Flesch-Kincaid Reading Grade Level Re-examined: Creating a Uniform Method for Calculating Readability on a Certification Exam

Abstract

Objective: This study attempted to establish a consistent measurement technique for calculating readability on a state-wide Certified Nursing Assistant's (CNA) certification exam. **Background:** Monitoring the readability level of an exam helps ensure all test versions do not exceed the maximum reading level of the exam, and that knowledge of the subject matter, rather than reading ability, is being assessed. **Method:** A two part approach was used to specify and evaluate readability. First, two methods (Microsoft Word® (MSW) software and published readability formulae) were used to calculate Flesch Reading Ease (FRE) and Flesch-Kincaid Reading Grade Level (FKRGL) for multiple standardized tests as well as a state-wide CNA certification exam. Statistics calculated by hand were compared to those computed by MSW. Second, due to inconsistencies in readability statistic calculations, a single method was developed to calculate readability in order create tests at or below an eighth grade reading level. **Results:** There were significant differences between readability statistics calculated by hand and those calculated using MSW for the standardized tests as well as the CNA certification exam. Hand calculations indicated an easier to understand document than did MSW. Subsequently, by removing identifying values (e.g. numbers and letters), calculated reading levels were then consistent across test versions. **Conclusion:** Reading grade levels calculated via unpublished formulae should be used with caution due to inconsistent results. Further, creating a standardized format for the CNA exams will aid in making sure readability statistics of the document fall within the certification exam's guidelines. **Application:** The reading grade level calculation should be used to ensure the maximum reading level on a certification exam is not exceeded. Evaluating influences that affect reading level calculations should be an integral aspect of creating standardized tests.

Introduction

Standardized tests and certification exams require both procedural knowledge of the content as well as the ability to read and comprehend the test questions and prompts. Ensuring readability level matches reading ability of test takers is of paramount importance in the creation of a test. Specifically, the Certified Nursing Assistant (CNA) Competency Evaluation is a certification exam in the medical field, and is required by the state to be written at an eighth grade reading level or lower. Test creators and Nurse Aide testing researchers are tasked with ensuring the readability requirements are satisfied. However, previous research on readability underlines the vast inconsistencies and problems that can arise throughout the readability calculation process. This study attempts to clarify a few key problems with readability calculation as well as develop a uniform method for ensuring readability requirements are met on the CNA exam.

Literature Review

Effective examinations are well organized, easy to understand (e.g. written at an appropriate reading grade level), and arranged in a way that appeals to the reader (Osborne, 2000). However, readability often becomes an issue in the creation of exams. Readability (comprehension difficulty) is one characteristic of an examination that determines its utility and usefulness as an assessment tool (Bormuth, 1966). This comprehension difficulty refers to how well the text's

readability matches the reading comprehension level of individual readers (Benjamin, 2012). Readability can be determined based on characteristics like syllables per word, number of words per sentence, word difficulty, and language complexity (McClure, 1987).

First examined through the lens of survey development, readability of text is important to assure that the target audience is reached (Cantril, 1944; Payne, 1951; Terris, 1949). If the comprehension level of the examination is too high, an individual may have the procedural knowledge to pass the exam, but still may not pass because the reading grade level of the exam is higher than that individual's personal reading grade level. For example, in a competency exam for specific job placement, the exam must be written on a level at or below the minimum education level required for that job. Questions that are too difficult produce higher levels of variance in test scores (Fowler, 1995). Furthermore, questions that are difficult for readers to comprehend can produce unnecessary measurement errors, which affects the psychometric properties of the exam (Groves, 1989). For this reason, readability must be calculated on examinations to guarantee reliable test results.

