Readability of Credentialing Examination Materials:

Evaluating a Proposed Measurement Model

Barbara A. Badgett

Alpine Testing Solutions, Inc.

Alice J. Corkill

University of Nevada, Las Vegas

Abstract

The purpose of this study was to collect validity evidence to evaluate the use of a newly developed readability model designed to assess the readability of credentialing examination materials while accommodating the multiple-choice item format and occupational-specific language. The proposed model and recalibrated, existing readability formulas were used to assess the readability of sample examination and occupational materials for a licensing examination. Correlational analyses indicated that the results obtained with the proposed model were not strongly related to the results of the recalibrated, existing formulas. Differential treatment of occupational-specific vocabulary by the proposed model and existing formulas appeared to be responsible for the low correlations. The relationships between the proposed model and recalibrated formulas strengthened when occupational-specific vocabulary terms were not included in the calculation of the semantic variables for the recalibrated, existing formulas. The results provide preliminary evidence to support the viability of the proposed model for assessing the readability of credentialing examination items and related occupational materials.

Introduction

The “first step” in the test development process is to adequately describe the content domain or the scope of the construct that will be measured (Linn, 2006). The test framework for credentialing examinations must include clear descriptions of the content and/or constructs to be measured to support the appropriate representation of the domain (Badgett, 2010; Badgett & Corkill, 2010; AERA/APA/NCME, 1999). Credentialing examinations are designed to determine whether prospective practitioners possess the appropriate levels of knowledge, skills, and abilities and to provide the public some level of confidence in the competence of certified professionals.

Within the representation of a content domain, readability levels of credentialing examination items are one source of evidence that the intended content, cognitive demand, and performance expectations of an examination are consistent with job-related requirements (Badgett and Corkill, 2010). Equivalence across the readability levels of credentialing examination items and job-related textual materials should be established as an additional source of evidence to evaluate the appropriateness of score interpretation. If the readability level of credentialing examination items exceeds the reading level requirements for job performance, program sponsors must consider whether they have potentially allowed the introduction of undue construct irrelevant variance. Such an introduction of unnecessary error has the potential to negatively influence the appropriateness of score interpretations

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. Although the issue of readability is sufficiently critical for inclusion in The *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), credentialing test development does not usually involve formal measures of readability

The importance of readability issues in testing is not disputed. However, the format and content of multiple-choice credentialing test items introduces complications that hinder reliable and valid estimations of readability (Plake, 1988). According to Badgett and Corkill (2010), existing readability formulas cannot be confidently and reliably applied to multiple-choice credentialing examination items for four reasons.

First, with the exception of the Homan-Hewitt formula (Hewitt & Homan, 2004; Homan, Hewitt, & Linder, 1994) and the Lexile Framework (e.g., Stenner and Burdick, 1997), existing formulas are only suitable for several samples of continuous prose of 100 or more words. Multiple-choice test items are typically shorter than 100 words and are not typically comprised of continuous prose.

Second, although the Homan-Hewitt formula was developed specifically for use with multiple-choice test items, the authors developed and validated the formula with elementary-school-level materials and populations. Credentialing examinations are created for adult examinees who have typically endured post-secondary-school training or education. Because of inherent semantic- and syntactic-complexity differences of the materials as well as the expected differences in the reading comprehension abilities of the examinees, readability formulas designed for use with elementary-school level materials and populations would be inappropriate for post-secondary level materials and populations.

Third, although the Lexile Framework offers readability measures for the passages included in multiple-choice, reading-comprehension test items (e.g., Smith, Stenner, Horabin, & Smith, 1989), the framework was not designed to estimate the readability of test item stems or options. In fact, the Lexile Framework has been unsuccessful in estimating the readability of passages that are comprised of non-continuous prose (e.g., recipes or lists) and can only be generalized to continuous prose (Smith, et al., 1989).

