8	DC	X	X	X
CL8	FOG1	X		
CL8	FOG2	X	X	X
CL8	FOG3			
CL8	HH	X	X	X

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = Homan-Hewitt. X = significant difference between formula results according to sign test results. Significance identified at the .00143 level.

The comparisons of the Sign test results across Books revealed that of the new-model formulas the #TU8 and TUL8 performed most consistently. Specifically, the results of new-model and recalibrated formula comparisons were perfectly consistent across combined Books 1 and 2, Book 1, and Book 2. This may indicate that the #TU8 and TUL8 are more stable than the other two new-model formulas. Interpreting these results in light of the correlation results, it appears that if one new-model were to be chosen as the most stable of the four, it would be the TUL8. When the occupational-specific vocabulary list was used with the recalibrated formulas the results of the TUL8 were

more strongly correlated than the results of the #TU8 with the results of the recalibrated Dale-Chall, FOG1, FOG3, and Homan-Hewitt.

Summary of post-hoc Sign tests: occupational-specific vocabulary list used with recalibrated formulas.

To determine whether addressing occupational-specific vocabulary in the same manner would result in a different pattern of significant differences between new-model and recalibrated formulas, additional analyses were conducted. The occupational-specific vocabulary list was used with the recalibrated formulas and Sign tests were then conducted within materials to compare the readability estimates derived with the new-model TUL8 and the recalibrated formulas.

The Sign test results revealed that significant differences still existed between the readability estimates of the TUL8 and recalibrated formulas. However, the differences shifted direction for the recalibrated Dale-Chall, FOG1, and FOG2. When the occupational-specific vocabulary list was not used with the recalibrated formulas, the recalibrated Dale-Chall, FOG1, and FOG2 formulas resulted in readability estimates reflecting significantly harder-to-read texts than the readability estimates derived with the TUL8.

The original Sign tests showed no significant difference between the readability estimates derived with the TUL8 and recalibrated FOG3 for the examination materials. However, the new sign tests revealed that the readability estimates of the FOG3 formula resulted in readability estimates reflecting significantly easier-to-read texts than the readability estimates derived with the TUL8. The results of the recalibrated Homan-

Hewitt still reflected that the passages were harder to read than was indicated by the new-model TUL8.

Results from Phase III, Step III: regression analyses.

This subsection includes a summary of the Phase III results for the regression analyses that were conducted for each set of examination materials (i.e., combined Books 1 and 2, Book 1, and Book 2). The summary begins with an explanation of the purpose of the regression analyses. The overall findings across all formulas are summarized.

The results obtained for the examination materials were further examined according to the number of occupational-specific vocabulary terms in the passages as well as the number of multisyllabic, occupational-specific vocabulary terms in the passages. Simple linear and stepwise multiple regression techniques were used to determine whether relationships existed between the results determined according to the formulas that required the use of lists of familiar words (i.e., recalibrated Dale-Chall and Homan-Hewitt) and the number of unfamiliar occupational-specific vocabulary terms that appeared in the passages. Simple linear regression was used to investigate relationships between the results determined according to the recalibrated FOG2 formula, which requires the number of multisyllabic occupational-specific vocabulary terms. It was expected that a significant amount of variance in the recalibrated Dale-Chall and Homan Hewitt formula readability estimates would be accounted for by the respective number of unfamiliar words that were included in the occupational-specific vocabulary list. It was also expected that a significant amount of variance in the recalibrated FOG2 formula readability estimates would be accounted for by the number of occupational-specific vocabulary terms comprised of three or more syllables.

Summary of regression results.

For all recalibrated formulas, the number or percentage of occupational-specific vocabulary words that were originally identified as contributors to semantic complexity accounted for a significant amount of variance in the estimated readability level of the passages. For the recalibrated Dale-Chall formula, the variance accounted for ranged from 49.0% to 65.0%. For the recalibrated FOG1 formula, the variance accounted for ranged from 72.6% to 73.7%. For the recalibrated FOG2 formula, the variance accounted for ranged from 76.2% to 79.7%. For the recalibrated FOG3 formula, the variance accounted for ranged from 82.3% to 85.4%. For the recalibrated Homan-Hewitt formula, the variance accounted for ranged from 27.5% to 65.0%.

For all sets of examination materials, the regression results for each formula were significant. For the combined Book 1 and 2, the variance in estimated readability explained by the independent variables ranged from 45.9% to 84.4%. For Book 1, the variance in estimated readability explained by the independent variables ranged from 36.2% to 82.3%. For Book 2, the variance in estimated readability explained by the independent variables ranged from 27.5% to 85.4%. For all recalibrated formulas, the relationship between the number of occupational-specific vocabulary initially identified as contributors to semantic complexity (i.e., unfamiliar, long, or multisyllabic) and the readability-estimates values derived with each respective formula was negative in all sets of examination items. This indicated that fewer instances of occupational-specific vocabulary terms initially identified as contributors to semantic complexity were related to higher readability-estimate values or easier-to-read text.

The regression results, taken in sum, may be interpreted to suggest that occupational-specific vocabulary made a significant contribution to estimated readability levels. This supports the idea that the differences in estimated readability obtained with the new-model and the recalibrated formulas were related to the occurrences of occupational-specific vocabulary. Specifically, the identification of occupational specific vocabulary as unfamiliar, multisyllabic, and long (i.e., more than six letters) significantly contributed to the low (difficult-to-read) readability values obtained with the recalibrated formulas.

Post-hoc comparisons of average readability estimates.

The mean readability estimation results derived from each formula were sorted from lowest to highest for Book1 and Book 2. The orders, or rankings, were then compared across results of Book 1 and Book 2 (see Table 57). In other words, it was determined whether the formula means fell in the same order for the two books of examination items. For both sets of examination items, the recalibrated Homan-Hewitt formula resulted in readability estimates indicating greatest reading level required (i.e., lowest readability values) and the four new-model formulas resulted in readability estimates indicating the lowest reading level required (i.e., highest readability values).

The formula results were consistent across examination materials (Book 1 and Book 2). Specifically, when formula mean readability estimates were sorted from lowest to highest for each book, they fell in the same order. This offers some evidence that the new-model and recalibrated formulas performed consistently across the two sets of examination items.

Table 57

Book 1 and Book 2: all formula mean readability estimates in ascending order

Book 1				Book 2			
Formula	Mean	*SEM*	*SD*	Formula	Mean	*SEM*	*SD*
HH	513.75	89.04	327.53	HH	570.84	81.21	275.62
DC	546.88	28.39	139.07	DC	630.84	19.99	97.91
FOG2	582.08	35.76	175.18	FOG2	651.74	34.35	168.28
FOG1	626.94	35.27	172.81	FOG1	691.80	36.88	180.65
FOG3	753.08	24.11	118.13	FOG3	790.33	27.53	134.89
#C10	776.27	47.32	231.84	#C10	845.36	33.21	162.69
CL8	816.46	35.43	170.79	CL8	858.93	23.50	115.27
TUL8	852.02	33.21	162.68	TUL8	887.42	23.23	113.79
#TU8	866.08	31.36	153.62	#TU8	899.05	22.22	108.86

Note. #TU8 = new model incorporating number of T-units and unfamiliar words at level 8; TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; #C10 = new model incorporating number of clauses and unfamiliar words at level 10; CL8 = new model incorporating clause length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt.

The new-model formulas consistently provided average readability estimates that reflected easier-to-read texts than did the recalibrated formulas. This was expected because the new-models treat occupational-specific vocabulary differently than the recalibrated formulas. The new-model formulas, with the use of the occupational-specific

vocabulary list, require occupational-specific vocabulary to be considered familiar to the respective audience of readers. Whereas the recalibrated formulas require occupational-specific vocabulary to be considered a contributor to increases in semantic complexity with their identification of the terms as unfamiliar, multisyllabic, or long. In addition, these new-model formulas showed consistent patterns of results across the two books of examination items. This indicates that the formulas were performing relatively consistently across different sets of examination materials.

Post-hoc comparisons of average readability estimates: occupational-specific vocabulary list used with recalibrated formulas.

The mean readability estimation results derived from the new-model TUL8 and recalibrated Dale-Chall, FOG1, FOG2, FOG3, and Homan-Hewitt formulas, with the use of the occupational-specific vocabulary list with the recalibrated formulas, were sorted from lowest to highest for each material set. This was done to determine whether mean formula rankings would be affected when occupational-specific vocabulary was treated in the same manner across all formulas. The orders, or rankings, were compared across Book 1 and Book 2. As was found when the occupational-specific vocabulary list was not used with the recalibrated formulas, of the recalibrated formulas the Homan-Hewitt resulted in readability estimates indicating greatest reading level required (i.e., lowest readability values).

The incorporation of the occupational-specific vocabulary list with all formulas had an effect on the order in which the other formula results fell. The new-model TUL8 no longer resulted in one of the easiest estimations of readability. The new-model TUL8 resulted in readability estimates that fell is second place, indicating the second greatest

reading level required. Conversely, when the occupational-specific vocabulary list was not used with the recalibrated formulas, the TUL8 resulted in easier estimations of readability than any of the recalibrated formulas.

The use of the occupational-specific vocabulary list with the recalibrated formulas also appeared to affect the level of consistency of the rankings. When the occupational-specific vocabulary list was not used with the recalibrated formula, the results for examinations materials Book 1 and Book 2 were perfectly consistent. The formula results fell in exactly the same order and no significance tests were necessary. However, when the occupational-specific vocabulary list was used with the recalibrated formulas, the recalibrated Dale-Chall, FOG1, and FOG3 results appeared to differ across Book 1 and Book 2.

Table 58 offers a side-by-side comparison of mean readability estimates for each formula for Books 1 and 2 in ascending order. The recalibrated Homan-Hewitt resulted in the most difficult estimation of readability (i.e., lowest readability value) for both books. The TUL8 resulted in the second most difficult estimation and the recalibrated FOG2 resulted in the third most difficult estimation of readability for both books. The orders in which the results of the recalibrated Dale-Chall, FOG1, and FOG3 fell differed. One-way between groups analysis of variance (ANOVA) was conducted for both books to compare the mean estimated readability values derived with the recalibrated Dale-Chall, FOG1, and FOG3 formulas. For Book 1, the results derived from the three formulas did not significantly differ from one another (n = 24, $F_{(2,72)}$ = .209, p = .812). The same was true for Book 2 (n = *24*, $F_{(2, 72)}$ = .527, p = .593). The results of the ANOVA analyses for Books 1 and 2 indicated that although the recalibrated Dale-Chall, FOG1, and FOG3

appeared to result in differential rankings across books, their mean values were not significantly different from one another. Therefore, the slight differences in mean values that resulted in different rankings did not indicate that the formula results were actually ranked differently across the two Books.

Table 58

Book 1 and Book 2: all formula mean readability estimates in ascending order—occupational-specific-vocabulary list used with recalibrated formulas

Book 1				Book 2			
Formula	Mean	*SEM*	*SD*	Formula	Mean	*SEM*	*SD*
HH	513.75	66.86	327.53	HH	570.84	56.26	275.62
TUL8	852.02	33.21	162.68	TUL8	887.42	23.23	113.79
FOG2	955.60	16.19	79.30	FOG2	954.89	17.03	83.44
DC	984.92	17.41	85.27	FOG3	981.44	17.72	86.81
FOG3	988.55	16.99	83.22	FOG1	994.94	19.16	93.84
FOG1	1000.44	18.82	92.22	DC	1008.49	18.99	93.03

Note. TUL8 = new model incorporating T-unit length and unfamiliar words at level 8; DC = recalibrated Dale-Chall; FOG1 = stepwise derived recalibrated FOG; FOG2 = hierarchically derived recalibrated FOG; FOG3 = simple derived recalibrated FOG; and HH = recalibrated Homan-Hewitt.

The mean estimated readability levels derived with the formulas fell in different orders when the occupational-specific vocabulary list was used with the recalibrated formula than when the list was not used with the recalibrated formulas. Not surprisingly, once occupational-specific vocabulary were treated in the same manner across all

materials, the new-model TUL8 no longer resulted in one of the easiest estimations of readability. Statistical analyses revealed that where the rank orders of the formula results differed across material sets, they did not significantly differ. Nevertheless, the simple rankings of the recalibrated formula means were not as consistent across material sets as was observed when the occupational-specific vocabulary list was not used with them.

CHAPTER 5

DISCUSSION

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, p.99), "In employment and credentialing testing, the proficiency level required in the language of the test should not exceed that appropriate to the relevant occupation or profession." In addition, Downing (2006) asserted that to offer acceptable validity evidence, the content of a credentialing exam should be determined with attention to curricular documents, teaching syllabi, instructional materials and content, and textbook content, as well as other pertinent sources. Regardless of these suggestions for practice, typically readability is not formally addressed in the development of credentialing examinations. This is because, although a variety of variables have been shown to affect the readability of text, no formal method exists that is appropriate for the nature of credentialing exams and their related materials.

Previous research has clearly established that semantic and syntactic characteristics of texts are valid and reliable indicators of readability level. Homan and Hewitt contributed to this research and extended it by creating and validating a formula appropriate for the format of multiple-choice, elementary-school-level test items. The current research differs from previous research, including that of Homan-Hewitt, in that it was designed to develop a readability estimation model that accommodates not only the multiple-choice item format, but also the occupational-specific language related to credentialing examinations. The model was created to be appropriate for learning, occupational, and examination materials related to credentialing examinations.

To create this new model, the variance in readability levels accounted for by several combinations of semantic and syntactic variables was investigated. This was done with the use of cloze scores of previously validated calibration passages (Miller and Coleman, 1967) as the dependent variable and combinations of semantic and syntactic variables as the independent variables. Four new models were devised from this method. Then, existing readability formulas were recalibrated against these same passages with their existing predictor variables serving as the independent variables. The new-model and recalibrated formulas were used to estimate the readability of examination materials related to a dental-licensing program. The new-model and recalibrated formula results were compared.

This discussion is organized as follows. The results of the analyses are discussed and presented according to the phases of the investigation: Phase I: Usefulness of variables; Phase II: Formula creation and calibration; and Phase III: External validity and reliability evidence. A general discussion of the results follows and includes details related to how the current investigation is a step toward the measurement of readability levels of materials related to credentialing examinations. Then, directions for the practical application of the new-model TUL8 are outlined. The implications of this study for the dental-licensing program are then discussed. Next, the limitations of the current study are addressed. In the final section, suggestions for future research are presented.

Phase I: Usefulness of Variables

During the first phase of the investigation, Miller and Coleman's (1967) passages were analyzed according to the semantic and syntactic variables under investigation. Specifically, the syntactic analysis for each passage included determining 1) number of

T-units; 2) T-unit length (i.e., average number of words per T-unit); 3) number of clauses; 4) clause length (i.e., average number of words per clause); 5) number of sentences; 6) sentence length (i.e., average number of words per sentence); 7) percentage of passive sentences, and 8) percentage of passive verb phrases. To analyze the passages for semantic complexity, the number of words not included in *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) were determined for grade levels 4, 6, 8, 10, 12, 13, and 16.

Multiple simple linear regression analyses were conducted using the Miller and Coleman (1967) total CT scores as the dependent variable and each of the syntactic variables and the semantic variable at each level as the independent variables. All of the syntactic variables accounted for a significant amount of variance in total cloze scores. The semantic variable at levels 4, 6, 8, 10, 12, and 13 also accounted for a significant amount of variance in the total cloze scores. Level 16 of the semantic variable, however, did not account for a significant amount of variance in the total cloze scores.

Although all of the syntactic variables accounted for significant variance in total CT scores, only seven of the eight variables were retained for further investigation. Percentage of passive verb phrases was not retained because percentage of passive sentences, which also addressed voice, was highly correlated with that variable and it accounted for more variance in total CT scores. It was, therefore, deemed redundant to retain both variables for further investigation. Because level 16 of the semantic variable did not account for significant variance in total CT scores, it was not retained along with levels 4-13 of the semantic variable for further investigation for exploratory purposes.

These results clearly indicated that all of the syntactic variables under investigation were, as expected, related to the complexity or readability of the passages. The results showed that measures of T-units and clauses appeared to be more strongly related to readability than sentence measures (i.e., number of sentences, sentence length). Specifically, although both accounted for a significant amount of variance, number of T-units accounted for more variance in the total CT scores than did number of sentences. In addition, T-unit length and clause length accounted for more variance in total CT scores than did sentence length (see Table 5).

This offered preliminary evidence that measures of T-units and clauses as indicators of readability were at least as predictive as sentence measures. This was likely because T-units and clauses offer more data points for investigation. In other words, with T-units or clauses the text under investigation is divided into finer components for investigation than is possible with the sentence measures. This finer delineation of syntactic characteristics was especially important for the purposes of the present study. Specifically, the model to-be-created was meant to be appropriate for multiple-choice examination items. These sorts of texts, even after conversion into pseudo-continuous prose, tend to include fewer than 150 words. For the estimation of readability level, most existing formulas require samples of at least 150 words (e.g., Dale-Chall, FOG). Because more data points are typically provided via T-units or clause measures, they provide more information about the syntactic nature of a text. This is especially important with shorter texts that offer fewer measurement opportunities.

It was also observed that as the level of the semantic variable increased, the variance in total CT decreased. This is not surprising considering that lower levels subsume the

words contained in higher levels but also include additional words. For instance, a particular set of words identified as unfamiliar at level 8 would also be included at levels 6 and 4, but the lower levels would include words not found at level 8. Then, the words included at level 6 would obviously be included at level 4, but level 4 would include words not found at level 6. Although level 4 clearly accounted for the most variance in total CT scores, it was not necessarily the most appropriate semantic variable level for the purpose of this study because the materials for which the new model was being developed would be expected to exceed grade level 4. In addition, the readership for credentialing examinations would be expected to have reading ability levels that far exceed grade 4. Nevertheless, all of the semantic variable levels were retained and further investigated.

Phase II: Formula Creation and Calibration

During this phase of the investigation, new-model formulas were created and calibrated with the use of the Miller and Coleman (1967) passages and corresponding total CT scores. Existing readability formulas, Dale-Chall (1995), FOG, and Homan-Hewitt, were recalibrated with the same passages and total CT scores. For the calibration of the new-model formulas, stepwise multiple regression was used to explore the variance in total CT scores accounted for by the semantic- and syntactic-variable combinations (see Table 7 for details of all possible variable combinations). The semantic and syntactic variables served as the independent variables and the total CT scores served as the dependent variables. To recalibrate the existing formulas simple-linear, stepwise, and hierarchical multiple regression techniques were conducted with the respective variables for each existing formula as the independent variables and total CT scores as

the dependent variable. This resulted in recalibrations of the existing formulas that included the original variables but revised constants and weightings.

New-model Formula Creation and Calibration

The following subsections include summaries and discussions of the variable-combination analyses according to the syntactic variables included in the combinations. First, the variable combinations that incorporated number of T-units as the syntactic variable are summarized and discussed. Second, the variable combinations that incorporated T-unit length as the syntactic variable are summarized and discussed. Third, the variable combinations that incorporated number of clauses as the syntactic variable are summarized and discussed. Fourth, the variable combinations that incorporated clause length as the syntactic variable are summarized and discussed. Fifth, the variable combinations that incorporated number of sentences as the syntactic variable are summarized and discussed. Sixth, the variable combinations that incorporated sentence length as the syntactic variable are summarized and discussed. Then a discussion is offered regarding the equations identified for retention in the next phase of the investigation.

Across the analyses, several passages were removed (see methods section for details). Four passages were initially determined appropriate for removal based on total CT scores (i.e., passages 1, 3, 10, and 15), which were all at least .75 standard deviations above the mean. These high total CT scores indicated that these were the easiest passages. Additional passages were removed based on their standardized residuals, studentized residuals, and characteristics as measured according to the independent-variables under examination. Therefore, the decisions to retain particular models were made with the

consideration that these passages were inappropriate for inclusion in the development of the new model.

Number of T-units as the syntactic variable.

Stepwise regression analyses were conducted to determine the variance in total CT scores accounted for by the combination of number of T-units, percentage of passive sentences, and number of unfamiliar words (at each retained grade level). In every analysis, the percentage of passive sentences did not account for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. When the five outlying passages (passages, 1, 3, 5, 10, and 13) were removed, inclusion of number of unfamiliar words at levels 8 and 10 as the semantic variables coupled with number of T-units as the syntactic variable accounted for a significant amount of variance in total CT scores and allowed semantic and syntactic variables to enter the equation. Number of T-units and unfamiliar words at level 8 accounted for the most variance in total CT scores (R^2 = .828).

T-Unit length as the syntactic variable.

The same stepwise regression analyses were conducted, but T-unit length served as the syntactic independent variable. In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. The percentage of passive sentences accounted for a significant amount of variance in the dependent variable in only one analysis. This variable was, therefore, removed from further consideration. When the five outlying passages (passages, 1, 3, 5, 10, and 13) were removed, inclusion of number of unfamiliar words at levels 8 and 10 as the semantic variables coupled with T-units length

as the syntactic variable accounted for a significant amount of variance in total CT scores and allowed semantic and syntactic variables to enter the equation. Once again, level 8 of the semantic variable resulted in the most variance explained: T-unit length and unfamiliar words at level 8 accounted for the most variance in total CT scores (R^2 = .831).

Number of clauses as the syntactic variable.

The same stepwise regression analyses were conducted, but number of clauses served as the syntactic independent variable. In all analyses, the percentage of passive sentences did not account for a significant amount of variance in the dependent variable. This variable was, therefore, removed from further consideration. When the four passages with the highest total CT scores (passages, 1, 3, 10, and 13) and an additional outlying passage (passage 5) were removed, number of unfamiliar words at levels 8, 10, and 12 as the semantic variables coupled with number of clauses as the syntactic variable accounted for a significant amount of variance in total CT scores and allowed semantic and syntactic variables to enter the equation. Number of clauses and unfamiliar words at level 10 accounted for the most variance in total CT scores (R^2=.818).

Clause length as the syntactic variable.

The same stepwise regression analyses were conducted, but clause length served as the syntactic independent variable. In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. When the four passages with the highest total CT scores (passages, 1, 3, 10, and 13) and additional outlying passages (passages 5 and 31) were removed, inclusion of number of unfamiliar words at levels 8 and 10 as the semantic variables coupled with number of clauses as the syntactic variable

accounted for a significant amount of variance in total CT scores and allowed semantic and syntactic variables to enter the equation. Once again, using level 8 of the semantic variable resulted in the most variance explained: clause length and unfamiliar words at level 8 accounted for the most variance in total CT scores (R^2 = .849).

Number of sentences as the syntactic variable.

The same stepwise regression analyses were conducted, but number of sentences served as the syntactic independent variable. In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. When the four passages (passages, 1, 3, 10, and 13) with the highest total CT scores and an additional outlying passage (passage 5) were removed, inclusion of number of unfamiliar words at level 12 as the semantic variable coupled with number of clauses as the syntactic variable accounted for a significant amount of variance in total CT scores and allowed semantic and syntactic variables to enter the equation. Number of sentences and unfamiliar words at level 12 accounted for the most variance in total CT scores (R^2 = .448).

Sentence length as the syntactic variable.

The same stepwise regression analyses were conducted, but sentence length served as the syntactic independent variable. In the analyses, the percentage of passive sentences never accounted for a significant amount of variance in the dependent variable. Therefore, it never entered the regression equations. When the four passages with the highest total CT scores (passages, 1, 3, 10, and 13) and an additional outlying passage (passage 5) were removed, inclusion of number of unfamiliar words at level 10 as the semantic variables coupled with sentence length as the syntactic variable accounted for a

significant amount of variance in total CT scores and allowed semantic and syntactic variables to enter the equation. Number of sentences and unfamiliar words at level 10 accounted for the most variance in total CT scores (R^2 = .772).

Summary of the calibration results.