There are a variety of techniques available to calculate the reading grade level of a printed document. These formulae serve as tools to guarantee the reading grade level of the document matches that of the individual readers. Readability formulae were originally designed for elementary school textbooks as a way to verify that the books were not too difficult for children to comprehend (DuBay, 2004). One of the first readability tools created for this purpose was the Flesch Reading Ease (FRE) formula (Flesch, 1948). This formula is based on a 100-point scale, where a higher score indicates an easier to understand document. This measure was added to in 1975, resulting in the Flesch-Kincaid Reading Grade Level (FKRGL) formula (Thomas, Hartley, & Kincaid, 1975). The FKRGL formula for determining readability is perhaps the most widely used across various disciplines and rates text based on U.S. school grade level ranging from one to12. While there have been documented flaws in the design of these techniques, the FRE and the FKRGL test are two of the most commonly used and available tools to evaluate readability (Ley & Florio, 1996).

The formula used to calculate FKRGL is based on two main criteria. The first, syntactic difficulty, is measured by evaluating the number of words per sentence. The second component, word difficulty, is quantified by syllables per word (Thomas et al., 1975). While these criteria are relatively easy for researchers to calculate "by hand," (e.g. inputting the values computed in a word count from a word processing program into the published formulae) creating a computer program that accurately incorporates syntactic difficulty and word difficulty has proved challenging (Hochhauser, 2005a). To calculate the number of sentences, programs rely on a count utilizing punctuation, although this is not always accurate.

The presence or absence of punctuation also influences the subsequent readability statistics. Further, abbreviations and lists can falsely increase the sentence count. For example, Coke and Rothkopf (1970) used a computer program to determine an algorithm for reading ease, and found that a "word" can be any number of alphanumeric symbols, and a sentence can be any words between two punctuation marks. In the medical field, titles such as "R.N." and "M.D." can falsely inflate word and sentence count in this manner. Similarly, in an attempt to write computer code encompassing the FRE formula, Fang (1968) noted that the general rule of one syllable per

vowel in a word has an extraordinary number of exceptions. Thus, these general rules create high variance and discrepancies in computer software programs meant to determine readability.

Certain computer programs like Microsoft Word® (MSW) produce reading grade levels for documents, but the algorithms used by the program to produce these numbers are not made public (DuBay, 2004). Researchers have made attempts at reproducing these results, but without the specific algorithms, it is impossible to determine if the readability scores generated by computer match those calculated by hand (Hcchauser, 2005a). Further, Mailloux, Johnson, Fisher, & Pettibone (1995), utilized four readability programs to analyze the same text, and found significantly different readability scores, despite the fact that the same FKRGL formula was reported as being used by each program. A more recent study also noted this major discrepancy, and even found that the same document scanned on two separate computers both using MSW yielded significantly different readability results (Benjamin, 2012). This is especially important in standardized tests and certification exams where multiple researchers are calculating readability and need to be sure that calculations are consistent across machines.

Statement of Problem

Certified Nursing Assistants (CNAs) are medical professionals who help patients with healthcare needs under the supervision of a Registered Nurse (RN) or a Licensed Practical Nurse (LPN). CNAs help fulfill basic quality of life needs of patients such as taking vital signs and in some cases, administering medications and treatments.

To become a CNA, individuals must complete an approved training program and pass a state-wide competency examination that tests knowledge and nursing skills. Typically, a qualified organization is responsible for creating and administering the written competency exam as required by their respective state's Department of Public Health (DPH). Further, the DPH requires all CNA exams to be written at or below a reading grade level of eight years to ensure exam scores reflect the knowledge of the individual rather than his or her ability to read at that level.

To meet this reading level requirement, readability statistics (e.g. Flesch-Kincaid Reading Grade Level, Flesch Reading Ease) were calculated for every CNA exam version administered by the qualified organization. The Flesch-Kincaid Reading Grade Level (FKRGL) represents the amount of education in years required to adequately understand the document. However, researchers for Nurse Aide Testing observed that FKRGL values calculated in Microsoft Word® (MSW) varied from a grade level of 3 to almost 12, even for an exam with identical questions. These discrepancies varied as a function of individual test imputers (i.e. individuals inputting the same test questions, but using slightly different formats regarding periods and numbering). With a state requirement of a reading level of eight years of education, any CNA exam with higher reading levels could not be given to students. Further investigation by Nurse Aide Testing researchers on FKRGL highlighted the problematic nature of this measure, especially the calculation provided by MSW. While formulae have been published regarding how to calculate readability statistics by hand, MSW does not give any information on how FKRGL and FRE are computed within the program. The framework originally created by Flesch (1948), while claimed by MSW to be the main source of readability calculation, may actually be misrepresented in the MSW algorithms.