Fourth, credentialing examination items often include occupational-specific, technical language that could skew the results of existing readability formulas. Such terminology is generally identified as "difficult" or as contributing to semantic complexity. This is inappropriate because a respective credentialing-examination audience could reasonably be expected to be familiar with such vocabulary.

Even though a readability model suitable for use with multiple-choice credentialing test items is not currently available, the fact remains that it is necessary to determine whether examination materials have readability levels that are congruent with the readability levels of occupational materials. If examination materials demand reading comprehension ability greater than that required for comprehension of the occupational materials, measurement error may be increased unduly. On the other hand, establishing that the readability levels, or reading comprehension demands, of the examination and occupation materials are congruent could provide construct-related validity evidence to credentialing programs (Badgett & Corkill, 2010).

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). For comprehension purposes, the readability level of a text should corresponds with, or not exceed, the reading ability of the reader. When the readability level of a text exceeds the reading ability of a reader, comprehension of the intended message is impeded.

A variety of mathematical equations have been developed to assess readability (McLaughlin, 1969). Readability formulas typically consist of predictor variables combined with constants and are generally used as a means of quantifying the reading ability required for an individual to comfortably read and understand a given text (Felker, 1980; Redish & Selzer, 1985; Stokes, 1978). Results are reported as numerical indices either as grade (e.g., Dale-Chall, 1948, 1995; FOG, 1952; FORCAST, 1973; Fry, 1965; Harris-Jacobson, 1974; SMOG, 1969; Spache, 1953) or scaled difficulty levels (e.g. Flesch, 1948; Lexile, 1987).

A host of syntactic and semantic variables have been investigated for their predictive power in estimating readability (DuBay, 2004; Fry, 2002; Klare, 1963; Oakland & Lane, 2004; Sharrocks-Taylor & Hargreaves, 1999; Sydes & Hartley, 1997). The most popular or widely used readability formulas address syntactic complexity via measures of sentence length (e.g., Bormuth, 1969; Chall & Dale, 1995; Dale & Chall, 1948; Flesch, 1948; Gunning, 1952; Lorge, 1939; McLaughlin, 1969; Spache, 1953; Stenner & Burdick, 1997) or number of sentences per passage (e.g., Chall & Dale, 1995; Coleman, 1965; Coleman & Liau, 1974; Fry, 1968, 1977). Measures of semantic complexity, or vocabulary load, most commonly used include: 1) average word length (e.g., Flesch, 1948; Fry, 1968, 1977); 2) number difficult words (identified according to familiar word lists; e.g., Chall & Dale, 1995, Dale & Chall, 1948, Homan, et al., 1994; Spache, 1953); and 3) syllable counts (e.g., Gunning, 1952; McLaughlin, 1969; Caylor, Sticht, Fox, & Ford, 1973; Coleman, 1965; Farr, Jenkins, & Patterson, 1951). Research shows that sentence length, word length, and the percentage of difficult words (vocabulary) have the most predictive power for estimating readability (Stenner & Burdick, 1997).

Readability in Testing

High-stakes test development does not typically involve formal measures of readability. This is likely because traditional readability formulas generally require several samples of continuous prose that include at least 100 words. Multiple-choice test items are not generally comprised of continuous prose and include stems that are usually between one and three sentences long with response options that are shorter.

Concatenating test items to create a single continuous prose segment that meets the length requirement of readability formulas is inappropriate for two reasons. First, prose subjected to readability formulas should be continuous and test items, regardless of whether they are artificially joined, are distinct pieces of text. Second, if items were combined to create quasi-continuous prose that meets length requirements and a traditional readability formula was applied, it would be impossible to assess the readability levels of individual items. Instead, the results would offer an overall estimate of the entire instrument (Badgett & Corkill, 2010; Homan, Hewitt, & Linder, 1994).

Homan et al. (1994) developed and validated the Homan-Hewitt readability formula for use with the individual sentences and phrases that occur in multiple-choice tests at the 2nd through 5th grade levels. Hewitt and Homan (2004) further investigated the use of their 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, 2) word length, and 3) sentence complexity. 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 the number of 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 (see Hunt, 1965).