Several insights may be offered based on the results of the new-model calibration analyses. First, across all analyses the incorporation of number of unfamiliar words at levels 4 or 6 as the semantic variable did not allow syntactic variables to enter the equations once outlying passages were removed. This was because the number of unfamiliar words at levels 4 and 6 were too strongly related to, or too predictive of, total CT scores to allow syntactic variables to enter the equations. In other words, these semantic variables consumed too much variance in total CT scores to allow for the consideration of syntactic characteristics of the passages. Regardless of the amount of variance accounted for by these semantic variables, it was inappropriate to consider allowing either of them to serve as the sole variable in the new model because the aim of the design was to create models that accounted for both semantic and syntactic characteristics.

In all of the analyses reported, percentage of passive sentences was the weakest predictor variable. Although the initial investigation of the predictor variables showed that percentage of passive sentences was significantly correlated with total CT scores when it was included along with the other two independent variables, it was not sufficiently predictive of total CT scores to enter the equations. The semantic variable and other syntactic variables were stronger predictors than percentage of passive sentences and accounted for so much variance in total CT scores that there likely was

insufficient remaining variance unaccounted for to allow percentage of passive sentences to enter.

It was possible that if higher levels of the semantic variable were incorporated, enough variance would have been available for percentage of passive sentences to enter the equations, but this approach would not have been appropriate. Specifically, the initial investigation of the relationships between the independent variables and total CT scores (dependent variable) revealed that the relationship between percentage of passive sentences and total CT scores was weaker than the relationships between number of unfamiliar words at levels 4, 6, 8, 10, and 12 and total CT scores. Furthermore, the relationship between percentage of passive sentences and total CT scores was weaker than the relationships between any of the retained syntactic variables and total CT scores. Therefore, settling on a level of the semantic variable that accounted for less variance in total CT scores in an attempt to allow percentage of passive sentences to enter the equation would have compromised the integrity of the formula. Voice (i.e., passive versus active) has received little attention in readability research and was primarily included in the present investigation for exploratory purposes. Results of the present investigation suggest that voice is likely insufficiently predictive of reading difficulty to warrant inclusion in readability formulas.

Whereas number of unfamiliar words at levels 4 and 6 accounted for too much variance, number of unfamiliar words at levels 8, 10, and 12 accounted for enough, but not too much, variance to allow both syntactic and semantic variables to enter the equation. Once outlying passages were excluded from the analyses and when coupled with either number of T-units, T-unit length, or clause length as the syntactic variable,

inclusion of number of unfamiliar words at level 8 resulted in regression equations that explained the most variance in total CT scores, as compared to other levels of the semantic variable. Once outlying passages were excluded from the analyses and when coupled with number of clauses or sentence length as the syntactic variable, inclusion of number of unfamiliar words at level 10 resulted in a regression equation that explained the most variance in total CT scores. Once outlying passages were excluded from the analyses and when coupled with number of sentences as the syntactic variable, number of unfamiliar words at level 12 accounted for the most variance in total CT scores.

It was not surprising that the number of unfamiliar words at the three middlemost levels performed the best. Unfamiliar-word totals at these levels included words expected to be unfamiliar to students in grades 8, 10, and 12. In contrast to the number of occurrences of unfamiliar words at level 13, the number of unfamiliar words for levels 8, 10, and 12 were sufficient and therefore explanatory of readability (level 8 $M = 4.64$, $SD = 5.60$; level 10 $M = 3.61$; $SD = 4.09$; level 12 $M = 2.17$; $SD = 2.47$). Furthermore, the corresponding close scores for the Miller and Coleman (1967) passages were derived from cloze-test responses of undergraduate students. Level 16, which reflects college-graduate- or professional-level vocabulary, would most likely include words that are foreign to undergraduate students. Level 13 reflects undergraduate college-level vocabulary, which ideally would be expected to be familiar to the undergraduate audience from whom the scores were obtained. Nonetheless, the number of unfamiliar words at level 13 was insufficient to allow the semantic variable at that level to enter the equations.

Equations for Phase II.

According to the a priori design of this study, one new-model formula that included either a T-unit or clause measure would be selected from Phase II of the investigation to be retained for use in Phase III. This plan was modified when it was discovered that at least one combination of variables for each of these syntactic variables accounted for more than 80% of variance in total CT scores. This made it difficult to identify just one new-model formula as superior to the others. Therefore, four formulas were selected for further analyses.

Three criteria were used to select new-model formulas for further study in Phase III of the investigation. First, regression equations that necessitated the inclusion of the four passages with the highest total CT scores, which were previously identified as inappropriate for the current calibration, were not further explored. Second, to be retained a formula had to account for at least 80% of variance in total CT scores after the removal of the four passages with the highest total CT scores. Third, when a set of analysis for a specific syntactic variable included more than one equation that excluded the four passages with the highest total CT scores and accounted for more than 80% of variance in total CT score, the equation with the greatest amount of variance explained was selected for further investigation.

Based on the above criteria, four new-model regression equations were selected (see Table 59). The first new-model formula (#TU8) accounted for 82.8% of variance in total CT scores and included number of T-units as the syntactic variable and number of unfamiliar words at level 8 as the semantic variable. The second new-model formula (TUL8) accounted for 83.1% of variance in total CT scores and included T-unit length as

the syntactic variable and number of unfamiliar words at level 8 as the semantic variable. The third new-model formula (#C10) accounted for 81.8% of variance in total CT scores and included number of clauses as the syntactic variable and number of unfamiliar words at level 10 as the semantic variable. The fourth new-model formula (CL8) accounted for 82.8% of variance in total CT scores and included clause length as the syntactic variable and number of unfamiliar words at level 8 as the semantic variable.

Table 59

New-model formulas retained for further investigation

Formula name	Syntactic variable	Semantic variable	R^2	Formula
#TU8	Number of T-units	UFW-level 8	.815	Y′ = 916.646 - (18.506*UFW) + (13.544*#TU)
TUL8	T-unit length	UFW-level 8	.819	Y′ = 1192.242 - (19.278*UFW) - (8.461*TUL)
#C10	Number of clauses	UFW-level 10	.805	Y′ = 944.244 - (26.154*UFW) + (8.424*#C)
CL8	Clause length	UFW level 8	.828	Y′=1169.09 - (19.92*UFW) - (9.597*CL)

Note. UFW = number of unfamiliar words; #TU = number of T-units; TUL = T-unit length; #C = number of clauses; and CL = clause length.

Although the combination of number of sentences as the syntactic variable and number of unfamiliar words at level 8 as the semantic variable accounted 81.8% of

variance in total CT scores with the exclusion of inappropriate passages, it was not retained for further analysis. In addition, none of the formulas that included sentence length as the syntactic variable were retained. None of the sentence length formulas that were derived without the use of the passages identified for removal accounted for more than 80% of variance in total CT scores. No formulas that included measures of sentence characteristics were retained because they were not suitable for the purposes of the new-model, which was to be appropriate for the multiple-choice format of examination items. These items, even after conversion to pseudo-continuous prose, tended to include fewer than 150 words. Therefore, a finer delineation of syntactic characteristics was desirable and better achieved with measures of T-units or clauses.

Existing Formula Recalibration

Regression techniques were used to recalibrate the Dale-Chall (1995), FOG, and Homan-Hewitt readability formulas. The predictor variables for each respective formula were retained and total CT scores for the Miller and Coleman (1967) passages served as the dependent variable. In the following subsection, summaries for and discussions of each of those recalibrations are offered in turn.

Dale-Chall (1995).

Stepwise and hierarchical multiple regression techniques were used to recalibrate the Dale-Chall (1995) formula that would account for the most variance in total CT scores while excluding the passages previously identified as inappropriate for inclusion. With the removal of those four passages (passages, 1, 3, 10, and 13) and one additional outlying passage (passage 31), the stepwise multiple regression technique delivered the best model (see Table 60). This recalibrated Dale-Chall formula accounted for 88.1% of

variance in total CT scores and included number of unfamiliar words and average sentence length. When applied to the Miller and Coleman (1967) passages, the results of the original and recalibrated Dale-Chall formulas were significantly correlated: when all 36 passages were included, r = .937, and when only the 31 passages used for the recalibration were included, r = .961.

Table 60

Original and recalibrated Dale-Chall (1995) formulas

Original Dale-Chall formula	Y′= 64 - (.95 * *UFW*) - (.69 * SL)
Recalibrated Dale-Chall formula	Y′ = 1046.50 - (8.849 * UFW) + (4.984 * SL)

Note. UFW = number of unfamiliar words and SL = average sentence length.

Although it accounted for a large amount of variance, the resulting recalibrated Dale-Chall formula was inconsistent in terms of positive and negative signs. Specifically, the original formula required the weightings of number of unfamiliar words and average sentence length to be subtracted in the equation. The consistency in the signs found in the original formula was intuitive because these two independent variables would be expected to be related to readability in the same way. As the number of occurrences of unfamiliar words and average sentence length increase, the reading skill required to comprehend the material could be expected to increase. For the original Dale-Chall formula, low readability values indicate higher levels of readability or more complex text; therefore, subtracting these variables, as required by the original formula, is also intuitive.

The recalibration of the Dale-Chall formula resulted in different signs for the weightings of the predictor variables. Specifically, it required subtracting the weighting of number of unfamiliar words and adding the weighting for average sentence length. This was clearly inconsistent with what would be expected, considering both of these predictor variables should have contributed to readability in the same way. To determine the source of this discrepancy, the stepwise analysis was dissected (see Table 61).

Table 61

Recalibrated Dale-Chall formula statistics

Variable	R^2 change	*F* change	Sig. *F* change	*sr*	sr^2	*pr*	pr^2
UFW	.862	181.156	.0005	-.858	.736	-.928	.861
SL	.019	4.507	.043	.138	.019	.372	.138

Note. UFW = number of unfamiliar words; SL = average sentence length; *sr* = semipartial correlation; sr^2 = semipartial correlation squared; *pr* = partial correlation; pr^2 = partial correlation squared.

The semipartial and partial correlations were of particular interest in this analysis. Semipartial correlation values indicate the proportion of variance accounted for by a particular independent variable when the effects of other independent variables are removed. In other words, these values show the unique contribution of an independent variable in the explanation of variance in the dependent variable (Cohen & Cohen, 1975). The value of the squared semipartial correlation indicates how much variance explained in the model would decrease if that variable were removed (Cohen & Cohen, 1975). Partial correlation values, on the other hand, show the amount of variance accounted for

by a particular independent variable over and above that accounted for by other independent variables in the model. The value of the squared partial correlation indicates the proportion of variance explained by a particular independent variable that is not explained by the other independent variables in the model (Cohen & Cohen, 1975).

As indicated in Table 62, number of unfamiliar words (*sr* = -.858) was a much stronger predictor variable and accounted for a far greater amount of unique variance in total CT scores than average sentence length (*sr* = .138). The squared semipartial correlation values indicated that if average sentence length were removed from the model, only 1.9% percent of variance explained in total CT scores would be lost (UFW sr^2 = .736; SL sr^2 = .019). The partial correlation values indicated that number of unfamiliar words accounted for nearly all the variance in total CT scores (*pr* = -.928; pr^2 = .861). Therefore, the remaining variance to be accounted for by average sentence length was negligible (*pr* = .372; pr^2 = .138). The significance value for average sentence length (*p* = .043) was also worthy of note. Although, average sentence length accounted for enough variance in total CT scores to enter the model, the variable just barely met the significance requirements (i.e., $p < .05$).

It was expected that the recalibrated formula would be consistent with the original version of the formula by requiring the subtraction of both the number of unfamiliar words and average sentence length weightings. When simple linear regression was used to analyze these variables separately, the signs were consistent with those of the original formula. When the variables were both included in stepwise multiple regression analysis, the resulting regression equation required the average sentence length weighting to be

added, instead of subtracted. Even after additional analyses, the reason for this inconsistency was unclear.

FOG.

Two approaches were taken to recalibrate the FOG formula. First, the independent variables, average sentence length and percentage of multisyllabic (hard) words, were treated separately and multiple regression techniques were used to determine the amount of variance they explained in total CT scores. Second, the measures for the two variables were combined to create a single independent variable and simple linear regression was used to determine the amount of variance in total CT scores accounted for by the variable.

Stepwise multiple regression was conducted with average sentence length and percentage of hard words as the independent variables and total CT scores as the dependent variable. When all 36 passages were included in the analysis, average sentence length and percentage of hard words accounted for 74% of variance in total CT scores (R^2 = .740). Once the passages with the highest total CT scores were removed from the analysis, only percentage of hard words entered the equation. Removal of outliers did not allow average sentence length to enter the equation. Although the model that held both independent variables included passages previously identified as inappropriate for inclusion in the study, it was important to determine a recalibrated formula that included the same independent variables as the original formula. Therefore, the model derived with the use of the four passages with the highest total CT scores was retained as the recalibrated FOG1 formula (see Table 62).

In an attempt to devise a model that would provide more consistent comparisons of regression results, the four passages with the highest total CT scores were removed and both independent variables were forced into the equation. Hierarchical multiple regression analyses were conducted with average sentence length and percentage of hard words as the independent variables and total CT scores as the dependent variable. Two orders of entry for the independent variables were used: 1) percentage of hard words entered first and 2) average sentence length entered first. From both full models, percentage of hard words and average sentence length accounted for 83.3% of variance in total CT scores. When percentage of hard words was entered first in the equation, it explained all 83.3% of variance in total CT scores. Average sentence length did not account for any additional variance in total CT scores. When average sentence length was entered first in the equation, it explained only 17.6% of variance in total CT scores. Percentage of hard words accounted for an additional 65.7% of variance in total CT scores over and above the variance accounted for by percentage of hard words. Both orders of entry resulted in the same regression equation, which was retained as the recalibrated FOG2 (see Table 62).

For the next set of FOG recalibration regression analyses, the independent variables (sentence length and percentage of hard words) were summed to create a single independent variable. Simple linear regression was conducted with the sum of sentence length and percentage of hard words as the independent variable and total CT scores as the dependent variable. Once the four passages with the highest total CT scores and one additional outlying passage were removed, the summed independent variable accounted

for 73.2% of variance in total CT scores. The resulting regression equation was used to create the recalibrated FOG3 formula (see Table 62).

When applied to the Miller and Coleman (1967) passages, the results of the original FOG and recalibrated FOG formulas were significantly correlated. Specifically, with all 36 passages included, the original FOG results were correlated with those of the recalibrated FOG1 (r = - .982), FOG2 (r = - .904), and FOG3 (r = - 1.0). When only the passages used for the recalibration of the formulas were included, the original FOG results were significantly correlated with the results of the FOG2 (r = - .907) and FOG3 (r = - 1.0).

Table 62

Original and recalibrated FOG formulas

Original FOG formula	Y′ = .4 (SL) + (HW)
Recalibrated FOG1 formula	Y′ = 1277.463 - (18.192 * HW) - (8.446 * SL)
Recalibrated FOG2 formula	Y′ = 1109.175 - (18.193 * HW) - (.412 * SL)
Recalibrated FOG3 formula	Y′ = 1257.188 - (11.469 * (HW + LS))

Note. SL = average sentence length; HW = percentage of multisyllabic (hard) words.

Homan-Hewitt.

Several multiple regression approaches were necessary to recalibrate the Homan and Hewitt formula. The original Homan-Hewitt formula includes three independent variables: 1) sentence complexity (average T-unit length; WNUM); 2) number of difficult words (number of unfamiliar words; WUNF); and word length (number of words comprised of seven or more letters; WLON). The independent variables from the original

formula were used to create the recalibrated formula. The number of difficult words was to be identified at level 4 with the use of *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) word list. Using level 4 of the semantic variable did not allow the semantic variable to enter the equation. It was, then, unclear as to the level at which the words should be identified. Therefore, several multiple regression analyses were conducted to determine which level of difficult words would best fit the model and allow for the explanation of the most variance in total CT scores.

Seven sets of stepwise multiple regression analyses were conducted with sentence complexity, number of difficult words, and word length as the independent variables and total CT scores as the dependent variable. Each of the seven sets of analyses was conducted with number of difficult words identified at a different level (4, 6, 8, 10, 12, 13, or 16). Regardless of whether all 36 passages were included in the equation or inappropriate and outlying passages were removed, none of the resulting equations included all three independent variables. It was important for the independent variables from the original formula to be included in the recalibrated formula. It was also important for the independent variables to enter the equation in the order specified by Homan and Hewitt (2004; 1994) for the recalibrated formula to be as similar to the original formula as possible. Therefore, hierarchical multiple-regression was used to force the independent variables into the equation in the order specified by Homan and Hewitt.

Three sets of hierarchical multiple regression were conducted. In each set of analyses, number of difficult words was entered first, sentence complexity was entered second, and word length was entered last. One set of the hierarchical regression analyses included the identification of difficult words at level 4, one set included the identification of hard

words at level 6, and the last set included the identification of hard words at level 8. Regardless of whether inappropriate or outlying passages were removed, when difficult words were identified at levels 4 and 6, word length did not account for a significant amount of variance. When difficult words were identified at level 8, all of the independent variables accounted for a significant amount of variance in total CT scores. With the removal of the four passages with the highest total CT scores and an outlying passage, the full model explained 86.3% of variance (see Tables 27, 28, and 33). The resulting regression equation was retained as the recalibrated Homan-Hewitt formula. See Table 63 for the original and recalibrated Homan-Hewitt formulas. When applied to the Miller and Coleman (1967) passages, the results of the original and recalibrated Homan-Hewitt formulas were significantly correlated: when all 36 passages were included, r = .909, and when only the 31 passages used for the recalibration were included, r = .902.

Table 63

Original and recalibrated Homan-Hewitt formulas

Original Homan-Hewitt formula	Y′ = 1.76 + (.15 * WNUM) + (.69 * WUNF) - (.51 * WLON)
Recalibrated Homan-Hewitt formula	Y′ = 1128.958 - (.881 * WNUM) - (14.081 * WUNF) - (23.722 * WLON)

Note. *WNUM* = sentence complexity; *WUNF* = number of difficult words; and *WLON* = word length.

Phase III: External Validity and Reliability Evidence

Four new-model formulas were created in Phase II and were retained for further investigation during Phase III. As explained in the results section, the new-model TUL8

appeared to show marginally more consistent performance than the other three formulas. Because the new-model TUL8 performed marginally better and because discussing the results for all new-models would be redundant and cumbersome for the reader, this portion of the discussion will be primarily focused on the performance of the new-model TUL8. In a few instances, however, the results for all formulas are referenced. Clear distinctions are made when the results of all new-models are being referenced as opposed to the results of the new-model TUL8 alone.

Evidence collected during Phase III of the investigation suggested that the new-model TUL8 showed promise as a means of establishing readability while accommodating the multiple-choice item format and occupational-specific language related to credentialing examinations. The results of the correlation analyses, Sign tests, regression analyses, and rank ordering of formula results all supported this notion. The results of each of these analyses sets and manner in which they support the utility of the TUL8 are discussed in turn in the following sections. A summary of these findings is then offered.

The initial correlation analyses revealed that only a one, very weak, significant relationship existed between the results of the TUL8 and any recalibrated formula. It was assumed that the failure to find significant relationships between the results of the new-model and recalibrated formulas was due to the differential treatment of occupational-specific vocabulary in the new-model and recalibrated formulas. Once the occupational-specific vocabulary list was used with the recalibrated formulas, the results of the TUL8 were significantly correlated with the results of all recalibrated formulas ($p < .01$). Finding these substantial increases in the relationships between the TUL8 and recalibrated formulas confirmed the assumption that the initial failure to find significant

differences was largely due to the occupational-specific vocabulary being treated differently in the new-model and recalibrated formulas.

The initial Sign tests conducted to compare the results of the new-model and recalibrated formulas were perfectly consistent across combined Books 1 and 2, Book 1, and Book 2. In other words, significant differences observed between the results of the TUL8 and recalibrated formulas were consistent across the three examination-item sets. Furthermore, where a significant difference was not observed between the TUL8 and a recalibrated formula, the results were also consistent across the three examination-item sets. This indicated that the TUL8 performed consistently when applied to two different books of items, which offers some credibility to the stability of the TUL8 model.

The initially conducted Sign tests revealed that the TUL8 resulted in significantly easier estimations of readability (higher readability values) than the recalibrated Dale-Chall, FOG1, FOG2, and Homan-Hewitt. As with the initial failure to find significant correlations between the results of the TUL8 and recalibrated formulas, the differential treatment of occupational-specific vocabulary by the new-model and recalibrated formulas was surmised to be the reason for the significant differences between results. This assumption was supported with the results of post-hoc analyses. Specifically, the occupational-specific vocabulary list was used with the recalibrated formulas and Sign tests were conducted to compare those results to the results of the TUL8. The effect of the use of the occupational-specific vocabulary list with the recalibrated formulas was so powerful that where significant differences remained between the results of the TUL8 and the recalibrated Dale-Chall, FOG1, and FOG2, they changed directions. In other words, when occupational-specific vocabulary words were treated in the same manner

across all formulas, the recalibrated Dale Chall, FOG 1, and FOG2 resulted in significantly easier estimations of readability (higher readability values) than the TUL8. Furthermore, the initial Sign tests showed no significant difference between the results derived with TUL8 and recalibrated FOG3. However, when the occupational-specific vocabulary list was used with the recalibrated FOG3, it resulted in significantly easier estimations of readability (higher readability values) than the TUL8. The recalibrated Homan-Hewitt continued to result in significantly more difficult estimations of readability (lower readability values) than the TUL8. However, the Homan-Hewitt formula included two, rather than one, semantic variable. The additive effect of the two semantic variables included in the Homan-Hewitt resulted in more substantial estimations of semantic complexity than the TUL8, which only included a single semantic variable.

The results of the regression analyses supported the conclusion that the significant differences observed between the new-model and recalibrated formula results were due to the manner in which occupational-specific vocabulary were treated. Specifically, occupational-specific vocabulary that was identified in the recalibrated formulas as contributors to semantic complexity (i.e., identified as unfamiliar, long, or multisyllabic) accounted for an extraordinary amount of variance in readability estimates derived with each recalibrated formula.

The post-hoc rank ordering of the formula results provided further evidence of the stability of the new-models. Two rank orderings were conducted. First, the mean readability estimates derived with each formula under investigation were sorted from low to high. For both Book 1 and Book 2, the TUL8 fell in last place, which indicated that it resulted in mean readability estimates that reflected that the materials were easier to read

than was indicated by the results of the recalibrated formulas. Interestingly, the rank ordering of the recalibrated formulas was perfectly consistent across Book 1 and Book 2. This indicated that the recalibrated formulas were also showing a good degree of stability.

The second post-hoc rank ordering of the mean readability results were more informative than the first set of rank orderings. In this set of rank orderings, the occupational-specific vocabulary list was used with the recalibrated formulas and those results were rank ordered along with the results of the new-model TUL8. Not surprisingly, the TUL8 no longer fell in last place or resulted in mean readability estimates that reflected that the materials were easier to read than was indicated by the results of the recalibrated formulas. Instead, the results of the TUL8 formula fell in second place for Book 1 and Book 2. With the incorporation of the occupational-specific vocabulary list, the recalibrated Homan-Hewitt resulted in mean readability estimates that were lower than the mean readability estimates derived with the TUL8. All of the other recalibrated formulas resulted in mean readability estimates that indicated the texts were easier to read (higher readability values) than was indicated by the mean readability estimates derived with the TUL8.

The use of the occupational-specific vocabulary list with the recalibrated formulas, however, appeared to slightly affect the stability of the recalibrated formula results. When the list was not used with the recalibrated formulas, the rank ordering of the readability estimates derived with them was perfectly consistent across Book 1 and Book 2. When the occupational-specific vocabulary was used with the recalibrated formulas, the simple rankings of the recalibrated Dale-Chall, FOG1, FOG2, and FOG3 were entirely different

across Book 1 and Book 2. Admittedly, one-way between groups analysis of variance (ANOVA) revealed that the mean readability values were not significantly different from one another. Nonetheless, the simple rankings were inconsistent under these circumstances.