The purpose of the present research had two main aims. First, the discrepancies between MSW calculations of readability statistics and hand calculations based on published formulae are explored. Subsequently, patterns in differences between the two calculation methods for various standardized tests are examined. Second, to address the problem of inconsistency in the CNA exam readability statistics, a uniform method that establishes stable and consistent reliability across tests while adhering to DPH requirements was developed. These findings help ensure that all tests administered meet the required reading grade level, and that CNA students are being tested on competency across duty areas rather than reading comprehension. Further, test creators can be more certain that those who fail the CNA exam lack the skillset and understanding required to pass. A stronger focus on content rather than readability will allow test creators to make the best possible questions for the CNA exam.

Methods (Part 1) Materials

External Exams

Online sources were used to obtain the 2010 New York State Regents Exam (NYSRE) for third, fourth, fifth, sixth, seventh, and eighth grade English. Twenty-nine passages and subsequent comprehension questions were utilized from the NYSRE. The number of passages for each grade level were as follows: four passages from the third grade exam, five passages from the fourth grade exam, four passages from the fifth grade exam, five passages from the sixth grade exam, six passages from the seventh grade exam, and five passages from the eighth grade exam.

However, as educational standards may differ across states, we also wanted to capture educational standards from the state where the CNA exam was administered. Thus, practice test questions from the second state's 2010 Standardized Achievement Test for third, fifth, six, and seventh grade English were located online from the state assessment office through the state board of education for analysis. Analysis included two passages and subsequent comprehension questions from each standardized achievement test grade level. Thus, the analysis included a total of 37 passage/question combinations.

MSW was used to calculate readability statistics for each document. Additionally, Microsoft Excel was used to complete the hand calculations. For the present study, "hand" calculations refers to those calculations done in Excel using the published readability formulas (DuBay, 2004; Flesch, 1948) instead of the numbers that MSW generates automatically with the word count. For each passage, MSW calculated number of words, characters, paragraphs, and sentences, as well as average number of sentences per paragraph, words per sentence, characters per word, and percentage of passive sentences. These numbers were inputted into Excel using the formula function. The Statistical Package for the Social Sciences (SPSS) version 18 was used to analyze data.

Procedure

External Exams

To compare readability statistics calculated by hand with the published formulas to those calculated in MSW, standardized tests were first located via online sources. The 2010 versions of the NYSRE and the state's standardized achievement test's English exams for grades three through eight were copied and pasted from the online source into MSW for analysis.

Each passage and subsequent questions of the exam were entered into a separate document. This created 29 passages for the NYSRE grades three through eight. The same process was repeated for the state's standardized achievement test's practice questions, where each passage / questions was entered into a separate word document. Two pairs of readability statistics were calculated for each document: the FRE generated in the readability feature in MSW and the FRE calculated by hand using the published formula and Excel, as well as the FKRGL calculated by MSW and the FKRGL calculated by hand in Excel. A total of 37 passages and questions were analyzed for readability.

Analysis

External Exams

Paired samples t-tests were implemented to compare hand calculations of FRE to Microsoft calculations of FRE as well as compare hand calculations of the FKRGL to Microsoft calculations of the FKRGL. Bonferroini corrections were utilized to control error rate associated with running multiple t-tests.

Results (Part 1)

External Exams

To compare hand calculations to MSW calculations, a paired samples t-test was utilized. Overall results indicated a statistically significant difference between FRE calculations done by hand (M=91.22, SD=10.18), and FRE calculations done in MSW (M = 80.40, SD = 10.04), $t(36) = -16.91$, $p < .001$. FRE hand calculations were higher than MSW values (See Tables 2 & 3).