In their investigations of the Homan-Hewitt readability formulas, Homan and colleagues determined that item readability affected (Homan et al., 1994) and was related to (Hewitt and Homan, 2004) item difficulty and, thereby, affected student performance. Students with lower reading comprehension abilities were most negatively affected by item readability. Essentially, unmatched levels of test-item readability and student reading ability likely introduced construct-irrelevant variance.

The Homan-Hewitt formula seems to be the only readability measurement model specifically designed for single-sentence, multiple-choice questions. Although Homan et al. (1994) and Hewitt and

Homan (2004) conducted validation studies for the Homan-Hewitt readability formula, it has not been adopted for use with standardized tests. It appears that no researchers, other than the developers, have published or presented studies using the formula.

The issue of item readability is particularly important in high-stakes, standardized testing. If test developers fail to address the issue of item readability, they run the risk of creating items that are not aligned with the reading abilities of examinees (Badgett & Corkill, 2010). If the readability level of a test item is beyond the reading ability of an examinee and what is required for the measurement of the intended construct, the measurement of the construct of interest is likely confounded. Specifically when the readability level of a test item exceeds the reading comprehension ability of examinees, the item is likely to measure dual constructs: the construct of interest and reading comprehension. Unless the construct of interest is 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).

Inappropriate readability levels of test items can affect overall test scores and those effects may not be consistent across candidates. Examinees who possess high levels of knowledge related to the construct of interest may incorrectly respond to a question because they are unable to decipher the intended message. This can result in different test performance outcomes for examinees with similar construct-related 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 score interpretation (Plake, 1988).

Readability in Credentialing Examinations

Credentialing examinations are often largely composed of multiple-choice items, which prevent them from being well suited for the use of traditional readability formulas (Badgett & Corkill, 2010).

Occupational-specific, technical language also interferes with the acquisition of readability estimations of credentialing examination items and their related materials (Allan, McGhee, & van Krieken, 2005). These words tend to artificially inflate readability estimates of credentialing examination items and related materials because readability formulas are designed to be sensitive to indicators of semantic complexity, such as word length or vocabulary load (Badgett & Corkill, 2010). In many instances occupational-specific terminology is lengthy, multisyllabic, and is not included in lists of familiar words that are commonly used to identify difficult vocabulary (e.g., Dale & O'Rourke, 1981; Chall and Dale, 1995). The identification of occupational-specific vocabulary words as contributing to the semantic complexity of credentialing materials, in turn, affects the readability estimates yielded from traditional readability formulas. The resulting readability estimates would be appropriate if the test were taken by examinees without domain-specific backgrounds, but candidates who take a credentialing examination could be expected to be familiar with most occupational-specific vocabulary. In short, valid measures of readability for credentialing materials should not allow occupational-specific vocabulary to affect the resulting readability estimates.

Plake (1988) recommends that readability checks be included in the validation process of credentialing examinations. Such measures can help control for potential threats to the validity of credentialing examination result interpretations introduced through construct-irrelevant variance. 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 readability levels low enough to ensure that a candidate 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).Unmatched levels of readability between materials could open the door for candidate appeal. If the developers of credentialing examinations do not adequately address issues of examination-item readability, the validity of score interpretations may, and perhaps should, be questioned.

Methods

The purpose of this study was to collect validity evidence for a set of procedures developed by Badgett (2010) to estimate readability of credentialing examinations and related materials. The procedures and equations were developed to evaluate readability while accommodating the multiple-choice item format and occupational-specific language of credentialing examinations and were designed to be appropriate for use with examination and occupational materials (see Badgett, 2010 for a complete description of model development).

The development of the model proposed by Badgett (2010) extended the work of Plake (1988) and the variables included in the equation were chosen with attention to the work of Homan and colleagues (e.g., 1994; 2004). To address the multiple-choice format of the examination items, procedures similar to those used by Plake (1998) were developed to convert the items from non-continuous to pseudo-continuous prose. To address the occupational-specific vocabulary included in the materials, an occupational-specific word list was created with the assistance of a subject matter expert from the credentialing examination content discipline. The list, which contained 4,902 terms assumed to be familiar to students of the discipline, was created by referencing 26 domain textbooks.