The results outlined above lend support to the utility of the new-model. Furthermore, they implicate the failure to account for occupational-specific vocabulary in the recalibrated formulas as the source for the initially observed non-significant correlations and significant differences between the new-model and recalibrated formulas. Finding that the differential treatment of occupational-specific vocabulary in the new-model and recalibrated formulas was responsible for the weak correlations and significant differences substantiates the importance of considering occupational-specific vocabulary in the estimation of readability of credentialing examination items. Furthermore, finding that the incorporation of the occupational-specific vocabulary list with the recalibrated formulas appeared to slightly affect the consistency of the rank orderings of the mean readability estimates derived with recalibrated formulas suggested that simply using the list with the existing formulas may not be appropriate.

The introduction of the occupational-specific vocabulary list with the recalibrated formula and subsequent analyses that were conducted with the resulting readability values certainly provided some explanation for the initially weak correlations and significant differences between the new-model and recalibrated formulas. However, although all of the correlation values observed in the post-hoc analyses of the examination materials were significant ($p < .01$), some were weak and none were better than moderate. Conversely, the validation portions of previous research conducted to

create or modify readability formulas also included correlation analyses of readability estimates across formulas (e.g., Farr, Jenkins, & Patterson 1951; Fry 1968) and very strong relationships were observed. Fry (1968), for example, found correlations between the result of his formula and the Flesch and Dale-Chall were r = .96 and .94, respectively. Farr et al. (1951) found that the relationships between the original Flesch and the revised version of the Flesch to be r = .93.

The correlations between formula results shown in previous research were clearly much stronger than correlations observed during the post-hoc correlation analyses of the readability estimates in this study, which was initially surprising. However, the weaker correlations observed in this study were not an artifact of imprecision of the new-models. Rather the weaker correlations were probably a result of the level of appropriateness of using the recalibrated formulas with examination items and the manner in which the formulas addressed text characteristics.

The Dale-Chall and FOG formulas, which were used in the current investigation, were designed for use with several 100-word samples of continuous prose. They were not designed to be used to estimate the readability of single samples of pseudo-continuous prose, many of which were comprised of fewer than 100 words. Even after the examination items were converted into pseudo-continuous prose, many included far fewer than 100 words. Specifically, the items included in Book 1 ranged from 41 to 249 words (M = 80.22, SD = 43.13), and the items included in Book 1 ranged from 44 to 378 words (M = 93.96, SD = 31.01). Therefore, the weaker than expected correlations between the new-model and recalibrated Dale-Chall and FOG formulas may have been

due to the inappropriateness of the materials for use with the recalibrated, existing formulas.

The correlations between the new-model TUL8 and recalibrated Homan-Hewitt were stronger than the relationships between the TUL8 and recalibrated Dale-Chall or FOG. Unlike the Dale-Chall and FOG formulas, the Homan-Hewitt formula was designed for use with smaller text samples with no specific guidelines for how many words should be included and the authors did not indicate that multiple samples were necessary for accurate estimation. Even so, the correlation between the new-model TUL8 and recalibrated Homan-Hewitt was only of moderate strength (i.e., $r = .714$). The Homan-Hewitt formula, however, includes two measures of semantic complexity; whereas all of the other models investigated here include only one measure of semantic complexity. The additional measure of semantic complexity in the Homan-Hewitt formula was likely what prevented the correlations from being stronger than what was observed.

The weaker than expected correlations between the new-model TUL8 and recalibrated formulas even after the use of the occupational-specific vocabulary list with the recalibrated formulas was likely also due to the methods used to measure text characteristics. The new-model and recalibrated formulas incorporated different measures of semantic complexity. For example, a word identified as unfamiliar according to the new-model might, or might not, be identified as unfamiliar according to the Dale-Chall or vice versa. In addition, a word identified as multisyllabic (a FOG formula) might, or might not, be identified as familiar according to the new-model specifications and vice versa. Whether these measures of semantic complexity resulted in different findings was likely also affected by the mere nature of a text. For instance, the multisyllabic words

included in a text might be the same words identified as unfamiliar, essentially by chance. Furthermore, representations of semantic variables differed across the new-models and some of the recalibrated formulas. Specifically, the new-model formulas identified semantic characteristics with frequency counts of the existence of unfamiliar words; whereas, the FOG formulas identified semantic characteristics with percentage values for multisyllabic words.

Semantic characteristics for the TUL8 formula were measured by identifying the number of unfamiliar words at level 8 according to *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) list of familiar words and the occupational-specific word list. For the Dale-Chall formula, semantic characteristics were measured by identifying the number of unfamiliar words according to the Dale-Chall (1995) list of familiar words. The correlation between the semantic complexity measures of the TUL8 and Dale-Chall was $r = .508$. For the FOG formula, semantic characteristics were measured by identifying the percentage of words comprised of three or more syllables (multisyllabic). The correlation between the semantic complexity measures of the TUL8 and FOG was $r = .651$. Semantic characteristics for the Homan-Hewitt formula were measured by identifying the number unfamiliar words at levels 8 according to *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) list of familiar words and the number of words comprised of seven or more letters (long). The correlation between the semantic complexity measures of the TUL8 and the combined values for the semantic complexity measures of the Homan-Hewitt was $r = .832$.

Identification of syntactic characteristics also differed across formulas. Syntactic characteristics for the TUL8 and Homan-Hewitt formulas were identified by measuring average T-unit length. The recalibrated Dale-Chall and FOG formulas used sentence length as the indicator of syntactic complexity. Obviously, because T-units and sentences are different measures of syntactic complexity, they potentially result in dissimilar findings. Identification of sentence properties results in a less precise characterization of syntactic complexity than is offered by T-unit and clause properties. In some cases, T-unit measures are equivalent to sentence measures; but in other cases, T-unit measures allow a more precise measure of syntactic properties within sentences. In other words, a single sentence often includes several T-unit, which can be identified within that sentence.

How the results of sentence, T-unit, and clause measures correspond is clearly affected by the nature of a text. Specifically, if a text is syntactically simplistic, these values are likely to correspond very well; whereas, if a text is syntactically complex, sentence-measure values are less likely to correspond as well with the T-unit- and clause-measure values. Consider, for instance, that a passage has the same number of sentences, T-units, and clauses. That would indicate that the sentences in that passage are rather simplistic in that they do not contain multiple combinations of subjects and verbs. (Recall that a T-unit is defined as "one main clause plus the subordinate clauses attached to or embedded within it" [Hunt, 1965, p. 49] and a clause is defined as "a structure containing a subject [or coordinating subjects] and a finite verb phrase [or coordinating verb phrases]" [Hunt, 1965, p. 40.]) Conversely, a passage with significantly fewer sentences than T-units or clauses is likely more syntactically complex. That is, at least some of the

sentences must include multiple T-units or clauses and therefore multiple subject-verb combinations.

Inspection of the syntactic characteristics of the examination items, as measured by the new-model TUL8 and recalibrated formulas, revealed that the materials were sufficiently syntactically simple that the measures used in the different formulas yielded very similar results. Specifically, 45 of the 48 passages had the same number of sentences and T-units. Furthermore, sentence length and T-unit length were nearly perfectly correlated (r = .998) and their mean values were practically identical in that they only differed by a tenth of a point (sentence length: *M* = 15.63; *SD* = 6.34; T-unit length: *M* = 15.53; *SD* = 6.39).

The inspection of the syntactic characteristics offered a great deal of insight into the nature of the examination items. Specifically, by design the examination materials were rather syntactically simplex. Very few sentences included more than one T-unit. This means that most of the sentences did not include multiple subject-verb combinations. In other words, nearly every sentence was identified as one complete T-unit. Therefore, the readability formulas resulted in nearly identical measures of syntactic complexity and essentially only differed by the semantic-complexity measure along with respective constants and weightings. Because the syntactic-complexity measures were essentially identical across the new-model and recalibrated formulas, differences in the readability estimates could be attributed primarily to the measurement of semantic complexity. The weaker than expected correlations between the results of the TUL8 and recalibrated Dale-Chall and FOG could not be confidently attributed to the different measures of syntactic complexity used in the formulas.

While the administrators of the dental licensing program were apparently diligent in ensuring that the items included in their licensing exam were devoid of complex syntactical structure, it would be erroneous to assume that examinations for other credentialing programs would be equally as syntactically simplex. When the dental licensing examination was constructed, the administrators were aware that the examination would be translated from English to French. It would follow that the program administrators would make every attempt to facilitate the most accurate translation possible. Syntactic simplicity of the examination items, therefore, would be of paramount importance.

Summary of Phase III Discussion

It was posited that differential treatment of occupational-specific vocabulary in the new-model TUL8 and recalibrated formulas was largely responsible for the initially observed weak and non-significant correlations and significant differences between the readability estimates derived with the respective formulas. This supposition was substantiated with the findings of post-hoc correlation analyses and Sign tests. When the occupational specific vocabulary list was used with the recalibrated formulas, the correlations strengthened and reached significance and the results of the Sign tests were dramatically different than when the occupational-specific vocabulary list was not used with the recalibrated formulas. The results of the TUL8 were still significantly different from the results of the recalibrated Dale-Chall, FOG1, and FOG2, but the differences were in the opposite direction than they were without the use of the occupational-specific vocabulary list.

The relationships between the new-model TUL8 and recalibrated formulas markedly increased when occupational-specific vocabulary words were treated in the same manner across all formulas. However, even after the use of occupational-specific vocabulary with the recalibrated formulas, the relationships between the readability estimates derived with the new-model TUL8 and recalibrated formulas were moderate at best and in one pairing the correlation was weak. These weaker than expected correlations observed in the post-hoc analyses were thought to be attributable to the different methods used to measure semantic and syntactic complexity. Inspection of the syntactic measurement values yielded by the new-model TUL8 (T-unit length) and the recalibrated Dale-Chall and FOG (sentence length) formulas revealed that due to the nature of the examination materials, the values were nearly identical and almost perfectly correlated. Therefore, the weaker than expected correlations observed in the post-hoc correlation analyses were attributed almost solely to the different methods used in the new-model TUL8 and recalibrated formulas to measure semantic complexity.

The correlation results between the semantic-complexity measures derived with the new-model TUL8 and recalibrated formulas support that the different methods used to measure semantic complexity were responsible for the weaker than expected correlations between new-model and recalibrated formulas. The semantic-complexity measure derived with the new-model TUL8 was most weakly correlated with the semantic-complexity measure derived with the recalibrated Dale-Chall. Correspondingly, the readability estimate derived with the new-model TUL8 was most weakly correlated with the readability estimate derived with the recalibrated Dale-Chall. The semantic-complexity measure derived with the new-model TUL8 was most strongly correlated

with the semantic-complexity measure derived with the recalibrated Homan-Hewitt. Accordingly, the readability estimate derived with the new-model TUL8 was most strongly correlated with the readability estimate derived with the recalibrated Homan-Hewitt. As compared to those correlations, the relationships between the readability estimates and semantic-complexity measures of the new-model TUL8 and FOG formulas fell in the middle.

To provide more compelling evidence for the suitability of the readability levels of the examination items of the dental-licensing program, further studies should be conducted. These studies could be conducted with new sample materials that also include learning and occupational texts and the same new-model formulas or the same sample materials and different readability formulas. This might help substantiate that the examination items are written at an appropriate readability level.

General Discussion

The purpose of this study was to develop a set of procedures to establish readability, including an equation, that would accommodate the multiple-choice item format and occupational-specific language related to credentialing examinations. The procedures and equation were to be appropriate for learning materials, examination materials, and occupational materials. The procedures developed as well as the semantic and syntactic variables explored in the current study appear to be appropriate for such a model.

The new-models are more appropriate, or better-refined versions of them will be more appropriate, for use with credentialing examination materials than existing readability formulas for four reasons. First, the new-models involve consideration of discipline-specific, technical language that appears in credentialing program materials.

With the use of existing readability formulas, technical language, or occupational-specific vocabulary, has the propensity to artificially inflate readability estimates of credentialing-related materials. Occupational-specific words are often multisyllabic, long, and not likely to appear on lists of familiar words. Therefore, with the use of existing readability formulas, these words are typically identified as contributors to semantic complexity. This is inappropriate because candidates who take a credentialing exam should be familiar with the relevant occupational-specific vocabulary.

During the investigation of external validity and reliability of the new-models developed in this study, an occupational-specific vocabulary list of nearly 5,000 words was created for the field of dentistry. This list was used in conjunction with *The Living Word Vocabulary: A National Vocabulary Inventory (*Dale & O'Rourke, 1981) to identify unfamiliar words. The time and resources required to create the occupational-specific vocabulary list were daunting, but it appeared to greatly contribute to the utility of the new-model. The effect of the occupational-specific vocabulary list was apparent when the list was used with the recalibrated formulas. The readability estimates derived with the recalibrated formulas differed considerably when the occupational-specific vocabulary list was used as opposed to when it was not used.

Second, aside from the development and implementation of the occupational-specific vocabulary list, the new-models provide a more comprehensive measure of semantic complexity with the use of *The Living Word Vocabulary: A National Vocabulary Inventory (*Dale & O'Rourke, 1981) to identify unfamiliar words. When used in conjunction with the occupational-specific vocabulary list, use of *The Living Word Vocabulary* is likely to result in more precise measures of semantic complexity of

credentialing-related materials than is possible with the methods used in existing formulas. This list includes a corpus of over 44,000 familiar words. In addition, multiple meanings of the words included in the list are delineated by the grade level at which they should be considered familiar. Lists of familiar words incorporated by other formulas (e.g., Dale-Chall, 1943; 1995) are much less exhaustive and words within them are not delineated by the grade level at which they are expected to be familiar. Furthermore, other lists of familiar words do not include multiple meanings; therefore, the context in which a word is used is not considered in its identification as familiar or unfamiliar.

*The Living Word Vocabulary: A National Vocabulary Inventory (*Dale & O'Rourke, 1981) also likely provides a more precise indication of semantic complexity than syllable or letter counts, which are used in some existing formulas. With the use of *The Living Word Vocabulary* words need not be short or monosyllabic to be identified as familiar, nor are they identified as unfamiliar simply because they are long or multisyllabic. Although it seems logical that longer or multisyllabic words would be more difficult, this is not always the case. For instance, with letter or syllable counts, the word "important" would be inappropriately identified as a contributor to semantic complexity.

Third, the syntactic variables investigated for use in the new-models offer more appropriate measures of syntactic complexity for the intended materials. Specifically, measures of T-units and clauses, which were investigated for the new-models, offer more measurement points than sentence measures, which are incorporated in most existing formulas. The use of sentence measures in the existing formulas is appropriate for their intended use. Existing readability formulas are typically intended to be used with several samples of more than 100 words for reliable evaluation. However, sentence measures are

less appropriate for use with examination items. Multiple-choice examination items are generally constructed to be concise and tend to include fewer than 100 words. Their stems are usually between one and three sentences long and the response options are typically shorter. With fewer pieces of data to investigate, it is greatly advantageous to have more precise measures and as many measurement opportunities as possible, which is more likely with the measurement of T-unit or clause properties.

The fourth reason the new-models are better suited for use with credentialing-related materials is also related to the nature of multiple-choice examination items. Not only are multiple-choice test items typically constructed to be concise, but incomplete sentences are also often provided as options. Furthermore, test items are not continuous prose. Existing readability formulas are intended to be used with continuous prose and are not suited for use with non-continuous prose that includes incomplete sentences. The new-model, however, provides methods to accommodate the nature of the examination items. Procedures similar to those used by Plake (1984) were created to convert examination items into pseudo-continuous prose. The use of these procedures enabled consistent syntactic-characteristic measurement. Without a procedure to convert the items into pseudo-continuous prose, several pieces of text would be impossible to analyze according to their syntactic characteristics because they would not include the necessary subject-verb combinations.

The new-models show four major advantages over existing readability formulas that suggest they are more appropriate for use with credentialing related materials: 1) they include a method to accommodate occupational-specific vocabulary; 2) they include a more precise measure of semantic complexity; 3) they include a more precise measure of

syntactic complexity; and 4) they incorporate a method to convert examination items into pseudo-continuous prose. However, one might posit that the procedures for existing formulas could simply be modified to include the development and incorporation of an occupational-specific vocabulary list and procedures to convert the non-continuous prose of examination items into pseudo-continuous prose. Although this a tempting alternative to using the new-model in its entirety, readability estimates derived in this manner may not be as accurate or stable as would be possible with the new-models or future versions of them.

The new-models, or future versions of them, are potentially superior to modified versions of the existing formulas for two primary reasons. First, the syntactic measures used in all but one (i.e., Homan-Hewitt) existing formulas are not capable of returning the level of detail that is possible with T-unit and clauses. More syntactically complex sentences tend to include multiple T-units and clauses. Therefore, using measures of T-units or clauses for syntactically complex sentences will result in more data points for investigation. Specifically, if sentence measures were used for that type of complex sentence, one piece of data would be obtained. However, if T-unit or clause measures were used to quantify the syntactic complexity of that same sentence, multiple pieces of data could be obtained. This would result in a more accurate estimation of syntactic complexity for passages in which syntactically complex sentences exist.

Precision is always a priority in the estimation of syntactic complexity, but the advantage of using more precise measures of syntactic complexity is especially important for examination items because they often include fewer than the minimum number of words required by most existing formulas. Granted, in the current investigation it was

revealed the sentence and T-unit measures for the examination materials did not significantly differ and were nearly perfectly correlated. However, this might not always be the case and is really an indication of the dental program's mindfulness in their creation of examination items. The dental program appropriately used syntactically simple language when they created their test items. It would inappropriate to assume that all programs do the same. The professional dental licensing examination investigated here is developed in English but is later translated to French. Therefore, it is likely that during item development great efforts are made to ensure that the syntactic complexity of the items is kept to a minimum to help ensure the most precise translation possible. Credentialing programs that do require item translation may not go to such efforts to ensure this syntactic simplicity of items if their respective examinations are only delivered in English and not subject to translation.

Second, aside from *The Living Word Vocabulary: A National Vocabulary Inventory (*Dale & O'Rourke, 1981) potentially providing a more comprehensive and accurate estimation of semantic complexity than the word lists or methods used in existing formulas; simply introducing the occupational-specific vocabulary word list for use with the existing formulas might not be appropriate. Data in this investigation suggest that using the occupational-specific vocabulary list with existing formulas may affect their results in an unpredictable way. When the occupational-specific vocabulary list was not used and the results of the recalibrated formulas were sorted, the rank orders of the formula results did not vary across Book 1 and Book 2. For example, for both books the readability estimates derived with Homan-Hewitt formula were lower (indicating harder-to-read text) than the readability estimates derived with any other formula. This indicated

that recalibrated formulas were performing rather consistently for the different types of materials. However, when the occupational-specific vocabulary list was introduced for use with the recalibrated formulas and the results of the recalibrated formulas were sorted, the simple rankings of the recalibrated formula results varied across Book 1 and Book 2. Specifically, the order in which the formula results fell for Book 1 were different from the order in which they fell for Book 2. Statistical analyses revealed that where the rank orders of the formula results differed across books, they did not significantly differ. Nevertheless, the simple rankings of the recalibrated formula means were not as consistent across material sets (Book 1 and Book 2) as was observed when the occupational-specific vocabulary list was not used with them. This suggested that the incorporation of the occupational-specific vocabulary list for use with recalibrated formulas might have divergent effects for the different material sets. Further investigations that ensure sufficient power would better elucidate whether this is actually a matter of concern.

The new-models showed a good degree of consistency throughout this investigation. The rankings of readability estimates across Book 1 and Book 2 showed that the results of the new-models were consistent for different sets of sample materials. Taken together, the results of this investigation suggest that the new-models show promise for use with credentialing-related materials. That is not to say that any of the formulas are in their final form, as the variables should be further investigated and the formulas should be subjected to further calibration studies. Nevertheless, the procedures developed as well as the semantic and syntactic variables investigated for the new-models appear to offer a

more suitable method for measuring the readability of credentialing examination materials than existing formulas.

Practical Application of the New-model

The new-model TUL8 is intended to be appropriate for readability estimation of credentialing materials. This model should be appropriate for the format and content of learning, occupational, and examination materials. To apply the TUL8 equation to estimate the readability of credentialing materials, a set of procedures should be followed. These procedures are outlined in the following sections. Procedures for selecting samples for investigation are explained first. Next, the identification of relevant semantic and syntactic variables is discussed. This discussion includes a description of the materials necessary to address semantic complexity and the methods to be used in the development of an occupational-specific vocabulary list. Finally, the TUL8 equation to be applied to semantic and syntactic data gathered for the material sets is provided along with a brief explanation of how the resulting readability estimation values should be interpreted.

If the readability of credentialing materials are to be addressed for a particular program, the issue of readability should be addressed prior to the development of the respective examination instead of being treated as an afterthought to test development. Attending to the readability levels of examination items for a respective credentialing program post-hoc, would likely inhibit the implementation of steps necessary to ensure essential equivalence across learning, examination, and occupational materials. A program would be well served by addressing readability in the early phases of examination development by assessing the readability levels of relevant learning and occupational materials prior to the development of examination items. The results of such

analyses would facilitate the program administrators' knowledge and understanding of the readability levels of texts to which examinees are exposed in learning and occupational environments. This, in turn, could provide information to help guide the development of examination items that are of appropriate readability levels. Furthermore, periodic checks of the readability of examination items during item development would help ensure that the items are being created at appropriate readability levels. Finding initial incongruence or unacceptably high readability levels of examination items during development phases would allow program administrators to make informed decisions regarding item-development practices that may require amendment.

Finding unequal levels of readability across learning and occupational materials may put program administrators in a precarious position. They must then determine to which readability-level-standard they should hold themselves. Specifically, they must decide whether to target their examination items to the readability level of the learning or occupational materials.

Estimating the readability of examination items used in credentialing examinations without also estimating the readability of related learning and occupational materials would not provide an investigator useful information. The purpose of assessing the readability of examination items is to enable the comparison of those readability levels with the readability levels of materials used during educational or training courses and materials used on the job. Establishing that the readability levels are essentially equal for the examination and occupational materials addresses the issue raised in *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) and those raised by Plake (1988). Establishing that the readability level are essentially equal for the

examination and learning materials addresses the issues raised by Plake (1988) and Downing (2006). Therefore, to effectively apply the new-model TUL8, samples should be collected from learning, examination, and occupational materials.

Material Samples

Learning-material samples should be collected from relevant text books, journal articles, and any other sources that are pertinent to educational or training programs in which candidates generally participate in preparation for the credentialing examination. A subject matter expert should be consulted to ensure proper identification of relevant sources. Sample of approximately 150 words should be extracted from each of the sources. Equal number of samples should be selected from the beginning, middle, and ends of these sources.

Collecting samples from multiple-choice examination materials requires access to relevant item-difficulty data. The data should be used to conduct stratified, systematic sampling to ensure equal representation of items at different difficulty levels. First, the items should be sorted according to difficulty level and then divided into three groups according to difficulty (high, medium, and low). Then, the items should be resorted within each group or strata according to their item identification codes, or the items should be un-sorted in some other way so that they no longer appear in order of difficulty. Starting at an n^{th} item, every n^{th} item within each stratum should be identified for selection. Once a representative sample of examination items is selected, the items should be converted from non-continuous prose into pseudo-continuous prose. Guidelines for conducting these conversions are outlined in the methods section of this document.

Occupational-materials samples should be extracted from texts that are representative of what a practicing professional would likely encounter on the job. These materials might include instruction manuals, product and equipment manuals, professional journal articles, memos, or professional journal editorials. Subject matter experts should be consulted to ensure the relevance of sources identified. A sample of approximately 150 words should be extracted from each of the collected sources. Equal numbers of samples should be selected from the beginning, middle, and ends of these sources.

Analyzing the Materials According to Semantic and Syntactic Characteristics

The new-model TUL8 involves the measurement of semantic and syntactic characteristics. The manner in which these characteristics should be addressed is discussed in the following sections. First, directions for semantic-complexity measurement are provided. This discussion begins with an explanation of the materials that are required to perform semantic-complexity estimations and the methods for developing an occupational-specific word list. Then, a description of the methods for assessing syntactic complexity is provided.