Similarly, the hand calculations and MSW calculations of FKRGL were significantly different, $t(36) = 13.58$, $p < .001$. MSW calculations (M=4.31, SD=1.82) were consistently higher than hand calculations (M=3.04, SD=1.82). These discrepancies indicate the inconsistency in the way MSW calculates readability statistics (See Tables 1 & 2).

Table 1

Mean Readability Statistics of Sample State Standard Achievement English Test Essays (2010)

Exam	FRE (calculated by MSW)	FRE (Calculated by hand)	FKRGL (calculated by MSW)	FKRGL (Calculated by hand)	Passive Sentences
Grade 3	89.25	97.78	2.60	1.57	3%
Grade 5	81.30	81.64	4.05	4.14	7%
Grade 6	79.55	89.11	4.75	3.55	8%
Grade 7	67.95	84.61	6.00	4.25	2.5%

Table 2

Mean Readability Statistics of New York State Regents English Exam (2010)

Exam	FRE (calculated by MSW)	FRE (Calculated by hand)	FKRGL (calculated by MSW)	FKRGL (Calculated by hand)	Passive Sentences
Grade 3	88.55	99.63	2.63	1.17	0.5%
Grade 4	87.04	98.89	2.80	1.33	1.4%
Grade 5	84.75	94.99	3.80	2.48	4.25%
Grade 6	77.74	89.59	4.98	3.49	6.6%
Grade 7	76.65	88.16	5.21	3.94	4.3%
Grade 8	72.36	83.78	5.76	4.61	6.8%

Methods (Part 2)

Materials

Results from the first section of the present study highlighted the major discrepancies in MSW's calculation of readability statistics. It is required that the certification exam is written at a reading grade level of eight or lower. FKRGL is used as a measure to ensure this requirement is met. However, researchers found that the exact same version of a certification exam yielded different FKRGL numbers, depending on how the exam was typed by the test creators. Although all questions and possible answer choices were identical, formatting and numbering differed only slightly (See Table 3). This was especially problematic in versions that were yielding a FKRGL of eight or higher. Thus, two modifications were used to determine the cause of inconsistent and unusually high readability statistics.

Table 3

Data Calculated from Original Test Version Using MSW

	Researcher 1	Researcher 2
Words	3799	4231
Characters	19079	20273
Paragraphs	426	425
Sentences	57	343
Sentences per Paragraph	1.8	1.0
Words per Sentence	15.3	8.7
Characters per Word	4.9	4.5
Passive Sentences	14%	6%
Reading Ease	53.2	68.6
Grade Level	9.6	5.8

Six test versions of the certification exams were used in the sample. Each version included three modifications of the same exam.

Procedure

Certification Exams

First, two researchers copied and pasted the same exam version into MSW and calculated readability statistics in MSW and by hand. The discrepancies between the two were high, and only one of the exams met the required FKRGL. Despite the test questions being identical, question format (e.g. period after question numbers or not, how multiple choice options were formatted, etc.) differed slightly.

Three modifications of each CNA exam version were then created. One was the original test version that had been administered multiple times to CNA students (i.e. no modification). The exam was typed exactly as written, including the question number and the lettered alternatives for the multiple choice answers. The second modification of the exam completely omitted numbers and letters denoting choices completely, and re-inserted these markers as images rather than typed text. The last modification used complete sentences, by duplicating the question stem to create complete sentences for the multiple choice options. This eliminated one word multiple choice responses (See Figure 1).

Readability statistics for each modification of each exam version were calculated both in MSW and by hand.

Figure 1

Example Test Questions Demonstrating Three Modifications

Original Format 1. There are three primary colors. One of the primary colors is: A. red. B. pink. C. green. D. purple.
No Numbers There are three primary colors. One of the primary colors is: red. pink. green. purple.
Complete Sentences There are three primary colors. One of the primary colors is red. One of the primary colors is pink. One of the primary colors is green. One of the primary colors is purple.