Badgett (2010) addressed syntactic and semantic characteristics in the proposed formula. Average T-Unit length was used as the indicator of syntactic complexity. The number of unfamiliar words at grade-level 8 according to *The Living Word Vocabulary: A National Vocabulary Inventory (*Dale & O'Rourke,

1981) was used to assess semantic complexity. The proposed formula was: Readability Estimate = 1192.242 + (19.278 * Number of Unfamiliar Words) + (8.461 * Average T-Unit Length).

Badgett (2010) also recalibrated two existing readability formulas with the same materials used to calibrate the proposed formula. Specifically, regression techniques were used to recalibrate the FOG and Homan-Hewitt formulas while maintaining their original predictor variables. The Homan-Hewitt formula was selected for comparison because the proposed model was developed with particular attention to the variables used. The FOG formula was selected for comparison because it was created specifically for use with adult-level reading materials and populations (Gunning, 1952). The recalibration of the existing formulas allows for a consistent comparison of results across the proposed and recalibrated formulas. Table 1 includes the original and recalibrated versions of the existing readability formulas.

Table 1

Original and recalibrated version of existing readability formulas

Gunning FOG formulas	
Original	Y′= 4 (SL) + (HW)
Recalibrated	Y′= 1257.188 + (11.469 * (HW+ SL))
Homan-Hewitt formulas	
Original	Y′= 1.76 + (.15 * WNUM) + (.69 * WUNF) – (.51 * WLON).
Recalibrated	Y′= 1128.958 + (.881 * WNUM) + (14.081 * WUNF) + (23.722 * WLON)

Note. SL = average sentence length; HW = percentage of words with more than two syllables, SL = average sentence length; WUNF = number of unfamiliar words; WNUM = t-unit length; and WLON = number of words with more than six letters.

The proposed model and recalibrated, existing formulas were applied to examination items from a dental licensing program and occupational materials related to dentistry. The results of the proposed and recalibrated formulas were compared within material sets using correlational analyses. Then, the occupational-specific vocabulary was used with the recalibrated formulas and they were applied to the examination and occupational materials, again. The results of the proposed and recalibrated formulas, with the use of the occupational-specific vocabulary list, were compared within material sets using correlational analyses.

Materials

The proposed formula 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) offers grade-level (4th grade through college) familiarity scores for multiple meanings of 44,000 words. The second list, an occupational-specific word list created by Badgett (2010), includes 4,902 terms assumed to be familiar to students of dentistry (see Badgett, 2010 for full list).

Occupational materials.

The occupational materials included 36 text samples of approximately 150 words each collected from resources dentists would be expected to reference during practice. The sources from which the sample passages were extracted were selected with the assistance of a subject matter expert from the credentialing examination content discipline. The sample occupational materials were drawn from 25 peer-reviewed professional journal articles, 4 professional magazine editorials, 6 product or equipment manuals, and 1 domain-practice, risk-management program manual. Following typical readability research methods, sample passages were randomly selected such that roughly one-third came from the beginning, middle, and final portions of the sources.

Examination materials.

Examination items (n = 300) and respective difficulty data for 100 candidates were provided by the licensure program. Multiple-choice items (n = 48) were selected from the two 150-item components (i.e., Book 1 and Book 2) of the examination: 24 examination items from each book. Stratified and systematic sampling procedures were used to ensure that broad range of difficulty was represented by the sample items. Each set of 150 items was sorted according to empirical item difficulty values and divided into three difficulty groups of 50 items each (see Table x for difficulty means, standard deviations, and ranges). The 50 items in each group were then resorted within their respective stratum according to their item identification codes. Staring at the first item in each difficulty stratum, every 6th item was identified for selection. This resulted in the selection of 16 items from each stratum (high, moderate, low difficulty; see Table 2).