The new-model TUL8 requires the use of two lists of familiar words for the assessment of semantic complexity or vocabulary load. The first word list, *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) can be acquired through university libraries. The second list of familiar words is the occupational-specific vocabulary list and it must be created by the investigator.

The use of the occupational-specific vocabulary list enables appropriate accommodation for occupational-specific vocabulary included in relevant materials (e.g., learning, examination, occupational). This list should include words that would

reasonably be assumed to be familiar to candidates expected to take the examination. To create this list, discipline-specific glossaries or textbook appendices should be referenced. Once seemingly appropriate sources are identified, a subject matter expert should be consulted to ensure that the sources are appropriate and that important sub-domains are represented. These sources should be used to create an exhaustive list of occupational-specific vocabulary. Once again, a subject matter expert should be consulted to ensure that the list is sufficiently comprehensive.

Use both word lists to assess the semantic complexity of the learning, occupational, and examination materials. *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) should be used first to identify words in the passages that are unfamiliar at grade-level 8. The unfamiliar words should be marked and counted. Second, the words identified as unfamiliar according to *The Living Word Vocabulary: A National Vocabulary* should be checked against the occupational-specific vocabulary list. More specifically, the words that were identified as unfamiliar according to *The Living Word Vocabulary: A National Vocabulary Inventory* but appear in the occupational-specific vocabulary should be removed from the unfamiliar-word totals. By this, only non-domain-specific vocabulary terms are subject to identification as unfamiliar and thereby contributors to semantic complexity. The number of words that were identified as unfamiliar with the use of both word lists should be totaled. To determine the unfamiliar word value for a passage, the sum of unfamiliar words should be divided by the number of words in the passage and that quotient should be multiplied by 150. For example, a passage consisting of 158 words, 14 of which are identified as unfamiliar (with the use of

both word lists) should have an unfamiliar word or semantic complexity value of 13.29 [(14/158)*150 = 13.29].

To estimate syntactic complexity, the learning, occupational, and examination materials should be analyzed according to average T-unit length. The first step in establishing average T-unit length is to enumerate the T-units in the passages. T-units include "one main clause plus the subordinate clauses attached to or embedded within it" (Hunt, 1965, p. 49). For example, the sentence, "This is normally the case in the spleen and the bone marrow, which are prominently affected by sickle cell disease" includes one main clause (i.e., "This is normally the case in the spleen and the bone marrow") and one subordinate clause (i.e., which are prominently affected by sickle cell disease") and is identified as a single T-unit. The sentence, "However, dental needs across large populations are uniform, and the costs are relatively small" includes two independent clauses (clause 1 is "However, dental needs across large populations are uniform"; clause 2 is "and the costs are relatively small") and, therefore, two T-units. To determine average T-unit length, divide the total number of words included in the passage by the total number of T-units in the passage. For instance, a passage comprised of 158 words and 7 T-units would have an average T-unit length of 22.57 (158 / 7 = 158).

Applying the Equation and Interpreting the Results

The unfamiliar word value (semantic complexity measure) and the average T-unit length (syntactic complexity measure) for each passage should then be included as the semantic and syntactic variables in the TUL8 equation. The TUL8 formula is as follows:

$$Y' = 1192.242 - (19.278 * UFW) - (8.461 * TUL)$$

(Where UFW = unfamiliar word value and TUL = T-unit length).

The equation result provides a readability estimate for a particular passage. The readability estimate values for passages included in a set of materials can be averaged to determine a mean readability estimate. Higher readability estimate values indicate easier-to-read text and lower readability estimate values indicate harder-to-read text. These readability estimate values do not correspond with grade-levels or the level of reading ability necessary to understand the texts. Instead, they should be used to rank order the learning, occupational, and examination materials in terms of readability levels.

Implications for the Dental-licensing Program

The readability level of the examination materials, as determined according to any of the new-models, in and of itself does not provide the dental-licensing program sufficient information to determine whether the items are of appropriate readability levels. Making that determination would require readability-level assessment to also be conducted for relevant learning and occupational materials. Obtaining readability estimates for all material sets (learning, occupational, and examination) would enable meaningful comparisons across readability levels and the ability to determine whether the readability levels of the examination items are appropriate.

The current readability level data, however, does offer the dental licensing program some insight into the nature of their examination items. More specifically, when the variables included in the new-model TUL8 were inspected, it was revealed that the examination items were syntactically straightforward. This information should provide the dental-licensing program with some confidence that any efforts made to ensure that the items were devoid of undue linguistic complexity were successful.

Whereas it was possible to make some determination about the syntactic complexity of the materials with comparisons of T-units and sentences, it was not possible to make similar determinations regarding the degree of semantic complexity for the examination items. It was possible, however, to elucidate the impact of accounting for occupational-specific vocabulary by treating such terminology as familiar. The measurement values for semantic complexity were dramatically affected by the use of the occupational-specific vocabulary list. One might surmise that because the readability estimates are merely used as a means to rank order materials (learning, examination, and occupational) and not as an indication of the reading ability required to understand the text (e.g., grade-level equivalents), the occupational-specific vocabulary should not be a matter of concern. However, it would be erroneous to assume that all material types (i.e., learning, examiniation, occupational) would include equal frequencies of occupational-specific vocabulary. Therefore, the failure to remove occupational-specific vocabulary from unfamiliar word totals would potentially result in inappropriate estimations of semantic complexity.

The next steps for the dental-licensing program are to collect and analyze sample sets of learning and occupational materials. The readability estimates for those materials should then be compared to the readability estimates of the examination items. Finding that the readability level of the examination materials is essential equivalent to the readability levels of the learning and occupational materials would offer the program an additional piece of validity evidence for their testing program. If essential equivalence is found between material types, the program would gain a degree of confidence that the readability level of the examination is such that undue construct-irrelevant variance is not

likely being introduced by the semantic or syntactic complexity of the items. If results indicate that the examination items are significantly more difficult to read than the learning or occupational materials, the program could take steps to amend future item-writing practices to help ensure that readability is addressed.

Limitations of the Current Study

The current study was constrained by obvious limitations. The first two phases of the investigation (i.e., Phase I: Usefulness of variables and Phase II: Formula calibration and recalibration) suffered from limitations related to the insufficient information provided by previous researchers, the use of a less than ideal set of calibration passages, and difficulties encountered during recalibration of existing formulas. The painstaking procedures required to implement the new-models in the third phase of the investigation presented further limitations. Moreover, some of the analysis results that were obtained during the third phase of the investigation (i.e., Phase III: External validity and reliability evidence) were questionable. These matters related to the limitations of the current investigation are discussed in turn in the following sub-sections.

Insufficient Information

Some of the research referenced during this study provided insufficient information to answer questions that came about during the investigation. In particular, Miller and Coleman (1967) did not provide the appropriate information to allow the use of their 36 passage for calibration purposes without referencing additional sources.

The calibration passages and their respective data were necessary to explore the variance accounted for by the semantic and syntactic variables under consideration, calibrate the new-model formulas, and recalibrate the existing formulas. It was difficult to

locate passages appropriate for calibrating equations; but it was much more difficult to locate the requisite data for the passages that were available. Miller and Coleman (1967) included the 36 passages they calibrated as an appendix to their study. They also included an abundance of data about those passages. However, they did not include corresponding cloze scores for the passages. Therefore, it was necessary to locate and reference a secondary source (i.e., Aquino, 1969) to obtain the cloze scores for Miller and Coleman's (1967) passages. The secondary source was relied upon for the total CT scores with some trepidation. It is possible that Aquino (1969) did not properly interpret or report these scores and it is not clear how they were obtained. Because the research of Miller and Coleman (1967) and Aquino (1969) was conducted more than forty- years ago, contacting the authors was not an option.

Appropriateness of Miller and Coleman Passages

Passages calibrated for level of readability according to cloze scores are not readily available. Therefore, the Miller and Coleman (1967) passages were the only viable option for this investigation. A few of Miller and Coleman's (1967) passages were written at a level appropriate for elementary-school students and were unsuitable for the purposes of this investigation. Specifically, four passages (1, 3, 10 and 15) were initially determined to be inappropriate for inclusion in the current study because they were the easiest of the passages according to their corresponding total CT scores. Although all 36 passages were initially investigated, the four passages with the highest total cloze scores were not included in the regression analyses conducted to calibrate the new-model formulas or recalibrate the most of the existing readability formulas that were retained for further investigation.

During the calibration of the new-model formula and recalibration of the existing formulas, additional passages tended to show high standardized residuals and their corresponding total cloze scores were not in accordance with their semantic- or syntactic-variable measures. These passages did not behave in this manner for every variable combination (i.e., formula) and were, therefore, removed when necessary to allow the relevant semantic and syntactic variables to enter the equation, to improve fit, and when the total cloze score and independent variable data did not correspond. Furthermore, the recalibration of one of the existing formulas required that all 36 passages be included. This resulted in slight differences in the passages that were used to calibrate the new-models and recalibrate the existing formulas that were retained for further investigation.

Passage 5 was not included in the calibration of any of the new models or the recalibration of the FOG3 or Homan-Hewitt formulas. Passage 31 was not included in the calibration of new-model formulas #C10 or CL8, nor was it included in the recalibration of the Dale-Chall formula. All 36 passages were included in the recalibration of the FOG1. The recalibration of the FOG2 was conducted with the removal of the four passages with the highest total CT scores; it was not necessary to remove any additional passages.

Although it would have been ideal to include exactly the same passages in the calibration and recalibration of all new-model and existing formulas, it was not possible. It was necessary to remove different passages for the different formula calibrations or recalibrations in order to allow all of the relevant variables to enter the equation and to address high residuals that were observed in some instances. It is not surprising that there was some variation between the passages that showed high standardized residuals in the

regression analyses conducted for the calibration of the new-models and recalibration of the existing formulas because they included different independent variables. Regardless, the regression analyses showed that passages 5 and 31 tended to misbehave for many of the new-model and recalibrated formulas.

A different set of calibration passages might not have required exclusion of different passages for the calibration of the new-models and recalibration of the existing formulas. A set of passages written at a higher grade level would likely have been more appropriate. Ideally, a set of passages would have been developed and calibrated with post-graduates. This would have offered more appropriate materials and corresponding cloze scores.

Recalibration of Existing Readability Formulas

A host of problems were encountered during the recalibration of the existing formulas. During the multiple regression analyses conducted to recalibrate the Dale-Chall formulas, it was difficult to find a solution that would hold both independent variable (i.e., sentence length and number of unfamiliar words). When all 36 passages were included and when the four passages with highest total CT scores were removed, the solutions did not include sentence length. It was necessary to remove an additional passage (31) to allow both variables to enter the equation.

The recalibration of the Dale-Chall formula resulted in different signs for the weightings of the predictor variables. Specifically, it required subtracting the weighting of number of unfamiliar words and adding the weighting for average sentence length. This was clearly inconsistent with what would be expected, because both of these predictor variables should have contributed to readability in the same way. The original

formula required the weightings of number of unfamiliar words and average sentence length to be subtracted in the equation. For the original Dale-Chall formula, low readability values indicate higher levels of readability or more complex text; therefore, subtracting these variable weightings are intuitive. When simple linear regression was used to analyze these variables separately, the signs were consistent with those of the original formula. However, when the variables were both included in stepwise multiple regression analysis, the resulting regression equation required the average sentence length weighting to be added, instead of subtracted. Even after the stepwise analysis was inspected and additional analyses were conducted, the reason for this inconsistency was unclear.

The recalibration of the FOG readability formula was problematic because, unlike the other existing formulas explored in the current study, the FOG formula is a linear equation but it is not a regression equation. The two independent variables, sentence length and percentage of hard words, are added and multiplied by a constant of .4. Due to the nature of this formula, a straight forward method of recalibrating it was not readily apparent. Because the original formula involved adding the two independent variables without weighting either of them, two approaches were used to recalibrate the formula, which resulted in three recalibrated versions of the FOG formula. First, the independent variables were entered independently and several multiple regression analyses were conducted with total CT scores as the dependent variable. Second, the independent variables were added together to create a single independent variable and simple linear regression was conducted with total CT scores as the dependent variable.

When a stepwise multiple regression approach was used, the equation would not hold both independent variables when the four passages with the highest total CT scores were not included. Removal of additional outliers did not allow both variables to enter. However, the equation did hold both variables when all 36 passages were included. The equation that resulted from the inclusion of all 36 passages was retained as the first recalibrated version of the FOG formula: FOG1.

Because the four passages with the highest CT scores were not included in the regression analyses conducted to calibrate and recalibrate the other formulas, additional regression analyses were conducted for the recalibration of the FOG in attempt to derive a solution that did not involve those four passages. Specifically, the four passages with the highest total CT scores were removed and hierarchical multiple regression was conducted to force both independent variables into the equation. The solution from this analysis was retained as the second recalibrated version of the FOG formula: FOG2.

For the last FOG recalibration analysis, the four passages with the highest total CT scores and outlying passage 5 were removed. The independent variables were summed and simple linear regression was conducted. The solution from this analysis was retained as the third recalibrated version of the FOG formula: FOG3.

At the conclusion of the recalibration analyses for the FOG formula, three recalibrated versions were created. The first, FOG1, included all 36 passages and was derived with stepwise multiple regression. The second, FOG2, included 32 passages, as the four passages with the highest total CT scores were removed, and was derived with hierarchical multiple regression. The third, FOG3, included 31 passages, as the four passages with the highest total CT scores and outlying passage 5 were removed, and was

derived with simple linear regression of the combined independent variables. Because it was unclear whether one recalibrated version of the formula was better than the others, all three versions were retained for further investigation. This resulted in the necessity of more analyses than were initially anticipated. Instead of conducting analyses for one recalibrated FOG formula, analyses had to be conducted for all three versions.

A great number of difficulties were encountered during the recalibration of the Homan-Hewitt formula. Homan et al. (1994) indicated that level 4 should be used to identify difficult vocabulary, or unfamiliar words, with *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) list of familiar words. However, using a stepwise multiple regression approach and identifying unfamiliar words at level 4 to recalibrate the Homan-Hewitt formula did not allow all independent variables to enter the equation. Therefore, additional stepwise multiple regression analyes were conducted in an attempt to identify a grade level for the semantic variable that would allow all independent variables to enter the equation. Regardless of the grade level at which the unfamiliar words were identified or the removal of outlying passages, none of the solutions derived with the stepwise approach would hold both semantic variables along with the syntactic variable. It was then clear that another method was necessary to allow all three variables included in the original formula to be included in the recalibrated version.

The results from the stepwise multiple regression analyses were inspected and several hierarchical multiple regressions were conducted in order to force all three independent variables into the equation in the order in which they entered during Homan and Hewitt's (1994) initial calibration. Grade levels 4, 6, and 8 were explored for the identification of

unfamiliar words. In the end, the recalibrated version of the Homan-Hewitt that was selected for retention and further investigation was that which incorporated the identification of unfamiliar words at level 8 and derived via hierarchical multiple regression with the passages with the highest total CT scores and outlying passage 5 removed.

Clearly, the recalibrated version of the Homan-Hewitt formula deviated from the original version in terms of the level at which unfamiliar words were identified. Using the same level of the semantic variable would have been ideal and was the original intent, but with both stepwise and hierarchical multiple regression approproaches, the use of level 4 did not allow all variables to enter the equation. A compromise was therefore necessary. It was more important for all three variables to enter the equation than it was for the level of the semantic variable in the recalibrated formula to exactly match the level used in the original formula.

The recalibration of the original existing formulas was far more difficult and time consuming than was anticipated. Because it was necessary for independent variables included in the original versions of the existing formulas to be included in the recalibrated versions, multiple approaches were necessary and far more analyses were conducted than was initially expected. In addition, in order to allow the requisite independent variables to enter the respective equations and to address standardized residuals, it was necessary to remove different passages for some of the recalibrated formulas. This might have affected the results obtained during the external validity and reliably portion of the investigation (Phase III).

Procedural Issues

A limitation of the new-models and the procedures required by them is the resource allocation necessary for proper implementation. An extraordinary amount of time and effort was required to obtain sample learning, occupation, and examination materials for analysis; convert examination items into pseudo-continuous prose; appropriately identify T-units and clauses of sample material sets; create an occupational-specific vocabulary list; and gain access to and use *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981). Furthermore, the input of a subject matter expert from the appropriate discipline was necessary at several points in the investigation. The following subsections include a discussion of the difficulties encountered during each of these steps in this investigation. Required input from subject matter experts is discussed for relevant steps. Then, alternatives to some of these steps that could require less time and effort are presented.

Collecting a Representative Sample of Examination Materials

Collecting samples from examination materials along with the requisite data can be especially challenging. Although many credentialing programs post retired examination items to their websites, they do not provide corresponding data for these items. Therefore, obtaining credentialing examination items and their corresponding data requires access to administrators of the respective credentialing program who are willing to share examination items and data. Because credentialing examinations are very expensive to create and are held in great confidence, credentialing programs are generally reluctant to share this information. Of course, retired items can be often be accessed through websites, yet without the relevant item-difficulty data an investigator would be unable to

ensure the collection of a representative sample. Specifically, an investigator would be incapable of ensuring that the examination items collected appropriately spanned the difficulty continuum. It is important that the collection include a concordant representation of items with high, medium, and low difficulty values.

To guarantee that a sample set of items drawn from an exam includes an appropriate representation of items at different difficulty levels, it is necessary to use stratified and systematic sampling. This process requires the items be sorted according to difficulty values and then divided into groups according to item difficulty (high, medium, low). The items within each group must then be unsorted or resorted within their respective stratum according to their identification codes. Every n^{th} item should then be selected for inclusion in the sample. Furthermore, the selected items must be converted into pseudo-continuous prose before they can be analyzed. The conversion procedures do not require nearly as much time and effort as identifying and collecting material samples, but it is still one more step than must be completed that requires additional time and effort.

Identifying syntactic characteristics of the sample materials.

The measurement of syntactic complexity by the new-models requires the identification of T-units and clause properties; whereas existing formulas typically require the identification of sentences properties. Because identifying T-units and clauses is not a straightforward and simplistic a task, training is required. Even with training, it is difficult to consistently identify T-units and clauses with precision. Therefore, if one of the new-models were to be implemented, it would be advisable to use multiple raters, all of whom would require hours of training. Inter-rater agreement should then be determined. The use of existing formulas requires only one rater and extensive training is

unnecessary to ensure accurate identification of sentence properties. Accordingly, the use of existing formulas is less demanding in terms of time and resources.

Creating an occupational-specific vocabulary list.

The new-models include the use of an occupational-specific vocabulary list to identify words or technical language in the texts that should be considered familiar to the respective audience. Such a list should be as exhaustive as resources will allow and must span the breadth of the discipline. For the current study, the list included nearly 5,000 words related to dentistry. Composing this list required accessing dozens of text book appendices and glossaries. Some, but not all, of these sources were available electronically and could be imported into word processing programs.

Furthermore, devising an occupational-specific vocabulary list for any credentialing program is best done with input from a subject matter expert. Such a person can recommend sources from which the words can be drawn or determine whether a list of sources collected by an investigator appropriately spans the discipline. Subject matter experts are not necessarily readily available or willing to advise an investigator and attempting such a task without their input would unadvisable.

Gaining access to and using The Living Word Vocabulary: A National Vocabulary Inventory (Dale & O'Rourke, 1981).

The list of familiar word used in the new-models, *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981), is not readily available for use. This book has been out of print for several years and is not located at public libraries or many university libraries. The book can be purchased, but it is rather expensive and very few copies are available for sale.

Gaining access to *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) for the current study was extremely difficult. After several attempts, the book was retrieved from other university libraries through inter-library loans, but those libraries did not offer renewal of the loan for consecutive months and the durations of the loans were insufficient to complete the necessary work. Subsequently, it was necessary to retrieve the book multiple times from different university libraries. It was important to use the same version of *The Living Word Vocabulary* throughout this investigation; therefore other versions of it were not accessed. However, similar versions of it appear to be available and might be easier to access through a university library.

With its corpus of 44,000 words, *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) offers an exceptionally comprehensive account of words that should be familiar at grades 4, 6, 8, 10, 12, 13, and 16. However, the book is over 700 pages long and using it with three sets of passages of approximately 150 words each can be daunting. Furthermore, because the list offers the grade levels at which different meanings of the same word should be familiar, it sometimes takes longer to identify whether a word should be deemed familiar or unfamiliar. In some instances the investigator must refer to the sample passages in order to determine the context in which a word is used and choose, from several very brief definitions, the appropriate grade level of familiarity.

It was unclear at the outset of this investigation how to best use *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) in conjunction with the occupational-specific word list. Prior to formula calibration, the grade levels at which the words in all of the calibration passages and sample materials were familiar was

identified with the use of *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981). Because the formulas were not yet calibrated, it was unclear which grade level would be used for the determination of word familiarity. Therefore, all grade levels were considered. *The Living Word Vocabulary* was used prior to consideration of the occupational-specific vocabulary list. This required that the semantic complexity data collected with the use of *The Living Word Vocabulary* be modified according to the occupational-specific vocabulary list. Specifically, the numbers of unfamiliar words that were identified with the use of *The Living Word Vocabulary* were altered to remove enumerations of words that existed in the occupational-specific vocabulary list.

Alternatives for applying the new-models.

It would likely be cost prohibitive for a credentialing program to implement the procedures required in the new-models. As it stands, the processes involved in the new-models would likely require several months to complete and would, therefore, be very expensive. However, some of the steps in the new-model could be modified to save time and effort. This abridgement of the process would still require input from subject matter experts, but credentialing programs have access to a great number of professionals who are sometimes willing to donate their time.

First, instead of creating an occupational-specific word list and using it to analyze the sample materials, subject matter experts could offer input regarding the sample materials. Specifically, the words in the passages would be converted to list form and presented to a number of subject matter experts. The subject matter experts would be asked to identify words that are specific to their field. The words identified by the different subject matter

experts would then be cross referenced. They would then discuss and come to a consensus about words for which their initial judgments did not concur. The final list of words identified as occupational-specific by the subject matter experts would then be identified as familiar in the passages. The words identified as occupationally specific and, therefore, familiar to the respective audience would not require further semantic-characteristic analyses with *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981).

This approach is far more time efficient for two reasons. First, it would be unnecessary to spend the time required to create an exhaustive occupational-specific vocabulary list, much of which would not be used. Second, the use of *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) would be far less daunting because a large number of the words that would have required grade-level-familiarity identification would have already been identified as occupational-specific vocabulary.

Questionable Results

During Phase III of this investigation, many analyses were conducted to collect external validity and reliability evidence to support the utility of the new-models. Some of the correlation analysis results failed to reveal relationships of the expected strength. In addition, all comparisons of readability estimates were subjected to very stringent significance criteria. This subsection includes discussions of these issues.

Correlation results for new-models.

Because it was apparent that the different ways in which occupational-specific vocabulary was treated was the culprit for the weaker than expected correlations between

the readability results derived with the new-model and recalibrated formulas, the occupational-specific vocabulary was used with the recalibrated formulas and the readability levels of the materials were again assessed. When the occupational-specific vocabulary list was used with the recalibrated formulas, the relationships between the readability estimates derived with the new-models and recalibrated formulas strengthened. However, these results were weaker than was expected in that they were moderate at best. It was presumed that the weaker than expected relationships observed after the incorporation of the occupational specific vocabulary list across all formulas, were due to both semantic and syntactic variables of the new-model and recalibrated formulas differing. In previous research, only the semantic variable tended to differ between formulas investigated. Analyses conducted for the syntactic-complexity measures of the examination materials according to the predictors used in the new-models and recalibrated formulas, however, revealed that indications of syntactic complexity did not differ between the new-models and recalibrated formulas and were nearly perfectly correlated. In essence, because of the nature of the examination materials, measures of T-unit and sentences were the same. It then becomes impossible to conclude that the weaker than expected relationships between the readability estimates of the examination materials, as determined according to the new-model and recalibrated formulas, resulted from both measures included in the formulas differing.