Analysis

Certification Exams

A multivariate analysis of variance (MANOVA) was used to examine the effect of test modification on both hand and Microsoft calculations of FRE and FKRGL on the certification exams. Additionally, paired samples t-tests were utilized to compare hand calculations to MSW calculations for each test modification. A Bonferroni correction was utilized to decrease the error rate associated with running multiple t-tests. Significant results reflect a p-value less than the corrected alpha level created by dividing standard alpha (.05) by the number of comparisons for each test.

Results (Part 2)

Certification Exam

Readability statistics for three modifications of the CNA exam were compared using a multivariate analysis of variance (MANOVA). Results indicated a significant main effect for type of modification (original, complete sentences, and no numbers) on MSW FRE calculation, F (2, 15)= 72.54, $p < .001$. The main effect for test modification on MSW FKRGL was also statistically significant, $F(2, 15) = 566.82$, $p < .001$.

Paired samples t-test revealed a statistically significant difference between hand calculations and MSW calculations of FRE values, $t(17) = -13.43$, $p < .001$. Overall, MSW calculations (M=58.62, SD=8.53) were lower than hand calculations (M=72.70, SD=8.02). Thus, MSW calculations of FKRGL (M=8.24, SD=2.55) were higher than hand calculations of FKRGL (M=7.07, SD=2.13).

Separate analyses for each version of the exam (e.g. original, complete sentences, and no numbers) were conducted adopting a Bonferroni correction to reduce error rate. Thus, the standard alpha level of .05 was divided by 3, yielding a new alpha level of .02 to compare to the computed p-values.

For tests in the original format, there was a significant difference between MSW FRE calculations and hand FRE calculations, $t(5) = -4.06$, $p < .02$. However, original test versions did not show a significant difference between MSW and hand calculations of FKRGL, $t(5) = 2.09$, $p = .20$. The complete sentences version of the exam showed significant differences between MSW calculations and hand calculations for both FRE, $t(5) = -94.88$, $p < .001$, and FKRGL, $t(5) = 52.26$, $p < .001$. No numbers (the final modification of the exam) showed similar results, with significant differences between MSW and hand calculations for FRE, $t(5) = -52.21$, $p < .001$ and FKRGL, $t(5) = 16.50$, $p < .001$. In all of the discrepancies between hand calculations and MSW calculations, MSW calculated a FKRGL that was higher than that found in the hand calculation.

Perhaps most notably, the No Numbers modification of the certification exam yielded identical results for readability, regardless of test inputter. Thus, numbers and letters were added back into the test version as images rather than text, as to not influence readability, but still allow test takers to see the question numbers. This method also helped ensure the requirement of an eight-year reading grade level was satisfied.

Discussion

The goal for the present research was to examine differences in readability statistics calculated both by hand and using MSW readability software in order to create a CNA exam that satisfies the requirement set by the DPH. Discrepancies were found in readability statistics between those calculated by hand and using MSW, with MSW reporting higher grade level text across the board. New formats of the CNA exam were constructed by the test creators using this data, scoring at or below the required reading grade level.

For CNA exams, hand calculations also resulted in higher FRE scores than did MSW. Hand calculations of FKRGL designated a lower grade level than that of MSW. Overall, MSW indicated a more difficult-to-read document than was calculated using original formulae. An implication for a higher reported reading grade level according to MSW could be that issues arise in attempts to lower the grade level of the document. An eighth grade reading level is the upper limit for the CNA exams. If MSW indicates that a document exceeds this boundary, changes must be made to the document in order to lower the reading level. If the MSW formula is artificially lowering the reading level, unnecessary simplification of the document could lead to confusion on the part of the test taker.

In external exams, NYSRE and a state's standardized achievement test, FRE hand calculations were higher than MSW statistics. Keeping with this trend, hand calculations for FKRGL were lower than MSW scores. Therefore, according to the official published formula, the NYSRE and a state's standardized achievement test were easier to read than reported by MSW. These discrepancies raise the question, how does MSW calculate readability?