Table 2

Item difficulty means, standard deviations, and ranges for difficulty strata

Difficulty Strata	Mean	*SD*	Range
High	.65	.09	.38 – .73
Moderate	.83	.04	.78 – .90
Low	.95	.03	.91 – 1.0

Procedures

Procedures reported by Plake (1988) and adapted by Badgett (2010) were used to convert the 48 examination items into pseudo-continuous prose. The six guidelines included in the procedures were:

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 treated in the same way as scenarios. If a set of instructions applied to a group of items, the instructions were included with each item and added to their pseudo-continuous prose.
6) Where the stem included options and the options referred back to the choices in the stem, the elements were combined to create as many complete sentences as possible.

The procedures devised for converting the items into pseudo-continuous prose yielded texts of at least four sentences each (see Table 3 for the number of examination items converted according to each guideline) for 47 items. One of the 48 selected items had fewer than four options and thus resulted in a pseudo-continuous text that included three sentences. After the items were converted into pseudo-continuous prose, the mean number of words was 88.54 (*SD* = 56.93, range = 41 – 378).

Table 3

Number of examination items converted according to each guideline

Guideline	Number of items
1	29
2	12
3	4
4	1
5	0
6	6

Note. 10 items that required the method 1 conversion also required method 4; 1 item that required method 6 also required method 4; 1 item that required method 2 also required method 4; 2 items that required the method 3 conversion also required method 4; and 1 item that required the method 3 conversion also required method 6.

The proposed formula and Homan-Hewitt formula require the identification of T-units and unfamiliar words. T-units are typically shorter than sentences, yet they possess, at minimum, a subject and a verb. Hunt (1965) explains that a T-unit is “a grammatically discrete unit intervening in size between the clause and what is punctuated as a sentence” and includes “one main clause plus the subordinate clauses attached to or embedded within it” (p. 49).

One rater identified the number of unfamiliar words in examination and occupational materials. Due to the complexity of the task, the identification of T-units involved two raters. The level of agreement between raters for T-unit identification was acceptable for the examination (r =1.0) and occupational materials (r = .931). Table 4 provides the means and standard deviations for numbers of unfamiliar words and T-unit length for the sample materials.

Table 4

Independent variable statistics for the proposed formula

	Material Type			
	Examination		Occupational	
	Mean	*SD*	Mean	*SD*
Number of Unfamiliar words	9.91	2.28	9.07	3.50
T-unit Length	15.53	6.39	20.64	7.45

Readability estimates for examination and occupational materials were calculated using the proposed formula, recalibrated FOG, and recalibrated Homan-Hewitt (See Table 5). Readability estimates for both sets of materials are reported in Table 6. High mean readability values indicate harder-to-read text and low mean readability values indicate easier to read texts.

Table 5

Proposed formula and recalibrated, existing formulas

Formula names	Equations
Proposed formula	$Y' = 1192.242 + (19.278*UFW) + (8.461*TUL)$
Recalibrated FOG	$Y' = 1257.188 + (11.469*(HW + SL))$
Recalibrated HH	$Y' = 1128.958 + (.881*WNUM) + (14.081*WUNF) + (23.722*WLON)$

Note: HH = Homan-Hewitt, UFW =number of unfamiliar words, TUL = average T-unit length, SL = average sentence length, HW = percentage of words with more than three syllables, WUNF = number of unfamiliar words, WNUM = T-unit length, and WLON = number of words with more than six letters.

Table 6

Descriptive statistics for all formulas

Material Type	Formula	Range	Mean	*SEM*	*SD*
Examination	PF	637.57	1514.76	20.21	140.03
	FOG	500.63	1742.67	18.31	126.84
	HH	2100.43	2946.62	61.57	426.55
Occupational	PF	466.21	1541.78	11.39	91.84
	FOG	858.57	1767.82	27.37	164.22
	HH	1301.94	2532.08	59.18	355.07

Note: PM = proposed formula, FOG = recalibrated FOG**,** and HH = recalibrated Homan-Hewitt.