It was possible, however, that some of the correlations were weaker than expected because the Dale-Chall and FOG formulas were not designed for use with materials such as those investigated here. More specifically, the Dale-Chall and FOG were designed for use with several passages comprised of 100 words. The examination items, even after

conversion to pseudo-continuous prose, tended to include fewer than 100 words. Furthermore, only a single estimation was possible for each item.

Failure to find significant differences.

This portion of the limitations section includes discussions about concerns regarding analyses that resulted in a failure to find significant differences between formula results. First, the matter of stringent alpha levels is discussed. Then the possibility that power was insufficient in the current investigation are discussed.

Non-parametric analysis methods were used to compare the readability estimates derived with the different formulas. A Bonferroni correction for familywise error was used to adjust alpha for each comparison. Because so many comparisons were required, the use of the Bonferroni method resulted in very stringent criteria for significance. It might be argued that where significant differences were not observed, the extremely conservative alpha level was responsible. Inspection of the comparisons, as discussed in the results section, did not reveal this to be the case. Specifically, where the readability estimates derived with new-models and recalibrated formulas were compared to one another and differences did not reach significance, they would still not have reached significance if alpha had been set at .01. Furthermore, differences for some of the formula pairs would not have been significant even if alpha had been set at .05. Therefore, where the most important comparisons were concerned, the strict alpha level was not responsible for the failure to find significant differences.

It was important to address the stringent alpha levels; however, this issue is not as relevant as it might first appear. Specifically, although the alpha levels were stringent; they were consistently stringent in the comparisons made for combined Books 1 and 2,

Book 1, and Book 2. Therefore, that they were set at very low levels was not a matter of concern. The within-material-set comparisons were conducted to collect information about the performance of the formulas. The results of the within-material-set comparisons were then compared across sets to determine if the same differences were observed for combined Books 1 and 2, Book 1, and Book 2. The ultimate objective was to determine whether a consistent pattern of differences was observed when the formulas were applied to different types of materials.

It was not possible to determine power necessary to for the current investigation because there was no way to estimate potential population effect sizes for the comparisons. Therefore, it was unclear what would suffice an adequate sample size. It is possible that power was insufficient in the current investigation. Therefore, significant difference may actually exist. This investigation, however, may provide the data necessary to conduct the appropriate a priori calculations to determine a suitable power level and corresponding sample size for future research.

Future Research

The findings of the current investigation indicate that, although promising, the new-models require further study. Specifically, the semantic- and syntactic-complexity measures included in the new-models appear to be valid indicators of readability for credentialing-examination materials, but further calibration or validation studies should be conducted. The following section includes a discussion of several approaches that might be taken in future research. Each will be discussed in turn. First, the issue of power and how greater power might lead to different findings in future research is briefly discussed. Second, a recommendation for studies involving different calibration passages

is discussed. Third, potential refinements of the procedures that were used in the current investigation to convert examination items into pseudo-continuous prose are addressed. Fourth, research ideas regarding the exploration of the new-models in their current forms are presented.

Power

Because data were not available to estimate potential population effect sizes for comparisons, a priori power analyses were not conducted for the current investigation. It was, therefore, impossible to determine an appropriate sample size or number of sample passages to be collected for examination materials. Some of the comparisons conducted during this study revealed non-significant differences, but it was possible that power was limited by sample size or the number of sample items that were included in the examination materials. Conducting an a priori power analysis would help ensure that an appropriate sample size is implemented for the credentialing materials as well as calibration passages.

The data collected during this investigation might be used in future research to conduct the appropriate a priori calculations to determine a suitable power level and corresponding sample sizes for the credentialing and calibration materials. It is possible that a priori power analyses will reveal that larger samples are necessary. It is further possible that larger samples would result in different findings when readability estimates derived with different formulas are compared.

Calibration Passages

Because calibration passages and their corresponding data are not readily available, a variety of options did not exist. The Miller and Coleman (1967) passages initially

appeared to be best suited for the purposes of the current investigation because they were calibrated according to cloze scores obtained from undergraduate students; whereas, other sets of calibrations passages were generally calibrated according to cloze scores obtained with students from grades K-12 (e.g., Bormuth, 1971). However, data obtained during the calibration of the new-models revealed that the Miller and Coleman (1967) passages were perhaps not ideal for the present purposes. Because four of the passages were written at such a low reading level, they were immediately removed from further analysis. Then, multiple passages continued to misbehave in terms of the correspondence between their total CT score and independent variable data. It was necessary to remove these passages as well.

Future research should be conducted with the same semantic and syntactic variables investigated in this study, because they show great promise, but that research should include a better-suited set of calibration passages. Ideally, such calibration passages would be written at a more sophisticated reading level than the Miller and Coleman (1967) passages. Correspondingly, the passages should be calibrated according to cloze scores obtained from participants assumed to have greater levels of reading ability than those who participated in Miller and Coleman's (1967) calibration study.

Investigators interested in continuing this research might approach the issue of the need for different calibration passages in one of three ways. It is necessary to consider the expected reading level of a respective credentialing program audience. The options presented here were developed with a post-baccalaureate audience similar to that of the dental licensing program in mind. First, a more sophisticated set of calibration passages that were calibrated with an audience of readers who were assumed to have greater

reading ability could be located and used, if such a set of passages exists. Second, an existing set of calibration passages could be collected and recalibrated with a new group of participants, who are assumed to have higher levels of reading ability. For instance, Bormuth (1971) calibrated a set of 32 passages related to academic topics. He extracted passages from biology, chemistry, civics, current affairs, economics, geography, history, literature, mathematics, and physics text books to create these passages. The breadth of content covered by these passages makes them an attractive option for future research. However, Bormuth (1971) calibrated these passages with cloze scores from students in grades 3-12. Therefore, the cloze scores for these passages are not ideal for the calibration of a readability formula suitable for post- baccalaureate level reading materials. Nevertheless, it is possible that these passages, or a similar set of passages, could be recalibrated with cloze scores obtained from post- baccalaureates or graduate-school students. This would require access to a participant pool that included graduate students.

Both the first and second alternatives are limited by the constraints they impose on a researcher. Specifically, a researcher would be bound with a sample size not of their choosing. One lesson learned during this investigation is that sample passages sometimes behave in unexpected ways. It would therefore be advisable that an investigator have the liberty to remove passages that misbehave. With a set of 32 passages (e.g., Bormuth, 1971), an investigator may not have the freedom to remove passages that are not contributing to their research.

The third and most arduous alternative would be to create and calibrate an entirely new set of calibration passages. This endeavor could be approached in a number of ways

and two possibilities are outlined here. First, samples could be collected from a variety of sources that cover diverse topic areas and be calibrated with post- baccalaureate or graduate-school students from any discipline. Second, calibration passages specific to a particular discipline (e.g., dentistry) could be created by extracting text samples from relevant textbooks and graduate-school students from a related program could be recruited for participation in the calibration process. This discipline-specific approach would be ideal if the readability formula to be calibrated were being designed exclusively for use with materials related to credentialing examinations in that field. However, the discipline-specific approach might hinder the generalizability of the newly created readability formula for use with materials related to different credentialing programs. Therefore, if the to-be-developed readability formula is being created for the purpose of measuring credentialing related materials in general, the first approach described is likely more appropriate.

In either option presented for the development and calibration of new passages, researchers would be at liberty to develop calibration passages related to whatever discipline they determined appropriate. They would also be able to recruit participants from whichever discipline and at whatever level of reading ability they determined to be suitable. Furthermore, researchers would enjoy the freedom to develop as many passages as they deemed fit for the purposes of their research. This, in turn, would allow them the luxury of removing passages during the calibration process, if necessary, without losing so much information as to thwart the calibration process.

Additional Samples: Learning and Occupational Materials

The current investigation was designed to develop a new-model to estimate the readability of materials related to credentialing programs. This new-model was intended to be suitable for learning, examination, and occupational materials. An obvious next step in collecting validity evidence for any of the new-models developed and investigated in this study would be to apply them to learning and occupational materials related to the same dental-licensing program. This would enable comparisons of a new-model and recalibrated formula results across all three sets of materials. Finding that the formulas perform consistently across the different types of credentialing materials would offer further evidence of the utility of the new-model for use with credentialing materials.

Further Investigation and Potential Refinement of Methods to Convert Examination Items into Pseudo-continuous Prose

The methods used in the current investigation to convert examination items into pseudo-continuous prose were developed by the current author as an adaptation of the methods used by Plake (1984). For the current investigation, these methods were devised with the purpose of transforming the non-continuous examination items that included many incomplete sentences and single terms as options, into texts that better resembled continuous prose. The line of logic incorporated during the development of this procedure was that it would be best if stems were used in conjunction with each corresponding option to create complete sentences appropriate for syntactical analyses. However, neither the author of the current investigation nor Plake have extensive background in text linguistics, text processing, or text comprehension. It is, therefore,

possible that the methods developed during the current investigation to convert examination items into pseudo-continuous prose could be refined.

Further research might be conducted regarding the most appropriate method of converting examination items into pseudo-continuous prose. This research should be conducted with attention to other relevant research regarding text linguistics, text processing, or text comprehension. An outgrowth of such research might be a better-developed set of conversion methods that result in pseudo-continuous prose that more strongly resemble more authentic prose. This would add value because it would facilitate accurate syntactic assessment of examination items.

Further Investigation of the New-models in their Current Forms

Although the new-models, in their current form, yielded readability estimates that were not as strongly correlated with readability estimates derived with recalibrated formulas as was expected, they might be worthy of further investigation. Accordingly, further external validity and reliability research might involve applying the new-models, in their current forms, and existing readability formulas to a different, yet similar set of sample materials. This would entail collecting sample learning, occupational, and examination materials related to a different credentialing program. Although the samples would be extracted from different sources, they should be at a reading level that could be reasonably assumed to be similar to that of the dental program materials that were examined in the current study. For instance, materials might be collected for a different health-care-industry licensing or certification program (e.g., pharmacist or physician assistant).

This approach would offer the opportunity to inspect relationships between the readability estimates derived with the new-models and existing formulas for an entirely different set of materials. It is possible that relationships of different strengths than were observed in the current study will be observed with new sets of materials. Furthermore, new rank orderings of formula results (new-model and existing formulas) would be obtained. The rank-ordering results could then be compared to those observed in the current investigation. Finding consistency between rank orderings determined in this study and future studies would offer some evidence that the formulas, in their current form, provide valid measures of credentialing materials that allow learning, occupational, and examination materials to be accurately sorted according to readability levels.

Another method of investigating the new-models in their current form might involve applying the new-models and existing formulas to materials that do not include occupation-specific vocabulary. With this approach, the new-models would not involve the use of an occupational-specific vocabulary list; instead, they would only involve the use of *The Living Word Vocabulary: A National Vocabulary Inventory* (Dale & O'Rourke, 1981) to identify unfamiliar words. The readability estimates derived from the new-models and existing formulas would then be compared.

This approach would be the converse of the methods used in the current investigation. In the current investigation, the readability estimates derived with the new-models and recalibrated formulas were compared. Then, the occupational-specific vocabulary list was used with the recalibrated formulas and the readability estimates derived with the new-models and recalibrated formulas were compared again. The strategy suggested for future research would offer information similar to that obtained when the occupational-specific

vocabulary list was used with the recalibrated formulas, but would approach the comparisons of the new-model and existing formulas from a different angle. Specifically, in the current investigation occupational-specific vocabulary list was added to recalibrated formulas to allow more consistent comparison of the new-model and recalibrated formulas. The future research suggested here would remove the consideration of occupational-specific vocabulary from the new-models and thereby offer a more consistent comparison of the new-models and existing, rather than recalibrated, formulas. If the results of the new-model and existing formulas corresponded well, it would support that the new-models measure readability in manner similar to well-established readability formulas. It would follow, then, that if the new-models were to include accommodations for occupational-specific vocabulary (i.e., reintroduce the use of occupational-specific vocabulary lists) and were applied to materials that included such vocabulary, they could reasonably be expected to perform in a fashion similar to how they did when occupational-specific vocabulary was neither included in the materials nor accounted for by the models.

Implementing any of the above-described research possibilities in conjunction with one another or independently would be a worthwhile endeavor. There is not yet sufficient evidence to warrant the use of the new-models to collect validity evidence for credentialing programs. However, the new-models, the variables they include, and the procedures they incorporate show great promise. Additional investigations should be conducted to either provide further evidence for the validity and reliability of the new-models in their current form or to recalibrate the new-models with the same independent variables.

APPENDIX 1

OCCUPATIONAL-SPECIFIC VOCABULARY LIST

%

1RAG-2

3b

5-Hydroxytryptamine

Abdomen

Abdominal

Abducens

Abducent

Abduction

Abductor

Ability

Ablation

Abnormal

Abnormalities

Abnormality

Abnormally

Abrasion

Abreviation

Abscess

Absence

Absent

Absolutely

Absorption

Abusers

Abutment

Abutted

Academy

Acaine

Accelerating

Accentuate

Acceptable

Access

Accessory

Accommodate

Accompany

Accumulates

Accumulation

ACE

Acellular

Acetazolaminde

Acetylcholine

Acetylcholinesterase

Achieved

Achilles

Acid

Acidic

Acine

Aciniform

Acinus

Acoustic

Acquired

Acrocephaly

Acromial

Across

Acrylic

Actin

Activate

Activated

Activation

Actively

Acupressure

Acupuncture

Acute

Acyanotic

Adamkiewicz's

Adapt

Adaptation
ADCC
Addiction
Adenocarcinoma
Adenohypophysis
Adenoids
Adequate
Adherens
Adhering
Adhes
Adhesion
Adhesive
Ading
Adiposum
Adjacent
Adjective
Adjunctive
Adjustments
Adjuvant
Administered
Administration
Admitted
Adolescence
Adopted
Adrenal
Adrenalin
Adrenergic
Adrenocortical
Adventitia
Adverse
Aenesthetic
Aesthetic
AF
Affect
Affected
Affecting
Afferent
Affinity
Agent
Aggregate
Agitation
Agnosia
Agonists
Agranulocyte
AHA
AICA
Aid
AIDS
Aims
Airborne
Airway
Akinetic
Al
Ala
Alaeque
Alar
Albumin
Albuterol
Alendronate
Alert
Alfentanil
Algenate
Aligned
Alignment
Allelic
Allelically

Allen’s
Allergic
Allergies
Allergy
Alloantigens
Allodynia
Allogeneic
Alloplastic
Allotype
Allotypic
Alloy
Alpert’s
Alpha
ALS
Alter
Alterations
Alternate
Alternative
Alveolar
Alveoli
Alveolitis
Alveolus
Alveus
Amalgam
Ambenonium
Ambient
Ambulatory
Ameloblast
Amelogenesis
Amelogenin
Amine
Aminergic
Amines
Amino
Aminophylline
Amiodarone
Ammon’s
Ammonis
Amnesia
Amobarbital
Amount
Amoxicillin
Ampicillin
Amplification
Amplify
Ampules
Ampulla
Ampullae
Amputation
Amygdala
Amygdaloid
Amyl
Amylase
Amyotrophic
Anal
Analgesia
Analgesic
Analogs
Analysis
Analyzed
Analyzes
Anaphase
Anaphylatoxins
Anaphylaxis
Anastomoses
Anastomosis

Anastomotic
Animal
Anteroposterior
Anatomic
Ankylosis
Anthrax
Anatomical
Anomalies
Antiallergy
Anatomy
Anomalous
Antianxiety
Andresenlines
Anomaly
Antibacterial
Anectine
ANS
Antibiotic
Anemia
Ansa
Antibodies
Anergy
Antagonist
Antibody
Anesthesia
Antecubital
Anticholinergic
Anesthetic
Antegonion
Anticoagulant
Anesthetics
Anteriolar
Anticonvulsant
Angina
Anterior
Antidotal
Angioblast
Anteriorand
Antidysrhythmic
Angioedema
Anteriorly
Antiemergence
Angiogenic
Anterior-posterior
Antiemetics
Angiography
Antero
Antigen
Angiotensin
Anterograde
Antigenic
Angle
Anteroinferior
Antigens
Angstrom
Antero-inferior
Antihelix
Angular
Anterolateral
Antihistaminic
Angulation
Anterolaterally
Antihypertens
Anguli
Anteromedial
Antihypoglycemic

Antihypoglyemic
Antiinflammatory
Antilirium
Antimicrobial
Antiplatelet
Antiseptic
Antisera
Antitragicus
Anulus
Anxiety
Anxiolytic
Anxiolytics
Anxious
Aorta
Aortic
Aorticopulmonary
AP
Aperture
Apex
Aphasia
Apical
Apically
Apicoectomy
Apnea
Aponeurosis
Apoptosis
Apoptotic
Apparatus
Appear
Appendage
Appliance
Application
Applied
Apply
Appointment
Appointments
Apposed
Appositional
Appropriate
Appropriately
Approximates
Apraxia
Apresthesia
Aqueduct
Aqueous
Ar
Arachnoid
Arangement
Arantius’nodules
ARAS
Arc
Arcade
Arch
Archform
Archicortex
Archwire
Arcuate
Area
Areola
Arise
Armamentarium
Arnold
Arod
Aroused
Arrange
Arrangement

Array
Arrest
Arterial
Arterioles
Arteriosus
Artery
Arthritic
Arthritis
Articular
Articulare
Articulate
Articulations
Articulator
Artificial
Ary
Aryepiglottic
Arytenoid
ASA
ASC
Ascending
Ascends
Ascmedulla
Asepsis
Asleep
Aspect
Aspiration
Aspirator
Aspirin
Assemble
Assemblies
Assessing
Assessment
Assimilation
Assistant
assistants
Assists
Associated
Association
Asters
Asthma
Asthmatic
Astral
Astrocyte
Asymptomatic
Asystole
Atarax
Ataxia
Ated
Atherosclerosis
Ation
Atlantic
Atlas
Atom
Atoms
Atopic
Atresia
Atria
Atrial
Atrioventricular
Atrium
Atrophied
Atrophy
Atropic
Atropine
Attach
Attached

Attachment
Attacks
Attempt
Attenuation
AUC
Audioanalgesia
Audiogram
Audiometer
Auditory
Auerbach's
Aural
Auricle
Auriculae
Auricular
Auricularis
Auriculo
Auriculotemporal
Autoimmune
Automatic
Autonomic
Autoreactive
Autoregulation
Auxiliary
Averaged
Avoid
Avoidance
Axial
Axilla
Axillary
Axis
Axon
Axons
Azithromycin
Azmacort
Azygos
Aα
Aγ
Aδ
Ba
Babinski's
Backward
Bacteremia
Bacteria
Bacterial
Bactericidal
Bacteriocidal
Bacteriostatic
Baillarger's
Band
Banding
Bankart's
Barbiturate
Barbiturates
Bari
Barr
Barret's
Barrier
Bartholin's
Basal
Base
Basfunctional
Basilar
Basioccipital
Basis
Basophil
Basophilic

Batson’s
Batteries
Battle’s
Bcl-2
Bcontaining
Beams
Beca
Becker
Beclomethasone
Behind
Bell’s
Bellini
Below
Bemoysis
Benadryl
Beneath
Beneficial
Benign
Benzodiazepine
Benzodiazepines
Bergmann
Bernouilli
Betz
Bevel
Bevelled
Beyond
Bezold’s
Bichat’s
Bicuspid
Bifurcating
Bifurcation
Bilateral
Bilaterally
Bile
Billroth’s
Bin
Bind
Biofeedback
Biologic
Biological
Biomaterials
Biomechanics
Biophysical
Biopsy
Biotransformation
Birbeck
Bisecting
Bisphosphonates
Bitartrate
Bite
Biteblock
Bitemark
Biteplane
Bitewing
Bizygomatic
Bjork
Blaschko
Blastocyst
Bleaching
Blocks
Bloodless
Bloodstream
Blotting
Bluish
Blunt
BMD

Bo
Bof
Bolton
Boltonmandibularbase
Boltonplane
Boltonstandardcorrelation
bond
Bonded
bonding
Bonds
Bones
Bony
Border
Bound
Boundary
Bow
Bowman’s
BOX
Braces
Brachial
Brachiocephalic
Brachium
Bracket
Bradycardia
Brainstem
Branch
Branchial
Branchiomeric
Breakaway
Bregma
Bretylium
Brevis
Brevital
Bridge
Brief
Briefest
Broad
Broca’s
Brödel
Brodmann
Bromide
bronchi
Bronchial
Bronchiectasis
Bronchiolar
Bronchiole
Bronchioles
Bronchodilation
Bronchodilator
Bronchomediastinal
Bronchospasm
Bronchus
Bruch’s
Bruise
Bruit
Brunner’s
Bruxism
BSC
Buccal
Buccinator
Buck
Bucy
Bud
Bulb
Bulbar
Bulbourethral

Bulging
Bulk
Bumper
Bundle
Bup
Burning
Bursa
BursaFabricius
Bypass
Bystander
C
C1
C1q
C2
C3
C3
C3–7
C3a
C3b
C4
C4b2a
C5
C5a
C5b
C6
C7
C8
Ca
Cable
Cad
Caecum
Cajal
Calcaneal
Calcarine
Calcification
Calcified
Calcium
Calcospherite
Calculi
Calculus
Callosal
Callosomarginal
Callosum
Calot’s
CAM
Canal
Canaliculi
Canaliculus
Canals
Cancellous
Cancer
Cancerous
Canine
Canoe
Canten
Cantilever
Cantilevered
Cap
Capable
Capacitance
Capacity
Capillaries
Capillary
Capitis
Capping
Capsule

Carabelli's
Carbamazepine
Carbohydrate
Carcinoma
Cardiac
Cardinal
Cardiopulmonary
Cardiorespiratory
Cardiotonic
Caretaker
Caries
Cariogenic
Carious
Caroticotympanic
Carotid
Carpenter's
Carrier
Cartilage
Cartilaginous
Cas
Cascades
Caspases
Cast
CAT
Catalytic
Catalyzes
Cataract
Catechol
Catecholamine
Categories
Cathelicidins
Catheter
Cauda
Caudal
Caudally
Caudate
Causal
Causative
Cauterize
Caution
Cautious
Cava
Caval
Cavernous
Cavitation
Cavity
CB
CC
CCP
Cd
CD system
CD117
CD11a
CD152
CD18
CD19
CD2
CD21
CD28
CD29
CD3
CD36
CD4
CD4 T
CD40
CD62E

CD62P
CD79a
CD79b
CD8
CD80
CD81
CD86
CDR
Cecum
Cefadrozil
Cefazolin
CEJ
Cell
Cellular
Celontin
Cemental
Cementation
Cemented
Cementicles
Cementoblast
Cementocyte
Cementoid
Cementum
Center
Centers
Central
Centralis
Centrally
Centric
Centriole
Centromedian
Centromere
Centrum
Cephalexint
Cephalic
Cephalogram
Cephalometer
Cephalometric
Cerac
Ceramic
Ceramics
Cerebellar
Cerebelli
Cerebellomedullary
Cerebellum
Cerebral
Cerebri
Cerebrospinal
Cerebrum
Cervical
Cervicalis
Cervicis
Cervix
Cessation
Chain
Chamber
Change
Channel
Chapters
Characteristic
Characterized
Charcot's
Charging
Charting
Chassaignac's
Chemical