Issues arise when readability statistics are measured by MSW. Programs like MSW do not publish the methods by which they determine syllable or sentence counts, and different programs use varying algorithms to measure exam items, which results in a range of readability statistics for the same document depending on which program is used (Hochhauser, 2005b). Some researchers advise against the use of MSW to calculate readability statistics, as it artificially caps the resulting grade level at 12, even though the original formula gives scores up to 17 (Hochhauser, 2005b). These findings are consistent with studies noting the major discrepancies in readability calculations in MSW (Dubay, 2004). However, the largest contribution of the present research lies in the removal of numerical identifiers and letters throughout the CNA exam. To date, no readability research has noted the differences found by removing the values denoting questions and answer choices. Implementation of this new approach may aid in the confusion often encountered by test creators in readability requirements and calculations.

Directions for Future Research

Through the course of this study, it has become apparent that future research into readability formulae and algorithms must be done. Statistically significant discrepancies on the same text between hand calculation and MSW readability scores can create confusion for those designing exams or simply writing a document. If readability data is used to drive decisions about documents, the statistics must be consistent and accurate.

Plans are currently in effect to create an online version of this state-wide certification exam. It is intended that in the near future the majority of candidates taking the state-wide certification exam

will do so using a computer-based format. In such a format, there will be one test item per screen. The sensitivity of readability formulae may become an issue when smaller amounts of text are evaluated. Therefore, future studies will have to be conducted in order to maintain the required reading grade level.

References

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63–88. http://doi.org/10.1007/s10648-011-9181-8

Bormuth, J. R. (1966). Readability: A new approach. *Reading Research Quarterly*, *1*(3), 79–132. http://doi.org/10.2307/747021

Cantril, H. (1944). *Gauging public opinion*. Princeton, NJ: Princeton University Press.

Coke, E. U., & Rothkopf, E. Z. (1970). Note on a simple algorithm for a computer-produced reading ease score. *Journal of Applied Psychology*, *54*(3), 208–210. http://doi.org/10.1037/h0029067

DuBay, W. H. (2004). The principles of readability. *Online Submission*. Retrieved from http://eric.ed.gov/?id=ED490073

Fang, I. E. (1968). By Computer: Flesch's Reading Ease score and a syllable counter. *Behavioral Science*, *13*(3), 249–251. http://doi.org/10.1002/bs.3830130312

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*(3), 221–233.

Fowler, F. J. (1995). *Improving Survey Questions: Design and Evaluation* (1st edition). Thousand Oaks: SAGE Publications, Inc.

Groves, R. (1989) *Survey errors and survey costs.* New York, NY: Wiley.

Hochhauser, M. (2005a). Liabilities of "unreadable" consent forms. In E. F. Gabriele & V. J. Ducker (Eds.), *2005 Symposium Proceedings* (pp. 115-122). Arlington, VA: Society of Research Administrators, International.

Hochhauser, M. (2005b). What readability expert witnesses should know. *Clarity*, *54*, 38-42.

Ley, P., & Florio, T. (1996). The use of readability formulas in health care. *Psychology, Health & Medicine*, *1*(1), 7-28.

McClure, G. M. (1987). Readability formulas: Useful or useless? *IEEE Transactions on Professional Communication*, *PC-30*(1), 12–15.

Mailloux, S. L., Johnson, M. E., Fisher, D. G., & Pettibone, T. J. (1995). How reliable is computerized assessment of readability?. *Computers in nursing*, *13*(5), 221-221.

Osborne, H. (2000). In other words… Assessing readability… Rules for playing the numbers game. *On Call, 3*, 38-39. Retrieved from http://www.healthliteracy.com/assessingreadability.

Payne, S. (1951). *The art of asking questions.* Princeton, NJ: Princeton University Press. SPSS Inc. Released 2009. PASW Statistics for Windows, Version 18.0. Chicago: SPSS Inc.Terris, F. (1949). Are poll questions too difficult?. *Public Opinion Quarterly*,*13*(2), 314-319.

Thomas, G., Hartley, R. D., & Kincaid, J. P. (1975). Test-retest and inter-analyst reliability of the Automated Readability Index, Flesch Reading Ease Score, and the Fog Count. *Journal of Reading Behavior*, *7*(2), 149–154.