Results

Correlational analysis was used to determine how well the results of the proposed and recalibrated formulas corresponded. The examination item correlation matrix shows that the results from the proposed formula were significantly correlated with the results of the recalibrated FOG (r = .244, p < .05), but were not significantly correlated with the results of the recalibrated Homan-Hewitt (see Table 7). The results of recalibrated Homan-Hewitt formula and FOG were significantly correlated (r = .596, p < .01).

Table 7

Examination items: correlations between formulas

	FOG	Homan-Hewitt
Proposed Formula	.244*	.159
FOG	--	.596**

Note: FOG = recalibrated FOG and HH = recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

The occupational material correlation matrix shows that the results from the proposed formula were significantly correlated with the results of both recalibrated formulas (FOG: $r = .531$, $p < .01$; Homan-Hewitt: $r = .342$, $p < .05$; see Table 8). The results of recalibrated Homan-Hewitt formula were significantly correlated with the result of the recalibrated FOG ($r = .540$, $p < .01$.

Table 8

Occupational materials: correlations between formulas

	FOG	Homan-Hewitt
Proposed Formula	.531**	.342*
FOG	--	.540**

Note: FOG = recalibrated FOG and HH = recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

For a second set of analyses, the occupational-specific vocabulary list designed for use with the proposed formula was used 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 materials. Specifically, during the calculation of the semantic variable for the recalibrated formulas, occupational-specific vocabulary terms were removed from the totals. In other words, if a word that was identified as multisyllabic (FOG) difficult, or long (Homan-Hewitt) was also in the occupational specific vocabulary list, it was not included in the respective variable total (e.g., percentage of multisyllabic words, FOG; number of unfamiliar words/number of long words, Homan-Hewitt). Therefore, the occupational-specific vocabulary terms were treated in a manner consistent with the way they were treated in the proposed formula. 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. Readability estimates for both sets of materials are reported in Table 9.

Table 9

Descriptive statistics for all formulas-- use of occupational-specific vocabulary list with all formulas

Material Type	Formula	Range	Mean	*SEM*	*SD*
Examination	PM	637.57	1514.76	20.21	140.03
	FOG	381.99	1529.38	12.15	84.20
	HH	1218.52	1716.04	43.42	300.84
Occupational	PM	466.21	1541.78	11.39	91.84
	FOG	803.75	1662.21	24.83	148.97
	HH	930.23	1964.84	35.06	210.36

Note:PM = proposed formula; FOG = recalibrated FOG; and HH = recalibrated Homan-Hewitt.

Correlational analysis was used to determine how well the results of the proposed and recalibrated models corresponded when the occupational-specific vocabulary list was used with the recalibrated formulas. The examination item correlation matrix shows that the results from the proposed formula were significantly correlated with the results of the recalibrated existing formulas (FOG: r = .535; Homan-Hewitt: r = .714, p < .01; see Table 40). The results of recalibrated Homan-Hewitt formula and FOG were significantly correlated (r = .312, p < .05; see Table 10). The occupational material correlation matrix shows that the results from the proposed formula were significantly correlated with the results of the recalibrated existing formulas (FOG: r = .618; Homan-Hewitt: r = .670, p < .01; see Table 11). The results of recalibrated Homan-Hewitt formula were significantly correlated with the result of the recalibrated FOG (r = .465, p < .01).

Table 10

Examination items: correlations between formulas-- use of occupational-specific vocabulary list with all formulas

	FOG	Homan-Hewitt
Proposed Formula	.535**	.714**
FOG	--	.312*

Note: FOG = recalibrated FOG and HH = recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

Table 11

Occupational materials: correlations between formulas-- use of occupational-specific vocabulary list with all formulas

	FOG	Homan-Hewitt
Proposed Formula	.618**	.670**
FOG	--	.465**

Note: FOG = recalibrated FOG and HH = recalibrated Homan-Hewitt. ** Correlation significant at .01 level (one-tailed); * Correlation significant at .05 level (one-tailed).