Chemokines
Chemokinesis
Chemotactic
Chemotaxis
Chemotherapy
CHF
Chiari
Chiasm
Chiasmatic
Chlor
Chloral
Chlordiazepoxide
Chlorhexidine
Chloride
Chlorpheniramine
Chlorpromazine
Cholestatic
Cholinergic
Choloxin
Chondroblasts
Chondroglossus
Chorda
Chorionic
Choroid
Choroidal
Chromatids
Chromatin
Chromosomal
Chromosome
Chronic
Chyli
Ciclosporin
Cilia
Ciliary
Cilium
Cinereum
Cingular
Cingulate
Cingulum
Circuinvolving
Circuit
Circular
Circulation
Circulatory
Circuleading
Circumferential
Circumflex
Circumpulpal
Circumstances
Circumvallate
Circumventricular
Cistern
Cisterna
Cisterns
c-Kit
Clamping
Clara
Clarithromycin
Clarke’s
Clasp
Class switching
Claudius
Claustrum
Cleavage
Cleave
Cleft

Cleland's
Clenching
Cleoid
Clindamycin
Clinical
clinically
Clinicians
Clinoid
Cloaca
Clonal selection
Clonazepam
Clones
Clonidine
Clonus
Cloquet's
Closure
Clot
Clusters
CMD
CMI
CN
CNS
Coarctation
Coats
Cobalt
Cochlea
Cochlear
Cocktail
Coded
Coefficient
Coeliac
Coenzyme
Cognition
Cognitive
Coherent
Coils
Coincident
Colic
Collagen
Collagenase
Collapse
Collateral
Collect
Collectively
Colles'
Colli
Colliculus
Colon
Colony
Colorless
Column
Columnar
Coma
Combination
Combine
Command
Commissural
Commissure
Common
Commonly
Communication
Compact
Compartment
Compazine
Compensates
Compensatory

Complaining
Complement
Complementarity
Complementary
Completing
Complex
Complexes
Complicated
Complication
Component
Composed
Composite
Composites
Compound
Comprehension
Comprehensive
Comprehensively
Compress
Compression
Compressor
comprises
Comprising
Compromised
Compute
ConA
Concanavalin
Concave
Concavity
Concentration
Concern
Condensation
Condition
Conduct
Conduction
Conductive
Conductivity
Conduit
Condylar
Condyle
Condylion
Cone
Configuration
Confluent
Confused
Congenic
Congenital
Conglutinin
Coniotomy
Conists
Conjugate
Conjunction
Conjunctival
Connect
Connective
Connector
Conner
Connexon
Conniventes
Conscious
Consciousness
Consensual
Consensus
Consent
Consented
Consequence
Consider

Considerably
Consideration
Considerations
Consistent
Consists
Consolidation
Conspicuous
Constant
Constantly
Constipation
Constitutes
Constitution
Constrict
Constriction
Constrictor
Construct
Construction
Consultation
Contact
Container
Containing
Contains
Contamination
Continuation
Continue
Continuing
Continuous
Continuously
Contour
Contracted
Contractile
Contractility
Contracting
Contraction
Contracture
Contraindicated
Contraindication
Contralateral
Contrast
Contribute
Controlled
Conus
Conventional
Convergence
Conversion
Convertase
Converting
Convexity
Convey
Convolute
Convolution
Cooper's
Cooperate
Coordination
COPD
Cord
Cordis
Cornea
Corneal
Cornu
Corona
Coronal
Coronally
Coronary
Coronoid
Corpus

Corpuscle
Correct
Corrections
Correlation
Correspond
Corresponding
Corrugator
Cortef
Cortex
Corti
Cortical
Corticobulbar
Corticofugal
Corticopontine
Corticospinal
Corticosteroid
Corticosteroids
Cortiscosteroid
Cortisol
Cortisone
Cosis
Cosmetic
Costal
Co-stimulation
Co-stimulatory
Costocervical
Cough
Couinard's
Counterbalanced
Coupled
Coupling
Course
Covalently
Cowper's
CR1
CR4C1qR
Cramp
Cranial
Craniofacial
Craniometric
Craniometry
Craniostat
Cranium
Crease
Crepitus
Crest
Cribriform
Crico
Cricoid
Cricothyroid
Cricothyrotomy
Crisis
Crista
Critical
Cromoglycate
Cromolyn
Crooked
Cross
Crouzon's
Crowded
Crowding
Crown
Crowning
Cruciate
Crura
Crus

Crypt
CSF
CSOM
CT
CTL
CTLA-4
CTLs
Cuboidal
Cuff
Cullen’s
Culture
Cuneate
Cuneatus
Cuneiform
Cuneocerebellar
Cuneus
Cur
Curative
Curettage
Curing
Current
Currently
Curvature
Curve
Curvilibridge
Cushing’s
Cushingoid
Cushion
Cusp
Cuspal
Cuspid
Cutaneous
Cutdown
Cuticle
Cuticular
Cutter
Cuvier’s
CXC group
Cyanotic
Cycle
Cyclopropane
Cylindrical
Cyst
Cystic
Cytokine
Cytokineact
Cytokines
Cytometry
Cytoplasm
Cytoplasmic
Cytosol
Cytotoxic
Cytotoxicity
Cytotrophoblastic
D
Dacryon
DAF
Damage
Dangerous
Darkschewitsch
DB
DD
DDS
Deafferentation
Debanding
Debonding

Debridement
Debris
Decadron
Decalcification
Decay
Decibel
Deciduous
deciduoustooth
Decomposition
Decrease
Decrement
Decussating
Decussation
Deepen
Defect
Defective
Defensins
Defibrillator
Deficiency
Deficit
Definite
Definitive
Deflection
Deformity
Degradative
Degrade
Degranulation
Degree
Degrees
Dehiscence
Deiters
Dejerine
Del
Delayed
Delineated
Delirium
Delta
Deltoid
Demarcating
Demerol
Demilune
Demineralization
Demonstrate
Demonstrates
Dendrite
Dendritic
Denonvillier's
Denotee
Dense
Density
Dental
Dentally
Dentate
Denticles
Denticulate
Dentin
Dentinal
Dentine
Dentinoenamel
Dentinogenesis
Dentist
Dentistry
Dentition
Dentitions
Dentoalveolar
Dentofacial

Dentoform
Denture
Dentures
Deoxygenated
Deoxyribonucleic
Depakene
Dependence
Dependent
Depolarization
Depolarizing
Deposit
Deposition
Depressant
Depressed
Depression
Depressor
Depth
DeQuervain's
Der
Derangement
Derivative
Derive
Dermatitis
Dermatomes
Dermatomyositis
Dermis
Descemet's
Descending
Described
Desiese
Designate
Designation
Designed
Desirable
Desmosome
Destined
Destroy
Detachment
Detailed
Detect
Detectable
Detected
Detection
Determinants
Determination
Determine
Determined
Develop
Developed
Development
Developmental
Develops
Device
Dexamethasone
Dextrose
Dextrothyroxine
Diabetes
Diabetic
Diageticorum
Diagnosed
Diagnoses
Diagnosis
Diagnostic
Diagonal
Diagram
Dialated

Diameter
Diamox
Diaphragm
Diaphragmatic
Diaphysis
Diastema
Diazepam
Diazoxide
Diencephalon
Dieretics
Differentiate
Differentiation
Difficulty
Diffuse
Diffuses
Diffusion
Diffusional
Digastric
Digestion
Digitations
Digoxin
Dilantin
Dilatation
Dilate
Dilator
Dilution
Dimensions
Diminish
Diminution
Dimple
Diphenhydramine
Diphyodont
Diplopia
Direct
Direction
Directional
Disability
Disappeared
Disarticulated
Discarded
Disclaimer
Disclosing
Discoid
Discolouration
Disconnection
Discussed
Discusses
Disease
Diseased
Disequilibrium
Disgusting
Disinfectant
Disinfection
Disinhibition
Disk
Dislocation
Disorder
Disorientation
Disparate
Displacement
Disposable
Dissé
Dissociate
Dissociation
Dissolve
Distal

Distance
Distended
Distinct
Distinguish
Distribution
Distrie
Disturbance
Diuretics
Diverge
Divergence
Diverse
Diverticulum
Divide
Division
Dizzy
DM
DMD
DNA
Dobutamine
Document
Domain
Dominant
Donor
Donut
Dopamine
Dopaminergic
Dorsal
Dorsalis
Dorsi
Dorsiflexion
Dorsolateral
Dorsomedial
Dorsomedially
Dorsum
Dosage
Dose
Douglas
Downgrowth
Downs
Drain
Drainage
Dramatic
Draped
Dreifuss
Dressing
Drift
Droperidol
Drummond
DTH
Duchenne
Ducts
Ductules
Ductus
Dullness
Duodenal
Duodenum
Dupuytren's
Dura
Dural
Duration
Dyes
Dyphylline
Dysfunction
Dysostosis
Dysplasia
Dyspnea

Dysrhythmias
Dystrophy
E
E/anesthetic
Ebstein's
ECG
Ectodermal
Ectomesenchyme
Ectopia
Ectopic
Eczema
EDDI
Edema
Edentulous
Edge
Edinger
Edition
EDMD
Edrophonium
Effect
Effectively
Effector
Efferent
Effort
Eg
Eicosanoids
Elastic
Elasticity
Electrical
Electroanesthesia
Electrocardiogram
Electrocardiograph
Electrocardioscope
Electromechanical
Electronic
Electropaste
Electrosedation
Electrotonic
Eleidin
Element
Elements
Elevated
Elevation
Eliciting
Eliminate
Eliminates
Eliminating
Elimination
Ellipsoid
Elongated
Embedded
Embolism
Embryologically
Embryonic
EMD
Emerge
Emergence
Emergency
Emery
Emesis
Emigrate
Eminence
Emissary
Emission
EMLA
Emotion

Emotional
Emphysema
Emulsion
Enable
Enamel
Enamelin
Enameloid
Encapsulates
Encircle
Enclose
Encoded
Encompasses
Encountered
Encourage
Ended
Endings
Endo
Endocardial
Endocarditis
Endochondral
Endocrine
Endocytosis
Endoderm
Endodontic
Endodontics
Endodontist
Endogenous
Endolymphatic
Endometrium
Endomysium
Endonasal
Endoneurium
Endoplasmic
Endosseous
Endosteal
Endothelial
Endothelium
Endothermal
Endothermalic
Endotoxin
Endotracheal
Endplate
Eness
Enflurane
Enhance
Enhancement
Eniculate
Enlarged
Enlargement
Ensheath
Ensure
Entails
Enteric
Entering
Entity
Entoderm
Entorhinal
Envelope
Enzymatic
Enzyme
Enzymes
Eosinophilia
Eosinophilic
Eosinophillic
Eosinophils
Ependyma

Ependymal
Ephedrine
Epicondyle
Epicranius
Epidemiological
Epidermal
Epidermis
Epidural
Epiglottic
Epiglotus
Epilepsy
Epileptic
Epimysium
Epinephrine
Epineurium
Epiphyseal
Epiphysis
Epiploic
Episcleral
Epithalamus
Epithelia
Epithelial
Epithelioid
Epithelium
Epitope
Eponymous
Eponyms
Eposteal
EPSP
Equatorial
Equilibrium
Equina
Equipment
Equivalent
ER
Erb
Erected
Erector
Ergic
Erosion
Erupt
Eruption
Eruptive
Erythrocyte
Es/anesthetics
E-selectin
Esmolol
Esophageal
Esophagus
Essential
Establish
Esthetic
Estosterone
Etch
Ether
Ethmoid
Ethmoidal
Ethmosphenoid
Ethosuximide
Euphemistic
Eurosyphilis
Eustachian
Euthyroid
Evaginations
Evaluate
Evaluation

Event
Eventually
Evidenced
Evident
Evoke
Exam
Examination
Examiners
Examining
Exceed
Excess
Excessive
Exchange
Excision
Excitation
Excitatory
Exclude
Exclus
Exclusion
Exclusive
Excreted
Excretion
Exerts
Exfoliat
Exfoliate
Exhibit
Exist
Existence
Existing
Exists
Exocrine
Exocytosis
Exogenous
Exon
Exothermal
Exothermalic
Expand
Expander
Expanse
Expansion
Expected
Expediently
Explorer
Expose
Exposed
Exposure
Express
Expression
Extend
Extension
Extensive
Extensively
Extensor
Extent
External
Externally
Extracellular
Extract
Extracted
Extraction
Extrafusal
Extraocular
Extraoral
Extrapyramidal
Extravasated
Extravascular

Extreme
Extremity
Extrinsic
Extrusion
Eyeball
Eyelid
F
Fabricated
Fabrication
Facebow
Facial
Facialangle
Facilitated
Facioscapulohumeral
Factor
Factors
FADI
FAGD
Fainting
Fallopian
Fallopio
Fallot
Falx
Fanning
Fascia
Fascial
Fascicles
Fasciculus
Fastening
Fastigial
Fatality
Faucial
Favored
FDC
Feature
Feeder
Feel
Feil
Fellow
Femoral
Fenestrated
Fenestration
Fentanyl
Fetal
Fetus
Fever
FH
Fiber
Fibre
Fibres
Fibrillation
Fibrinolytic
Fibroblast
Fibroblastoclasts
Fibroblasts
Fibrosis
Fibrous
Field
Figure
Filaments
Filiform
Filling
Film
Filtration
Filtrum
Filum

Fimbria
Fin
Findings
Fine
Fingerbreadth
Firmly
Fissure
Fistula
Fitted
Fixation
Fixed
Flap
Flare
Flavour
Flavoured
Flexed
Flexing
Flexion
Flexor
Flexure
Flocculonodular
Flocculus
Flomerulonephrosis
Floppier
Flora
Flouride
Flow
Fluent
Fluid
Flumazenil
Flunitrazepam
Fluorescence
Fuoridation
Fluoride
Fluorosis
Flush
FMA
FMRI
Focus
Focusing
Foliate
Follicle
Follicular
Following
Follow-up
Fontanelle
FOP
Foramen
Foramina
Force
Forceps
Forcibly
Fordyce's
Fore
Foreactivity
Forearm
Forebasal
Forebrain
Foreign
Forel’s
Foreroughly
Forgiving
Form
Formal
Formalin
Formation

Formina
Forming
Formyl
Fornix
Forward
Fossa
Fovea
Foveae
Fraction
Fracture
Fractured
Fractures
Fragment
Frankfort
Franulomas
Frenum
Frequency
Frequently
Freund's
Frey’s
Fringe
Froehse
Frontal
Frontobasal
Frontonasal
Frontotemporale
Froriep’s
Frustrated
FSH
FSHD
Full-Mouth
Function
Functional
Functionally
Functioning
Functionless
Fundamental
Fundic
Fungiform
Funicular
Funiculi
Funiculus
Furcula
Furosemide
Furunculosis
Fuse
Fusiform
Fusimotor
Fusion
G
G1
G2
GABA
Gadget
Galen
Galeni
Galenic
Gallamine
Gallbladder
Galli
Gallstone
GALT
Gametes
Gamma
Ganglia
Ganglion

Ganglionic
Gangrenous
Gartner’s
Gary
Gastric
Gastrointestinal
Gauze
Gehrig's
Gel
Gelatinous
Gene
General
Generalized
Generated
Generator
Genetic
Genetic
Genetically
Geniculate
Geniculocalcarine
Genioglossus
Geniohyoid
Genital
Gennari’s
Genome
Genotype
Genu
Gerdy’s
Geriatric
Germinal
Germline
Germs
Gerota’s
Gestation
Giacomini’s
Gingiva
Gingivae
Gingival
Gingivectomy
Gingivitis
Gingivoplasty
Ginglymoarthrodial
Girdle
Gl
Glabella
Gland
Glands
Glandular
Glaucoma
Glenoid
Glhabenula
Glia
Glisson’s
Globular
Globus
Glomerular
Glomerulonephritis
Glomerulus
Glomus
Glossopalatine
Glossopharyngeal
Glossus
Glucagon
Glucocorticosteroid
Glue
Glutamate

Gluteus
Glycine
Glycopyrrolate
Glycosylphosphatidylinositol
Gnarled
Gnathion
Goethe
Goiter
Golgi
Gonion
Gonorrhea
Goodpature's
Gopharyngeal
Gprotein
Graafian
Gracile
Gracilis
Gradient
Graduated
Graft
Granular
Granulation
Granule
Granulocyte
Granulomatous
Grapevine
Grayson's
Grinding
Groin
Groove
Growth
GTP
GTR
Guage
Guanethidine
Guardian
Guaze
Gubernacular
Guedel
Guerin's
Gum
Gurgling
Gustatory
Guyon's
GVHD
Gyri
Gyrus
H
H-2
Habenula
Habenular
Habenulointerpeduncular
Haemorrhage
Halitosis
Haller's
Halothane
Halves
Haplotype
Hapten
Hardens
Harmoniously
Harris's
Hartmann's
Haversian
Hayfever
Headgear

Heavily
Helical
Helicis
Helix
Hemagglutination
Hematoma
Hematopoietic
Hematoylin
Hemidesmosome
Hemispheral
Hemisphere
Hemoglobin
Hemolytic
Hemophilia
Hemorrhage
Henle’s
Hensen
Hepatic
Hepatitis
Hering’s
Hering–Brewer
Hernia
Herniation
Herpes
Hertwig's
Heschl’s
Heterogeneous
Heterologous
Heuser’s
HGF
Hguanine
Hiatus
Hillocks
Hilton’s
Hilus
Hindbrain
Hindgut
Hinge
Hippocampal
Hippocampus
Hirschsprung’s
Histamine
Histocompatibility
Histologic
HIV
Hoboken’s
Hofbauer
Holden’s
Holder
Homeostasis
Homocytrotopic
Homologous
Hook
Horizontal
Horizontally
Hormone
Horn
Horner’s
Horseshoe
House
Howship’s
Humeral
Humerus
Humor
Humoral
Humour

Humphrey's
Hunt
Hunter's
Huschke
Huxley's
Hyaline
Hyalinization
Hyaloid
Hyaluronic
Hyaluronidase
Hybridoma
Hydrate
Hydrocephalus
Hydrochloride
Hydrocortisone
Hydrodynamics
Hydrogen
Hydrolysis
Hydrolyzes
Hydroxyapatite
Hydroxylation
Hydroxytryptamine
Hydroxyzine
Hygiene
Hyoglossus
Hyoid
Hypdroxyquin
Hyperacusis
Hypercarbia
Hyperglycemia
Hyperkalemia
Hyperplasia
Hyperpolarization
Hypersensitivity
Hyperstat
Hypertens
Hypertension
Hyperthermia
Hyperthyroid
Hyperthyroidism
Hypertonic
Hypertrophy
Hyperventilation
Hypnotic
Hypoblast
Hypocalcified
Hypogal
Hypoglossal
Hypoglycemia
Hypohalites
Hypomaturation
Hypopharyngeal
Hypophyseal
Hypophysial
Hypoplasia
Hypoplastic
Hypotension
Hypothalamic
Hypothalamus
Hypothetical
Hypothyroid
Hypothyroidism
Hypoxia
Hyrdralazine
Hyrtl's
Iatrosedation

ICAM-3
Iccosomes
Id
Identical
Identifying
Idiotype
Ie
IFNs
IgA
IgD
IgE
IgG
IgM
Ii
IL-1
IL-22
Ileo
Ileum
Ilio
Iliocostalis
Illness
Illustrated
IM
Image
Images
Imaginary
Imaging
Imbalances
Imbrication
Immediate
Immediately
Immobilize
Immune
Immunity
Immunoassays
Immunoblotting
Immunodeficiency
Immunofluorescence
Immunogenic
Immunoglobulin
immunohistochemistry
Immunological
Immunoreceptor
Immunosuppressed
Immunosuppressive
Impacted
Impaction
Impaired
Impairment
Impar
Imperfecta
Imperfections
Impingement
Implant
Implantation
Implanted
Implants
Implementation
Implies
Impressions
Improper
Impurities
Ina
Inact
Inactive
Inanimate

Inanterior
Inatrium
Inbreeding
Incapacitating
Incerebellum
Inchoroid
Incidence
Incipient
Incirculation
Incisal
Incision
Incisive
Incisor
Incisure
Inclination
Include
Including
Inclus
Inclusion
Incompletely
Inconclusive
Inconspicuous
Inconstant
Increment
Incremental
Incrus
Indentation
Independently
Inderal
Index
Indicate
Indicators
Indirect
Individual
Induce
Inducible
Induction
Infarction
Infect
Infection
Infectious
Inferior
Inferioris
Inferiorly
Inferius
Infiltration
Inflamed
Inflammation
Inflammatory
Infolding
Informed consent
Infra
Infraclavicular
Infradentale
Infrahyoid
Infraorbital
Infraspinatus
Infratemporal
Infratrochlear
Infundibulum
Infusate
Infusion
Inhalation
Inherent
Inherited
Inhibit

Inhibition
Inhibitorsmatrix
Inhibitory
Inhypothalamus
Ininferior
Inion
Initial
Initiate
Inject
Injectables
Injected
Injuries
Injury
Inlateral
Inlay
Inlumen
Inmedial
Inner
Innervate
Innervation
Innocuous
Innominata
Innominate
Innovar
Inoccipital
Inoverlying
Inphagocytes
Inprimary
Input
Inquiry
Insect
Inseptal
Inserted
Insertion
Insomnia
Inspection
Instances
Instantaneous
Instraight
Instrument
Insufficiency
Insula
Insular
Insulin
Intact
Intake
Integral
Integrate
Integration
Integrins
Intended
Intense
Intensity
Interact
Interaction
Interactions
Interalveolar
Interatomic
Intercalated
Intercavernous
Intercellular
Interception
Interceptive
Interconnect
Intercostal
Intercuspation

Interdental
Interfering
Interferons
Interim
Interincisal
Interleukin
Interlocking
Intermedia
Intermediate
Intermediomedial
Intermedium
Intermingled
Intermolecular
Internal
Internalize
Internasal
International
Interneurones
Interneurons
Interneuronsly
Interoccipital
Interocclusal
Interosseous
Interpalatal
Interpeduncular
Interphalangeal
Interposed
Interpretation
Interproximal
Interradicular
Interruption
Intersection
Interspecies
Interspinal
Interstitial
Interstitialt
Interstitium
Intertransverse
Intervene
Intervention
Interventricular
Intervertebral
Intervillous
Interwoven
Intestinal
Intestine
Intima
Intraarterial
Intracellular
Intracoronal
Intracranial
Intraepithelial
Intrafusal
Intralobular
Intramedullary
Intramembranous
Intramuscular
Intranasal
Intraocular
Intraoperat
Intraoral
Intraperitoneal
Intrathecal
Intratubular
Intravenous
Intravenously

Intricate
Intrinsic
Introduction
Intropin
Intrusion
Intubation
Invade
Invariant
Invasion
Invasive
Invenous
Inverted
Investing
Involuntarily
Involuntary
Involved
Involvement
Inwall
Inward
Ion
Ions
Ipsilateral
Iris
Irradiated
Irregular
Irrigation
Irritate
Ischemia
Ischemic
Islets
Isolate
Isoproterenol
Isoproternol
Isotype
Isuprel
ITAMs
Ito
IV
J
Jackson's
Jacobson's
Jaundice
Jaw
Jelly
Joint
Joints
Jugular
Jugulo
Jugulodigastric
Junction
Junctional
Juxta
K
Kappa
Kartagener's
Karyotype
Kcell
Kent
Keratinized
Keratinocytes
Kerckring's
Ketamine
Kidney
Kiesselbach's
Killian's
Kilo

Kilogram
Kilograms
Kinases
Kinin
Kinocilium
Klippel
Klonopin
Klumpke’s
Klüver
Kn
Knee
Koch
Kohn
Kölliker
Kraissl’s
Kulchitsky
Kupffer
L
L1
L1–2
L2
L3
Labbé’s
Labeled
Labial
Labially
Labii
Labrum
Labyrinth
Labyrinthine
Lacerum
Lacrimal
Lacrimation
Lacunae
Ladd’s
LAKs
Lambda
Lamella
Lamellae
Lamina
Laminated
Landmark
Landmarks
Landouzy
Landsmeer’s
Langer’s
Langerhans
Langhans
Lanterman
Laryngeal
Laryngitis
Laryngoscope
Laryngospasm
Larynx
Lasix
Latency
Lateral
Lateralis
Laterally
Latex
Latin
Latissimus
Lattice
Layer
Layers
Lead