Discussion

The goal of this study was to collect validity evidence for a set of procedures developed by Badgett (2010) to estimate readability of credentialing examinations and related materials. According the results of the correlation analysis conducted here, the proposed model shows promise as a means of establishing readability while accommodating the multiple-choice item format and occupational-specific language related to credentialing examinations. Insights into the effects of occupational-specific vocabulary are provided in the results of the correlational analysis.

When applied to the examination materials, the proposed formula was not significantly correlated with the recalibrated Homan-Hewitt. Although the proposed formula was significantly correlated with the recalibrated FOG formula when applied to the examination materials, the relationship was weak. When applied to the occupational materials, the proposed formula was significantly correlated with the results of both recalibrated formulas. The relationship of the FOG was moderate and the relationship for the Homan-Hewitt was weak.

The second set of correlation analyses were conducted for the results that were obtained with the use of the occupational-specific vocabulary list in all formulas. For both material sets, the proposed formula was significantly correlated with the recalibrated FOG and Homan-Hewitt formulas, and both relationships were moderate. These results show that when the occupational-specific vocabulary list was used with the recalibrated formulas, the correlations between the proposed and recalibrated formulas increased, as compared to when the use of the recalibrated formulas only involved the use of their prescribed materials or method for assessing semantic complexity.

The increases in relationships between the proposed formula and recalibrated formulas with the use of the occupational-specific vocabulary list highlights the effects and importance of the occupational-specific vocabulary when evaluating the readability of materials of this nature. Without the use of the occupational-specific vocabulary list, many of the job-related terms included in the dentistry examination and occupational materials were identified as contributors to semantic complexity. The proposed model treats the occupational-specific vocabulary as familiar, which is appropriate considering the intended audience. Finding stronger relationships between the proposed and recalibrated formulas with the use of the occupational-specific vocabulary list across all materials lends support to the viability of the proposed formula and procedures. However, further research should be conducted to collect additional validity evidence to evaluate the proposed formula and procedures.

Future Directions

The proposed formula and procedures have not been thoroughly or broadly evaluated. The semantic- and syntactic-complexity measures included in the new-models appear to be appropriate indicators of readability for credentialing-examination materials, but further studies should be conducted.

Studies should be designed to link the results obtained with the proposed model to grade-level equivalents or other scaling alternatives that can provide substantive interpretability or broader utility. A variety of techniques could be considered (e.g., regression approaches, linear equating, equipercentile equating). In addition, with the results of whichever technique proves most useful, a concordance table could be created that would show how results obtained with the proposed formula correspond with results of other readability formulas.

Further external validity and reliability studies should also be conducted. Research might involve applying the proposed formula and recalibrated formulas to a different, yet similar set of sample materials. This would entail collecting sample examination and occupational 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., physical therapist, physician assistant). This approach would offer the opportunity to inspect relationships between the readability estimates derived with the proposed formula and recalibrated 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.

Another method of investigating the proposed procedures and formula might involve applying the proposed formula and existing formulas to materials that do not include occupation-specific vocabulary. With this approach, the proposed formula and procedures 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 proposed formula 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 proposed formula and recalibrated formulas were compared. Then, the occupational-specific vocabulary list was used with the recalibrated formulas and the readability estimates derived with the proposed formula 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 proposed formula 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 proposed formula and recalibrated formulas. The research suggested here would remove the consideration of occupational-specific vocabulary from the proposed formula and thereby offer a more consistent comparison of the proposed formula and existing, rather than recalibrated, formulas. If the results of the new-model and existing formulas corresponded well, it would lend support to the proposed model as a reasonable measure of readability. It would follow, then, that if the proposed model included accommodations for occupational-specific vocabulary (i.e., reintroduce the use of occupational-specific vocabulary lists) and was 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.