Leading
Leaveable
Lectin
Leeway
LeFort
Lemniscus
Lengthening
Lens
Lenticular
Lenticularis
Lenticulostriate
Lenunomtde
Lesion
Lesions
Lesser
Lethargy
Leucocytes
Leukemia
Leukocyte
Leukotrienes
Levator
Level
Levodopa
Leydig
LFA
LFA-2
LFA-3
LGLs
LGMD
Liability
Lidocaine
Lieberkuhn
Ligament
Ligand
Ligate
Ligating
Ligation
Ligature
Likelihood
Limb
Limbic
Limit
Limitans
Limitation
Limitations
Line
Lineage
Lined
Lingual
Lingula
Lining
Link
Linkage
Liotrix
Lipid
Lipoidica
Lipolysaccharide
Lipopolysaccharide
Liquid
Lisfranc's
Lissauer
Lister's
Little's
Littré's
Liver
Lobe

Lobule
Local
Localized
Locate
Location
Loci
Lockwood's
Locus
Lodging
Logan
Logarithmic
Longissimus
Longitudinal
Longitudinally
Longus
Loop
Loosened
Lorazepam
Lordotic
Lou
LPS
Lumbar
Lumborum
Lumen
Lunate
Lung
Lupus
Luschka
Lutea
Luted
Ly
Lymph
Lymphadenopathy
Lymphatic
Lymphocyte
Lymphocytese
Lymphoepithelial
Lymphoid
Lymphokine
Lymphoma
Lymphotoxin
Lysis
Lysosome
Lysozyme
Lytic
Lyze
M
M4
Mackenrodt's
Macrodontia
Macroglossia
Macrognathia
Macromolecule
Macrophage
Macroscopically
Macula
MAdCAM-1
MAGD
Magendie
Magill
Magna
Magnetic
Magnum
Maintain
Maintainer
Maintenance

Major
Mal
Malamed
Malar
Malassez's
Malbuphine
Malformation
Malignant
Malleus
Malocclusion
Malpighian
Malt
Mammalian
Mammillary
Management
Mandatory
Mandible
Mandibular
Manifestation
Manifestations
Manifested
Manikin
Mannan
Mannikin
Mannose
Mantle
Manually
Margin
Marginal
Marrow
Mass
Masses
Masseter
Masseteric
Massive
Mastership
Masticate
Mastication
Masticatory
Mastoid
Mater
Material
Materials
Maternal
Mathieu
Matrices
Matrix
Matter
Maturation
Maturational
Mature
Maxilla
Maxillae
Maxillary
Maxillofacial
Maxillomandibular
MBL
MC
McMinn's
MCP
Measure
Meatus
Mechanical
Mechanism
Mechanoreceptive
Meckel's

Medial
Medially
Median
Mediate
Mediators
Medical
Medication
Medicine
Medihaler
Mediobasal
Medius
Medulla
Medullaris
Medullary
Megacolon
Meibomian
Meiosis
Meissner’s
Melanin
Melanocytes
Membrane
Membranous
Memory
Meningeal
Meninges
Meniscofemoral
Meniscus
Mental
Mentalis
Menti
Menton
Mep
Meperidine
Mephentermine
Meprobamate
Mercury
Merkel
Merocrine
Mesantoin
Mesencephalic
Mesencephalon
Mesenchyme
Mesial
Mesoappendix
Mesoderm
Mesonephric
Mesonephros
Mestinon
Metabolic
Metabolism
Metabolites
Metal
Metalloproteases
Metaphase
Metaphysis
Metaprotereno
Metaproterenol
Metaraminol
Metastasis
Metatarsal
Methemoglobinemia
Methionyl
Method
Methods
Methohexital
Methotrexate

Methoxamine
Methoxyflurane
Methsuximide
Methyldopa
Meyer’s
Meynert
MHC
Microangiopathy
Microbial
Microdontia
Microglia
Microglobulin
Microglossia
Micrognathia
Microlamellae
Microorganism
microscopic
Microscopically
Microtubules
Microvascular
Microvilli
Mid
Midazolam
Midbrain
Midconnecting
midfacial
Midinferior
Midline
Midpons
Midreticular
Mids
Midsagittal
Midway
MIF
Migrate
Migration
Migratory
MIIC
Milli
Millogram
Mineralized
Minimae
Minimally
Minimization
Minimize
Minimus
Minor
Miscellaneous
Mitochondria
Mitochondrion
Mitogen
Mix
Mixture
MLF
MMD
MN
Modality
Model
Modification
Modifier
Modify
Modulate
Module
Modulus
Molar
Mold

Molecular
Molecule
Molecules
Moll’s
Monitor
Monitoring
Monocyte
Monolayer
Monolayers
Mononuclear
Mononucleosis
Monro
Montgomery
Morbidity
Morgagni
Moribund
Morison’s
Morphine
Morphologically
Morula
Mosby
Motifs
Motion
Motivation
Motor
Mound
Mount
Mouthguard
Movement
MP
MPD
MR
MRA
MRI
mRNA
Mucin
Mucoceles
Mucoceole
Mucogingival
Mucosa
Mucosae
Mucosal
Mucous
Müller’s
Müllerian
Multi-disciplinary
Multifidus
Multiforma
Multimodal
Multinucleated
Murine
Murmur
Murmurs
Muscarinic
Muscle
Musclelocal
Muscular
Muscularis
Musculocutaneous
Musculotubular
Musculus
Mutans
Mutate
Mutation
MX
Mycobacterium

Mycoses
Mycotic
Mydriasis
Myelin
Myelinated
Myelinating
Myeloid
Myeloma
Mylohyoid
Myoblast
Myocardial
Myoclonic
Myoepithelial
Myofacial
Myofibrils
Myometrial
Myopathies
Myosin
Myotendinous
Myotome
Myotonic
Mysoline
Mytelase
N
N2O
Na
Nabothian
Nalbuphine
Naloxone
Nalozone
Narcan
Narcotic
Naris
Narrow
Nasal
Nasalis
Nasi
Nasion
Naso
Nasociliary
Nasofrontal
Nasolacrimal
Nasopalatine
Nasopharyngeal
Nasopharynx
Nausea
Nearcortical
Nearinterneurons
Nearseptal
Nearventral
Necessary
Neckpad
Necrobiosis
Necrosis
Nembutal
Neo
Neocortex
Neonatal
Neonatorum
Neoplasm
Neostigmine
Nephritis
Nephron
Nephrosclerosis
Nerve
Nervous

Nervousness
Nests
Neural
Neuroanatomical
Neuroblasts
Neurocranium
Neuroendocrine
Neurogenic
Neuroglia
Neuroglial
Neurohypophysis
Neuroimaging
Neurological
Neuron
Neuronal
Neurones
Neuropeptides
Neurotransmitter
Neurovascular
Neutrophils
NF-k B
Nickel
Nicotinic
Night guard
Nigra
Nigral
Nisentil
Nitabuch's
Nitric
Nitrite
Nitroglycerin
Nitrolingual
Nitroprusside
Nitrostat
Nitrous
NK
NO
Nociceptors
NOD
Node
Nodule
Nodulus
Non
Nonapeptide
Nonbacterial
Nondrug
Nonfluorosis
Nonfunctional
Nonintravenous
Nonkeratinized
Nonkeratinocytes
Nonmalignant
Nonmyelinated
Nonpharmacologic
Nonrunning
Nonsteroidal
Nontelencephalic
Norepinephrine
Nose
Nostril
Notably
Notation
Notch
Novocaine
Noxious
Nregions

NSAIDs
Nuclear
Nuclei
Nucleolus
Nucleotide
Nucleus
Nude
Nuel
Numerical
Numerous
Nystagmus
O
O2
Obex
Obicularis
Objective
Objects
Oblique
Obliquely
Oblongata
Observation
Obstruct
Obstruction
Obstructive
Obtunded
Occasion
Occasional
Occipital
Occipitofrontalis
Occipitotemporal
Occlusal
Occlusion
Occupies
Occur
Occurrence
Occurs
Ocular
Oculi
Oculomotor
Oculopharyngeal
Oddi
Odontoblast
Odontoblastic
Odontoclast
Odontogenesis
Odontogenic
Odontoid
Odorless
Oesophageal
Oesophagus
Offs
Olfaction
Olfactory
Oligodendrocytes
Oligomerization
Olivary
Omohyoid
Onlay
Onset
Ontogeny
Op
Opacity
Openings
Operation
Operative
Operator

Operatory
Opercular
Ophthalmic
Opinion
Opioid
Opisthion
OPMD
Oppose
Opposing
Opposite
Opsonins
Opsonization
Opsonized
Opthalmus
Optic
Optimal
Optional
Optokinetic
Oral
Orbit
Orbital
Orbitale
Orbitalis
Orbitoethmoidal
Orbitofrontal
Organ
Organelles
Organic
Organism
Organized
Organs
Orient
Orifice
Origin
Originate
Oris
Oro-naso-optic
Oropharyngeal
Oropharynx
Ortex
Orthodontia
Orthodontic
Orthodontist
Orthognathic
Orthostatic
Oscillate
Osseointegration
Osseous
Ossicle
Ossified
Ostectomy
Osteitis
Osteoblasts
Osteoclasts
Osteocytes
Osteodentin
Osteonecrosis
Osteons
Osteoplasty
Osteoporosis
Osteotomy
Ostia
Ostium
Otalgia
Otic
Outer

Outermost
Outline
Outnumbered
Outpatient
Output
Outside
Outward
Outweigh
Ovale
Ovarian
Ovary
Overbite
Overdenture
Overdose
Overgrowth
Overhydration
Overjet
Overlap
Overlapping
Overlay
Overlies
Over-riding
Oversedation
Overturned
Oxidative
Oxide
Oximeter
Oximetry
Oxtriphylline
Oxygen
oxygenated
Oxygenation
Oxymoronically
Oxytalan
P
PABC
Pacchionian
Pacemaker
Pacinian
Paciniform
Pad
PAF
Paired
Pal
Palatal
Palate
Palatine
Palatini
Palatoglossus
Palatomaxillary
Palatopharyngeus
Palatovaginal
Paleocortex
Palliative
Pallidus
Pallor
Palmar
Palmer's
Palpate
Palpebrae
Palpebral
Palpitation
Palsies
Palsy
PAMP
Pancoast

Pancreatic
Pancreatitis
Pancytopenia
Paneth
Panoramic
Papez
Papilla
Papillae
Paracentral
Paracrine
Parahippocampal
Paralysis
Paramedian
Paramesonephric
Parameters
Parasitic
Parasympathetic
Parathyroids
Paratonsillar
Paratopes
Paratracheal
Paravaginal
Parenchyma
Parenchymal
Parenteral
Paresis
Parietal
Parieto
Parietooccipital
Parkinson's
Parotid
Partial
Partially
Participate
Particle
Particular
Partmucosa
Passage
Passavant’s
Passenger
Passing
Passive
Patch
Patency
Patent
Paternal
Pathogen
Pathogenic
Pathological
Pathologically
Pathology
Pathway
Patient
Patient's
Patterns
PC
PCA
Peaked
Pearls
PECAM
Pectineal
Pectoral
Pectoralis
Pectoris
Pediatric
Pedo

Pedodontist
Peduncle
Peduncular
Pedunculi
Pehenytoin
Pellicle
Pellucida
Pelvic
Penetrate
Penetration
Penicillin
Penile
Penten
Pentobarbital
Pentothal
Peptide
percent
Perforate
Perform
Periapical
Periaqueductal
Pericallosal
Pericardial
Pericytes
Periimplant
Perikaryon
Perikymata
Perilymphatic
Perimysium
Perinephric
Perio
Period
Periodic
Periodontal
Periodontist
Periodontitis
Periodontium
Periodontology
Periosteal
Periosteum
Peripheral
Periphery
Periradicular
Perisinusoidal
Peritoneal
Peritoneum
Peritubular
Perivascular
Periventricular
Permanent
Permit
Peroxide
Perpendicular
Persist
Personal
Pertaining
PET
Petit
Petrosal
Petrosquamous
Petrotympanic
Petrous
Peyer’s
PFC
Ph
Phagocyte

Phagocytic
Phagocytose
Phagocytosis
Phagosome
Phalangeal
Phalanges
Pharmacokinetics
Pharmacologic
Pharmacological
Pharmacologically
Pharmacology
Pharmacosedat
Pharyngeal
Pharyngotympanic
Pharynx
Phase
Phenacemide
Phenergan
Phenobarbital
Phenol
Phenomena
Phenomenon
Phenotype
Phenurone
Phenylephrine
Philtrum
Phospholipid
Phosphorylation
Phosphorylcholine
Photographic
Photographs
Photons
Photoreceptors
Phrenic
Physical
Physician
Physostigmine
Pia
Pial
Pigmented
Pin
Pineal
Pisiform
Pituitary
Placement
Placenta
Plane
Planing
Planning
Plantar
Plaque
Plasma
Plasmalemma
Plasmin
Plasmon
Plastic
Plate
Platelet
Plateletsa
Platysma
Plexus
Plier
PM
PNS
Pocket
Pog

Pogonion
Point-to-point
Polarization
Polarized
Polars
Polishing
Pollar
Pollicis
Pollutants
Pollution
Polygon
Polymeric
Polymorphonuclear
Polymorphs
Polypeptide
Polyribosomes
Polysomes
Pons
Ponsten
Pontic
Pontine
Pontomedullary
Population
Porcelain
Pores
Poria
Porion
Porionic
Porosity
Portal
Portion
Posit
Position
Positional
Positron
Possess
Possibility
Possible
Post
Postcentral
Postcommissural
Postcommunicating
Postcondylar
Posterior
Posteriorly
Posteroinferio
Posterolateral
Posteromedial
Posterosuperior
Postganglionic
Postoperat
Post-operative
Postoperatively
Postponed
Postprocessing
Postrema
Postsynaptic
Posttreatment
Postural
Potassium
Potential
Potentially
Pouch
Pr
Practice
Prazosin

Preal
Precapillaries
Precaution
Precede
Precentral
Precipitated
Precipitously
Precision
Precommissural
Precommunicating
Precuneal
Precursors
Predentin
Predictive
Prednisione
Prednisone
Predominate
Preeruptive
Preexisting
Preformed
Prefunctional
Preganglionic
Pregnancy
Pregnant
Prelaryngeal
Preloaded
Premature
Premaxilla
Premedication
Premolar
Premotor
Preoperative
Preoptic
Preparation
Prepared
Pre-processing
Prescription
Presence
Present
Presentation
Preserving
Pressure
Prestroke
Pretectal
Pretracheal
Prevent
Prevention
Prevertebral
Previous
Prilocaine
Primarily
Primary
Primed
Primidone
Primitive
Primum
Principal
Prior
Prismatic
Prismless
Privileged
Probability
Probe
Procainamide
Procaine
Procedure

Procedures
Procerus
Process
Processes
Processing
Prochlorperazine
Produce
Product
Production
Products
Profile
Progress
Progression
Progressive
Progressively
Project
Projectile
Projecting
Projection
Projectly
Proliferation
Proliferative
Prolongation
Prolonged
Promazine
Promethazine
Prominence
Prominent
Prominently
Promotes
Prone
Propagate
Propanolol
Proper
Property
Prophase
Prophylactic
Prophylaxis
Propofol
Proprandolol
Propranolol
Propria
Proprietary
Proprioceptive
Prostate
Prostatic
Prostheses
Prosthesis
Prosthetic
Prosthion
Prosthodontic
Prosthodontis
Prosthodontist
Protect
Protection
Protective
Protein
Proteins
Protocol
Protrude
Protrusive
Protuberance
Proventil
Provide
Provides
Prow

Proximal
Prudent
Pseudoephedrine
Pseudohypertrophic
Pseudostratified
Psychiatric
Psychological
Psychomotor
Psychosedate
Psychosedation
PT
Pterygoid
Pterygomaxillary
Pterygopalatine
PTM
Ptosis
Pt-vertical
Ptyalin
Pudendal
Pulmonary
Pulp
Pulpal
Pulpectomy
Pulpotomy
Pulsating
Pulse
Pupil
Pupillary
Pure
Purkinje
Purposefully
Pursuit
Putamen
Pyelonephritis
Pyramid
Pyramidal
Pyridostigmine
Q
Quadrant
Quality
Quantify
Quartz
Quiescent
Qv
R
Radial
Radiata
Radiating
Radiation
Radicular
Radio
Radiograph
Radiographic
Radiographically
Radiography
Radiolabeled
Radiological
Radiology
Radius
RAG
Rami
Ramsay
Ramus
Random
Randomize
Randomized

Range
Ranula
Raphe
Rapid
Rapidly
Raschkow
Rash
Rate
Rathke's
Rational
Rationally
Raw
Ray
RBC
Reabsorption
React
Reaction
Reactions
Reactive
Readily
Rebase
Rebound
Recall
Received
Recent
Receptor
Receptors
Recess
Recipient
Reciprocal
Recognition
Recognizable
Recognize
Recombinant
Recombination
Recombine
Recommend
Recommendations
Reconstruction
Recorded
Recovery
Rectal
Rectouterine
Rectum
Rectus
Recurrence
Recurrent
Recycled
Redistribution
Reduce
Reduction
Refer
Reference
Referred
Reflex
Reformulation
Regenerate
Regeneration
Regimens
Region
Regional
Registration
Regular
Regularly
Regulate
Regulation

Reidel's
Reinke's
Reissner's
Relate
Relation
Relationship
Relative
Relatively
Relaxant
Relaxation
Relay
Release
Relieves
Reline
Remaining
Remarkable
Remifentanil
Remnant
Remodeling
Removable
Removal
Remove
Remover
Renal
Rendered
Renshaw's
Renumeration
Repair
Reparative
Replace
Replacement
Replant
Replanted
Replication
Reposition
Represent
Reproduction
Require
RER
Resemble
Reserpine
Reserves
Reshape
Resident
Residual
Resin
Resistance
Resistant
Resonance
Resorb
Resorption
Respectively
Respiration
Respiratory
Respond
Response
Restimulated
Restoration
Restorations
Restorative
Restoring
Restrictaction
Restriction
Resulting
Results
Resurface

Retain
Retainer
Retardation
Retention
Reticular
Reticulum
Retina
Retinacular
Retinaculum
Retinal
Retinopathy
Retractors
Retro
Retroflexus
Retrograde
Retromandibular
Retropharyngeal
Retropubic
Retrorenal
Retrovascular
Retruded
Retzius
Reused
Reveal
Revealed
Reversal
Reverse
Reversible
Rexed's
Rhesus
Rheumatic
Rheumatism
Rheumatoid
Rheumatology
Rhinitis
Rhomboid
Rhythmic
Ribonuclear
Ribonucleic
Ribosomal
Ribosomes
Ridge
Rigid
Rigidity
Rims
Ring
Riolan
Risedronate
Risk
Risorius
RNA
RNI
Robertson
Robin
Robinul
ROC
Rod
Roentgen
Roentgenographic
Rohr's
ROI
Roilan
Rolando
Romazicon
Root
Rooted

Rootlets
Rosenmüller
Rosenthal
Rosettes
Rosetting
Rostral
Rostrally
Rostrum
Rotate
Rotation
Rotatores
Rotter’s
Rotundum
Rough
Roughly
Rounded
Route
Routine
rRNA
Rubber
Ruffini
Ruffled
Rugae
Ruptured
S
Sac
Saccade
Saccule
Saddle
SAdrenergic
Sagittal
Sait
Sal
Salicylate
Saliva
Salivary
Salivation
Salpingopharyngeus
Sanitization
Santorini
Sarachidonic acid
Sarcomeres
Sarcoplasm
Sassouni's
Satellite
Satisfaction
Saturation
SC
Scalenes
Scalenus
Scaler
Scaling
Scalp
Scalpel
Scan
Scaphoid
Scapulae
Scapular
Scarlet
Scarpa’s
Scavenger
SCF
Schedule
Schlemm
Schmidt
Schreger

Schutz's
Schwann
Scissors
Sclera
Scleroderma
Sclerosis
Sclerotic
Sclerotomes
Scoog's
Scopolamine
Scraped
Screening
Scrotum
SE
Sealants
Sebaceous
Secobarbital
Secondary
Seconds
Secrete
Secretion
Secretory
Section
Sectional
Sections
Secundum
Secure
Securely
Sedate
Sedation
Sedative
Segment
Seizure
Selectin
Selectively
Sella
Sellaturcica
Semicanal
Semicircular
Semiovale
Semispinalis
Sensation
Sensitive
Sensitivity
Sensitize
Sensory
Separate
Seperator
Sepsis
Septa
Septal
Septi
Septum
Sequence
Sequences
Sequentially
Series
Serotonin
Serous
Serrated
Serratus
Sertoli
Serum
Serumal
Serving
Sesconal

Severe
Severely
Severity
Shaft
Shallow
Shape
Sharpey's
Sheaf
Sheath
Shedding
Sheet
Sheetextensions
Shelf
Shunts
Si
Sialogram
Sialography
Sickle
Sigmoid
Sign
Signaling
Significant
Significantly
Simplex
Sinus
Sinusoid
Site
Sites
Situated
Situation
Sjogren's
Skeletal
Skeletally
Skeleton
Skull
Slated
Slender
Sliding
Slightly
Slot
SMA
Smallpox
Smear
Smooth
SN
Sneezing
Snoring
SO
Socium
Socket
Sodium
Sole
Solitary
Solu
Soluable
Solution
Soma
Somata
Somatic
Somatosensory
Somites
SOr
Sores
Sought
Source
Sp

Space
Spared
Sparing
Spastic
Spatial
Specialist
Specialized
Specially
Specialty
Specific
specifically
Specificity
Specify
Specimen
Spectral
Spee
Speech
Speed
S-Phase
Spheno
Sphenoccipital
Sphenoethmoidal
Sphenoid
Sphenoidal
Sphenomandibular
Sphenopalatine
Sphenoparietal
Sphincter
Spinae
Spinal
Spinalis
Spindle
Spine
Spinocerebellar
Spinosum
Spinothalamic
Spiral
Spite
Spleen
Splenic
Splenium
Splenius
Splint
Spoken
Sponges
Spores
Spring
Squamous
Stabilize
Stable
Stahl's
Stain
Staining
Stains
Standardize
Standardized
Stanley
Stapedial
Stapedius
Stapes
Staphylion
Staphylococci
Staphylococcus
STATs
Status
Steel

Steinert's
Stellate
Stem
Stenosing
Stenosis
Stensen's
Stereocilia
sterile
Sterilization
Sternal
Sternocleidomastoid
Sternohyoid
Sternothyroid
Sternum
Stethoscope
Stick
Sticky
Stimulate
Stimulation
Stimuli
Stimulus
Stitch
Stoma
Stomach
Stomatitis
Stomodaeum
Stomodeum
Straddling
Straight
Strains
Strand
Strands
Strap
Stratified
Stratum
Straw
Streak
Strength
Strengthen
Streptococci
Stress
Stretch
Stretchable
Stria
Striae
Striated
Striations
Striatum
Strictly
Stripper
Stripping
Strof
Stroke
Stroking
Stroma
Strong
Structural
Structure
Struther
Stylets
Styloglossus
Stylohyoid
Styloid
Stylomandibular
Stylomastoid
Stylopharyngeus

Subarachnoid
Subclavia
Subclavian
Subclavius
Subcortical
Subcostal
Subcutaneous
Subdivision
Subendocardial
Subepithelial
Subgingival
Subhepatic
Subiculum
Sublingual
Submandibular
Submental
Submucosa
Submucosal
Suboccipital
Subperiosteal
Subsartorial
Subscapular
Subscapularis
Subserve
Subspinale
Substance
Substantia
Substantial
Substitute
Substrate
Subthalamic
Succedaneous
Succeptable
Successional
Succinate
Succinylcholine
Suction
Sufentanil
Sufficiency
Suggest
Suggestive
Suitable
Sulci
Sulcus
Sulfate
Sulphate
Summation
Supercilii
Superfamily
Superficial
Superficiale
Supergene
Superior
Superioris
Superius
Supernumerary
Supinator
Supine
Supplement
Supplementary
Supply
Support
Supported
Suppress
Suppression
Suppressive

Suppressor
Supra
Supraclavicular
Suprahyoid
Supramarginal
Supramentale
Supraorbital
Supraorbitale
Suprascapular
Supraspinatus
Supratrochlear
Supreme
Surface
Surfactant
Surgeon
Surgery
Surgical
Surround
Survival
Suspend
Suspension
Suspensory
Sustentacular
Suture
Swallow
Sweating
Swell
Sylvian
Sylvius
Sympathetic
Sympathomimetic
Symphyseal
Symphysis
Symptomatic
Symptoms
Synapse
Synaptic
Synchondrosis
Syncope
Syncytiotrophoblast
Syndactyly
Syndesmosis
Syndrome
Synephrine
Synergeneic
Synergism
Synergistic
Synonymous
Synovial
Synthase
Synthes
Synthesis
Synthetic
Synthroid
Syphilis
Syringe
System
Systemic
Systolic
T
T1
T6
Tactile
Tails
Tangent
Tantigen

TAP
Tapering
Tarsal
Tartar
Taste
Taurodontism
TB
Tcell
T-cell
Tcells
TCR
TCR-1
TCR-2
Te
Teardrop
Tearing
Technique
Tectorial
Tegmen
Tegmental
Telencephalic
Telencephalon
Telophase
Temperature
Temporal
Temporale
Temporalis
Temporary
Temporomandibular
Temporoparietalis
Tendency
Tendinous
Tendon
Tenfold
Tenon's
Tenovaginitis
Tensilon
Tension
Tensor
Tentorial
Tentorium
Teratogen
Teres
Term
Terminal
Terminale
Terminate
Termination
Terminology
Tertiary
Testes
Testis
Testut's
Tetralogy
TGFs
Th
Th0
Th1
Th2
Thalamic
Thalamostriate
Thalamus
Thalamusly
Thebesian
Theophylline
Theory

Therapeutic
Therapy
Thiamylal
Thiazide
Thickening
Thickness
Thind
Thiopental
Thoracic
Thoracis
Thoracodorsal
Thorazine
Threat
Threatening
Threshold
Throat
Thrombophlebitis
Thrombus
Thymic
Thymocytes
Thymus
Thyro
Thyrocervical
Thyroglossal
Thyrohyoid
Thyroid
Thyroidea
Thyrolar
Tibia
Tibial
Tightening
Tightly
Tilt
TIMP
Tint
Tip
Tipping
Tissue
Titanium
TLR
TMD
TMJ
TNFs
Todaro
Toicity
Tolerance
Tomes
Tomogram
Tomography
Tongue
Tonic
Tonofibrils
Tonsil
Tonsillar
Tonus
Tooth
Toothless
Topical
Torque
Tortuous
Torus
Total
Tounge
Tourniquet
Toward
Towne's

Toxins
Tprioritizing
Trabeculae
Traced
Trachea
Tracheal
Tracings
Tract
Traction
Tracts
Tragicus
Tragus
Transcription
Transcutaneous
Transdermal
Transduce
Transduction
Transection
Transendothelial
Transfer
Transform
Transformation
Transforming
Transfusion
Transgenic animal
Transient
Transiently
Transit
Transition
Translation
Translocations
translucent
Transmembrane
Transmissible
Transmission
Transosseous
Transosteal
Transparent
Transplantation
Transporter
Transposition
Transseptal
Transversarial
Transverse
Transversely
Transversospinalis
Transversus
Trapezius
Trauma
Traumatic
Traumatize
Traveling
Traversing
Treated
Treatment
Treitz
Tremor
Trendelenburg's
Treves
Triangle
Triangular
Triaries
Triarnterene
Tributaries
Tricuspid
Trigeminal

Trigger
Trimenton
Trimester
tRNA
Trochlear
Truncus
Trunk
Ttyrosine
Tubal
Tube
Tuber
Tubercle
Tuberculin
Tuberculosis
Tuberculum
Tubing
Tubocurarine
Tubular
Tubules
Tuft
Tumor
Tumour
Turcica
Turkishsaddle
Turner's
Twirl
Twitches
TWL
Tympani
Tympanic
Typical
Typically
Typodont
Tyrosine
U
Ubiquitination
Ulcers
Ulnar
Ultimately
Ultimobranchial
Ultralight
Umbilical
Umbilicus
Uncinate
Unclean
Uncommon
Unconscious
Unconsciousness
Uncovertebral
Uncrossed
Uncus
Undamaged
Underbite
Underdeveloped
Undergo
Undergone
Underlying
Undiagnosed
Undifferentiated
Undue
Unerupted
Uniform
Unilateral
Unimodal
Union
Unitary

United
Universal
Unknown
Unmyelinated
Unpaired
Unresponsiveness
Unusual
Unwanted
Upgoing
Upper
Upturned
Ureter
Urethra
Urothelial
Urticaria
Usage
Uterine
Uterus
Utricle
UV
Uvulae
V
V genes
Vaccination
Vaccine
Vaccinia
Vacuole
Vagal
Vaginal
Vaginalis
Vagolytic
Vagus
Valium
Valproic
Valsalva
Valve
Valveless
Valves
Valvulae
Valvular
Variability
Variable
Variation
Variety
Varying
Vascular
Vascularity
Vascularized
Vasculature
Vasoactive
Vasoconstriction
Vasodilation
Vasodilator
Vasopressor
Vast
Vater
Vault
VCAM
Vein
Veli
Velocity
Vena
Venae
Veneer
Venereal
Venipuncture

Venom
Venospasm
Venous
Ventilation
Ventolin
Ventral
Ventricle
Ventricular
Ventrolateral
Venule
Verapamil
Verbal
Verbally
Vermal
Vermilion
Vermis
Verrill
Versed
Versus
Vertebra
Vertebrae
Vertebral
Vertex
Vertical
Verticillium
Vesalius
vesicle
Vesicles
Vessel
Vessicular
Vestibular
Vestibule
Vestibulocochlear
Vestige
Via
Vibration
Vicinity
Vidian
Vieussens
Viewed
Viewer
Villi
Villus
Viral
Virally
Virchow
Virus
Viscera
Visceral
Viscerocranial
Vision
Vistaril
Visual
Visualization
Visualized
Vital
Vitality
Vitelline
Vitello
Vitreous
Vitro
Vivo
VLA-1
VLA-6
Vocal
Vocalis

Voigt
Voit’s
Volar
Volitional
Volkmann’s
Voltage
Voluntary
Volvulus
Vomer
Vomeronasal
Vomerovaginal
Vomiting
Vomitus
Von Ebner's
VonBrunn’s
vonEbner’s
Vregion
Vulva
W
Waldeyer’s
Wallenberg’s
Walnut-sized
Warfarin
Wax
Weak
Weakening
Weakens
Weakness
Wedge
Weil's
Welded
Wernicke’s
Western blotting
Westphal
Wharton's
Wheezing
Whene
Whitens
Whitnall’s
Widely
Wider
Widespread
Wiebel
Willis
Winslow
Wire
Wisdom
Wisps
Within
Wolffian
Wound
Wrisberg’s
Xanthine
Xenogeneic
Xeroestomia
Xerostomia
X-ray
Y
Y-axis
Z
Zarontin
Zcerebellum
Zinc
Zinn
Zone
Zonula

Zonule	Zygomatic	Zymogen
Zoster	Zygomatico	A
Zuckerkandl's	Zygomaticofacial	B
Zvia	Zygomaticotemporal	Γ
Zygion	Zygomaticus	Δ
Zygoma	Zygote	Λ

APPENDIX 2

IRB APPROVAL

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Approved
UNLV IRB
FEB 18 2009
Expires

Social/Behavioral IRB – Exempt Review
Approved as Exempt

DATE: March 2, 2009

TO: **Dr. Alice Corkill,** Educational Psychology

FROM: Office for the Protection of Research Subjects

RE: Notification of IRB Action by Dr. J. Michael Stitt, Chair
Protocol Title: **The Development of a Model to Estimate the Readability of Credentialing- Examination Materials**
OPRS# 0902-3016

This memorandum is notification that the project referenced above has been reviewed by the UNLV Social/Behavioral Institutional Review Board (IRB) as indicated in Federal regulatory statutes 45CFR46.

The protocol has been reviewed and deemed exempt from IRB review. It is not in need of further review or approval by the IRB.

Any changes to the exempt protocol may cause this project to require a different level of IRB review. Should any changes need to be made, please submit a **Modification Form**.

If you have questions or require any assistance, please contact the Office for the Protection of Research Subjects at OPRSHumanSubjects@unlv.edu or call 895-2794.

Office for the Protection of Research Subjects
4505 Maryland Parkway • Box 451047 • Las Vegas, Nevada 89154-1047

REFERENCES

Abedi, J. & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education 14(3)*, 219-234.

Abedi, J. (2006). Language issues in item development. In S.M Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp.377-398). Mahwah, NJ: Lawrence Erlbaum Associates.

Abedi, J., Lord, C., & Plummer, J. (1995). Language *background as a variable in NAEP mathematics performance: NAEP TRP task 3D: Language background study*. Los Angeles: UCLA Center for the study of evaluation/National Center for Research on Evaluation, Standards, and Student Testing.

Allan, S., McGhee, M., & van Krieken, R. (2005). *Using readability formulae for examination questions.* SQA Research and Information Services. Retrieved from http://ofqual.gov.uk/files/allan_et_al_using_readability_formulae_for_examination_questions_pdf_05_1607.pdf.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME).(1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Aquino, M.R. (1969). The validity of the Miller-Coleman readability scale. *Reading Research Quarterly 4(3)*, 342-357.

Bormuth, J. R. (1966). Readability: A new approach. Reading Research Quarterly, 1(3), 79-132.

Bormuth, J. R. (1967). Comparable cloze and multiple-choice comprehension test scores. *Journal of Reading, 10(5)*, 291-299.

Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement, 5(3)*, 189-196.

Bormuth, J. R. (1969). Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly, 4(3)*, 358-365.

Bormuth, J. R. (1971) *Development of readability analyses.* Final Report, Project No. 7-0052, Contract No. 1, OEC-3-7-070052-0326. Washington, DC: U. S. Office of Education.

Carroll, J. B. (1976) Psychometric tests as cognitive tasks: a new structure of intellect. In L.B. Resnick (Ed.), *The nature of intelligence* (pp. 27-56). Hillsdale, NJ: Erlbaum.

Carroll, J. B., Davies, P. & Richmond, B. (1971). *The word frequency book*. Boston: Houghton Mifflin.

Caylor, J. S., Sticht, T. G., Fox, L. C., and Ford, J. P. (1973, February). *Development of a simple readability Index for job reading material.* Paper presented at the meeting the annual meeting of the American Educational Research Association, New Orleans, LA.

Chall, J. S. & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Chall, J. S. (1988). The beginning years. In B. L. Zakaluk and S. J. Samuels (Eds.), *Readability: Its past, present, and future.* Newark, DE: International Reading Association.

Cohen, P. & Cohen, J. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Coleman, E. B. (1965). On understanding prose: some determiners of its complexity. NSF Final Report GB-2604. Washington, D. C. National Science Foundation.

Coleman, E. B. & Miller, G. R. (1968). A measure of information gained during prose learning. *Reading Research Quarterly, 3(3),* 369-386).

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60,* 283-284.

College Entrance Examination Board (1981). *Degrees of Reading Power brings the student and the text together*. New York: DRP Services of The College Board.

Crabbs, L. M. & McCall, W. A. (1925). Standard test lessons in reading. *Teachers College Record, 27(3),* 183-183. http://www.tcrecords.org ID Number: 5914, Date Accessed: 1/30/2007.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.) *New Directions for Testing and Measurement: Measuring Achievement: Progress Over a Decade* No. 5, San Francisco: Jossey-Bass.

Cunningham, J. W. & Cunnigham, P. M. (1978) Validating a limited-cloze procedure. *Journal of Reading Behavior, 10(2),* 211-213.

Dale, E & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin, 17*, 1-20, 37-54.

Dale, E. & Chall, J. S. (1949). The concept of readability. *Elementary English 26*, 19-26.

Dale, E. & O'Rourke, J. (1981). *The living word vocabulary: A national vocabulary inventory*. Chicago: World Book–Childcraft International.

Dilworth, C. Reising, R., & Wolfe, D. (1978). Language structure and thought in written composition: Certain relationships. *Research in the Teaching of English, 12(2),* 97-106.

Downing, S. M. (2006). Twelve steps for effective test development. In S.M Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp.3-25). Mahwah, NJ: Lawrence Erlbaum Associates.

DuBay, W. H. (2004). *The principles of readability*. Impact Information, Costa Mesa, CA. Retreived from www.impact-information.com.

Dunn, L. M. & Markwardt, F. C. (1970) *Peabody Individual Achievement Test.* Circle Pines, MN: American Guidance Service.

Farr, J. N., Jenkins, J. J., & Paterson, D. G. (1951). Simplification of Flesch reading ease formula. *Journal of Applied Psychology, 35(5),* 333-337.

Felker, D. (1980). *Document design: A review of the relevant research*. (Report No. AIR-75002-4/80-TR). Washington, DC: American Institute for Research. (ERIC Document Reproduction Service No. ED192331).

Flesch, R. (1943). *Marks of a readable style.* Columbia University contributions to education, (No. 897). New York: Bureau of Publications, Teachers College, Columbia University.

Flesch, R. (1948). The new readability yardstick. *Journal of Applied Psychology, 32,* 221-233.

Freedman, A. (1980). Writing in college years: some indices of growth. *College Composition and Communication, 31(3),* 291-295.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher, 56(3),* 286-291.

Fry, E. B. (1968). A readability formula that saves time. *Journal of Reading, 7,* 513-516.

Fry, E. B. (1977). Fry's readability graph: Clarifications, validity, and extensions to level 17. *Journal of Reading, 21,* 242-252.

Golub, L. S., & Frederick, W. C. (1971). *Linguistic structures in the discourse of fourth and sixth graders.* (Tech. Rep. No. 166). Madison, Wisconsin: The University of Wisconsin, Wisconsin Research and Development Center for Cognitive Learning.

Gray, W. S. & Leary, B. (1935). *What makes a book readable*. Chicago: Chicago University Press.

Gunning, R. (1952). *The technique of clear writing*. New York: Mc Graw Hill.

Harris, A. J. & Jacobson, M. D. (1974). A comparison of the Fry, Spache, and Harris-Jacobson readability formulas for primary grades. *The Reading Teacher*, 28, 920-923.

Hewitt, M. & Homan, S. (1991). Readability: A review of past research with a look at a new beginning. *Journal of reading education, 16(3),* 6-18.

Hewitt, M. A. & Homan, S. P. (2004). Readability level of standardized test items and student performance: The forgotten validity variable. *Reading Research and Instruction, 43(2),* 1-16.

Homan, S., Hewitt, M., & Linder, J. (1994). The development and validation of a formula for measuring single-sentence test item readability. *Journal of Educational Measurement, 31(4),* 349-358.

Hoeng, J. M. & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55(1),* 19-25.

Ho-Peng, L. (1983). Using T-unit measures to assess writing proficiency of university ESL students. *RELC Journal, 14(2),* 35-43.

Hull, L. C. (1979). *Measuring the readability of technical writing*. Paper presented at the 26th International Technical Communications Conference, Los Angeles, CA.

Hunt , K. W. (1965). *Grammatical structures written at three grade levels*. (Res. Rep. No. 3). Champaign, IL: National Council of Teachers in English.

Hunt, K. W. (1970a). Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development, 35(1),* 1-67.

Hunt, K. W. (1970b). Do sentences in the second language grow like those in the first? *TESOL Quarterly, 4(3),* 195-202.

Kammann, R. (1966). Verbal complexity and preferences in poetry. *Journal of Verbal Learning and Verbal Behavior, 5,* 536-540.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L. & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Research Branch Report, 8-75, Millington, TN: Naval Technical Training, U. S. Naval Air Station, Memshis, TN.

Kintsch, W. & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Pychological Review, 85(5),* 363-394.

Kistulentz, A. C. (1967). *Five readability ratings compared to comprehension test scores on ten high school literature books.* Unpublished master's thesis, Rutgers, The State University of New Jersey, Brunswick, NJ.

Klare, G. R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

Klare, G. R. (1966). Comments on Bormuth's 'readability: a new approach'. *Reading Research Quarterly, 1(4),* 119-125.

Klare, G. R. (1974-1975). Assessing readability. *Reading Research Quarterly, 10(1),* 62-102.

Klare, G. R. (1984). Readability. In P. David Pearson (Series Ed.) & R. Barr, M. L Kamil, & P. Mosenthal (Sec. Eds.) *Handbook of Reading Research: Vol. 1* (pp. 681-744). New York: Longman.

Klare, G. R. (1988). The formative years. In B. L. Zakaluk and S. J. Samuels (Eds.), *Readability:Its Past, Present, and Future,* (pp. 14-34). Newark, DE: International Reading Association.

Lively, B., and Pressey, S. "A Method for Measuring the Vocabulary Burden of Textbooks," Educational Administration and Supervision, Volume 9, 1923, pp. 389--398.

Lorge, I. (1944 a). Predicting Readability. *Teachers College Record*. *45,* 409-419.

Lorge, I. (1944 b). Word lists as background for communication. *Teachers College Record. 45,* 543-552.

Lorge, I. (1939). Predicting reading difficulty of selections for children. *Elementary English Review, 16,* 229-233.

MacGinitie, W. H. & Tretiak, R. (1971). Sentence depth measures as predictors of reading difficulty. *Reading Research Quarterly, 6(3),* 364-377.

MacGregor, I. D., Balding, J. W., & Regis, D. (1997). Motivation for dental hygiene in adolescents. *International Journal of Pediatric Dentistry, 7,* 235-241.

Maimon, E. P. & Nodine, B. F. (1978). Measuring syntactic growth: errors and expectations in sentence-combining practice in college freshmen. *Research in the Teaching of English, 12,* 233-244.

McCall, W. A. & Crabbs, L. M. (1926, 1950, 1961, 1979). *Standardized test lessons in reading.* New York: Teachers College, Columbia University Press.

McLaughlin, G. H. (1969). Smog-grading—a new readability formula. *Journal of Reading, 13,* 639-646.

Miller, G. R. & Coleman, E. B., (1967). A set of prose passages calibrated for complexity. *Journal of Verbal Learning and Verbal Behavior, 6,*851-854.

National Dental Examining Board of Canada (2009). *Sample questions and answers to illustrate examination format.* Retrieved January 2009, from http://www.ndeb.ca/en/accredited/osce_examination.htm.

National Dental Examining Board of Canada (n.d.). *Released written examination book.* Retrieved January 2009, from http://www.ndeb.ca/en/accredited/documents/2006ReleasedEnglishBookII.pdf.

Oakland, T. & Lane, H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4(3),* 239-252.

O' Donnell, R. C. (1975). A critique of some indices of syntactic maturity. *Research in the Teaching of English, 10,* 31-38.

O'Donnell, R. C., Griffin, W. J., & Norris, R. D. (1967). *Syntax of kindergarten and elementary school children: A transformational analysis.* (Res. Rep. No. 8). Urbana, IL National Council of Teachers of English.

Pellegrino, J. W. & Glaer, R. (1982). Analyzing aptitudes for learning: inductive reasoning. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 269-345). Hillsdale, NJ: Erlbaum.

Plake, B. S. (1988). Application of readability indices to multiple-choice items on certification/licensure examinations. *Educational and Psychological Measurement, 48,* 543-551.

Popham, J. W. (1981, April). *Development of Readability Controlled Basic Skills Tests.* Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, CA.

Powers, R. D., Sumner, W. A., & Kearl, B. E. (1958). A recalculation of four readability formulas. *Journal of Educational Psychology, 49,* 99-105.

Redish, J. C. & Selzer, J. (1985). The place of readability formulas in technical communication. *Technical Communication, 4,* 1-8.

Sharrocks-Taylor, D., & Hargreaves, M. (1999). Making it clear: A review of language issues in testing with special reference to the National Curriculum Mathematics Test at Key Stage 2. *Educational Research, 41(2),* 123-136.

Smith, E. A. & Senter, R. J. (1967). *Automated readability index.* (AMRL-TR-66-22). Wright-Patterson AFB, OH: Aerospace Medical Division.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *Elementary School Journal, 53,* 410-413.

Stenner, J. A & D. S. Burdick.(1997). *The objective measurement of reading comprehension in response to technical questions raised by the California department of education technical study group.* (Report No. CS013755). Durham, NC: MetaMetrics, Inc. (ERIC Document Reproduction Service No. ED435978).

Stenner, J. A. & Smith, M. (1982). Testing construct theories. *Perceptual and Motor Skills, 55, 415-426.*

Stenner, J. A., Burdick, H., Sanford, E. E. & Burdick, D. S. (2006). How accurate are Lexile Text Measures? *Journal of Applied Measurement, 7(3),* 307-322.

Stenner, J. A., Smith, M. & Burdick, D. S. (1983). Toward a theory of construct definition. *Journal of Educational Measurement, 20(4),* 305-315.

Sternberg, R. J. (1977). *Intelligence, information processing and analogical reasoning: the componential analysis of human abilities.* Hillsdale, NJ: Erlbaum.

Stevens, K. C. (1980). Readability formulae and McCall-Crabbs Standard Test Lessons in Reading. *The Reading Teacher, 33,* 413-415.

Stokes, A. (1978). The reliability of readability formulae. *Journal of Research in Reading, 1,* 21-34.

Sydes, M., & Hartley, J. (1997). A thorn in the flesch: Observations on the unreliability of computer-based readability formulae. *British Journal of Educational Technology, 28,* 143-145.

Szalay, T. G. (1965). Validation of the Coleman readability formulas. *Psychological Reports, 17,* 965-966.

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

Thorndike, E. L. & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Bureau of Publications, Teachers College, Columbia University.

Thorndike, E. L. (1916). An improved scale for measuring ability in reading." *Teachers college record, 17,* 40-67.

Thorndike, E. L. (1921). *The teacher's word book.* New York: Bureau of Publications, Teachers College, Columbia University.

Thorndike, E. L. (1932). *A teacher's word book of 20,000 words.* New York: Bureau of Publications, Teachers College, Columbia University.

Tretiak, R. (1969*). Readability and two theories of English grammar*. Doctoral dissertation, Columbia University, *Dissertation Abstracts International, 30,* 1441.

Whiteley, S. E. (1981). Measuring aptitude processes with multicomponent latent trait models. *Journal of Educational Measurement, 18,* 67-84.

Williams, A. R., Siegel, A. L. , Burkett, J. R. and Groff, S. D. (1977). *Development and evaluation of an equation for predicting the comprehensibility of textual materials*. AFHRL-BT-77-8. Brooks AFB, Tx: Air force Human Resources Laboratory, Air force Systems Command.

Yngve, V. H. (1960). A model and hypothesis for language structure. *Proceedings of the American Philosophical Association, 404,* 444-466.

VITA

Graduate College
University of Nevada, Las Vegas

Barbara A. Badgett

Degrees

Bachelor of Science, Elementary Education, 2001
University of Nevada, Las Vegas

Master of Science, Educational Psychology, 2003
University of Nevada, Las Vegas,

Publication:

Hoffman, B. H., Badgett, B. A., Parker, R. P. (2008). The effect of single-sex instruction in a large, urban, at-risk high school. *The Journal of Educational Research, 102(1),* 15-35.

Dissertation Title: Toward the Development of a Model to Estimate the Readability of Credentialing-Examination Materials

Dissertation Examination Committee:

Chairperson, Alice J. Corkill, Ph.D.
Committee Member, CarolAnne M. Kardash, Ph.D.
Committee Member, Gregory Schraw, Ph.D.
Graduate Faculty Representative, Mark H. Ashcraft, Ph.D.