The collection of additional evidence to support of the utility of the proposed readability model as a valid and reliable measurement method for materials related to credentialing examinations could support the viability of its use by credentialing programs. Such a model would allow programs to determine whether congruency exists between the readability levels of their respective examinations and related occupational materials. Finding evidence of essential equivalence in readability levels across the materials would provide program sponsors critical evidence to support the validity of corresponding score interpretations. Alternatively, should a credentialing program discover that the readability levels of examination items are significantly higher than the readability levels of job-related materials, this knowledge could be used to inform future test development practices.

References

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.

Badgett, B. A., (2010). *Toward the development of a model to estimate the readability of credentialing-examination materials*. Unpublished doctoral dissertation, University of Nevada, Las Vegas.

Badgett, B. A. & Corkill, A. J. (2010, May). *A proposed model to evaluate the readability of credentialing examinations.* Paper presented at the annual meeting of the Natinal Council on Measurement in Education, Denver, CO.

Bormuth, J. R. (1969). Factor validity of cloze tests as measures of reading comprehension ability. *Reading Research Quarterly, 4(3)*, 358-365.

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 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.

Coleman, E. B. (1965). On understanding prose: some determiners of its complexity. NSF Final Report GB-2604. Washington, D. C. National Science Foundation.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60,* 283-284.

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.

Dale, E & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin, 17*, 1-20, 37-54.

Dale, E. & O'Rourke, J. (1981). *The living word vocabulary: A national vocabulary inventory*. Chicago: World Book–Childcraft International.

DuBay, W. H. (2004). *The principles of readability*. Impact Information, Costa Mesa, CA. Retreived from www.impact-information.com.

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. (1948). The new readability yardstick. *Journal of Applied Psychology, 32*, 221-233.

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.

Fry, E. (2002). Readability versus leveling. *The Reading Teacher, 56(3),* 286-291.

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.

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.

Hunt , K. W. (1965). *Grammatical structures written at three grade levels*. (Res. Rep. No. 3). Champaign, IL: National Council of Teachers in English.

Klare, G. R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

Linn, R. L. (2006.) The standards for educational and psychological testing: Guidance in test development. In S.M Downing & T.M. Haladyna (Eds.), *Handbook of test development* (pp.27 - 38). Mahwah, NJ: Lawrence Erlbaum Associates.

Lorge, I. (1939). Predicting reading difficulty of selections for children. *Elementary English Review, 16,* 229-233.

McLaughlin, G. H. (1969). Smog-grading—a new readability formula. *Journal of Reading, 13*, 639-646.

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.

Plake, B. S. (1988). Application of readability indices to multiple-choice items on certification/licensure examinations. *Educational and Psychological Measurement, 48,* 543-551.

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, D. R., Stenner, A. J., Horabin, I., & Smith, M. (1989). *The Lexile Scale in theory and practice*: *Final report* (Report No. HD-19448).Washington, D. C.: MetaMetrics. (ERIC Document Reproduction Service No. ED307577).

Spache, G. (1953). A new readability formula for primary-grade reading materials. *Elementary School Journal, 53,* 410-413.

Stenner, A. J. & Burdick, D. S. (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).

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.

Author’s Note

Barbara A. Badgett, Alpine Testing Solutions, Inc.; Alice J. Corkill, Department of Educational Psychology, University of Nevada, Las Vegas.

A portion of this research is one component of a large-scale dissertation study conducted by Barbara A. Badgett and submitted in partial fulfillment of the requirements for the Ph.D. in Educational Psychology, Department of Educational Psychology, College of Education, University of Nevada, Las Vegas. The authors wish to thank Jack D. Gerrow, DDS, MS, Cert Pros, MEd; Marcia A. Boyd, DDS, MA, LHD(Hon), DSc(Hon); and the National Dental Examination Board of Canada for their continued support.

Correspondence concerning this paper should be addressed to Barbara A. Badgett, Alpine Testing Solutions, Inc. Email: barbara.badgett@alpinetesting.com.