Using Readability Tests to Improve the Accuracy of Evaluation Documents Intended for Low-Literate Participants

Julien B. Kouamé
Western Michigan University

Background: Readability tests are indicators that measure how easy a document can be read and understood. Simple, but very often ignored, readability statistics cannot only provide information about the level of difficulty of the readability of particular documents but also can increase an evaluator's credibility.

Purpose: The purpose of this article is two-fold: (1) to provide readers with logical reasons for using readability tests and (2) how to choose the right test for a project.

Setting: United States.

Research Design: A comparative framework is used to present the need for readability testing.

Keywords: *Readability tests, evaluation instruments, survey research, low-literacy survey*

Increasingly, survey critics reproach researchers for using language at levels too advanced and therefore inappropriate to their intended audience. In 1993, the National Adult Literacy survey indicated that about 50% of the public cannot read and understand information in even short publications (Association of Medical Directors, 2004). An article published by Calderón et al. (2006) stated that major survey tools such as the quality-of-life survey" used in health institutions, present variations in readability levels between items. To enable their patients to understand most of the evaluation documents, the Association of Medical Directors (2004) suggested that researchers and evaluators adapt their instruments to between a seventh and eighth grade level, which is the average adult American reading ability (Kirsh et al., 1993). More than ever, new evaluators as well as professional evaluators borrow instruments from other languages. Often, these instruments cannot be used in other linguistic settings without major modifications that account for comprehension. Consequently, evaluation instruments lose their ability to accurately measure outcomes because of linguistic inaccessibility. Most of these concerns, however, can easily be addressed by considering the use of a readability test as part of the content analysis process of the evaluation instrument.

Today, despite the multiplicity of readability formulas, only a few of these are being used, some of which are incorporated into computer software to facilitate their use. Provocative questions concern the selection of the appropriate formula, the method by which to interpret a result, and the need to continue the conversation regarding the use of readability testing. This paper presents attempts to address these questions.

Readability Tests for Test Validity and Report Clarity

Readability tests are indicators that measure how easy a document is to read and understand. For evaluators, readability statistics can be solid predictors of the language difficulty level of particular documents. The essential information in an evaluation document should be easily understandable. A proper readability level of the evaluation document will greatly prevent frustration for the project's participants. In short, evaluation documents presenting difficult items in a survey could lead to nonresponse, missing data points, or "unreliable responses because of a mismatch between item readability and the reading skills of the respondent" (Calderón et al., 2006). The implementation of readability tests prior to pilot testing results in the more efficient use of evaluators' time, a critical resource. Readability testing can also increase the validity and reliability of data collection instruments as well as the credibility of the evaluator.

To illustrate this point, I use an example from a study I conducted in fall 2006. The goal of this project was to develop and evaluate a simple and understandable survey for formative evaluation and to assess the effect of the readability test on low-literate participants.

A child abuse evaluation survey was borrowed for this assessment. The evaluation was conducted with 65 low-literate participants (10 years of formal schooling) for whom English was their second language. Participants were randomly assigned into two groups of 33 and 32 individuals. One group used a form of the survey in which the content was tested to suit the readability level by using the Flesch–Kincaid formula. Table 1 shows the Flesch-Kincaid grade level for each question across the two forms of the survey. Participants were also asked to evaluate instructions and the understandability of each item on a scale of 1 to 10, where, 1 describes an item that is easy to read and 10 describes an item that is difficult to read. For each group, the understanding level was calculated. The descriptive statistic is provided in Table 2. Also a frequency of rating is provided in the Figure 1 and Figure 2. On average, version 1 has a grade level (GL) of 9.8 (between 9 and 10 grade) compared to 5.22 for version 2. The participants' rating shows that the two documents were generally well understood (see Table 2). However, the document with the readability test presents a better understandability score.

Table 1
Flesch-Kincaid Grade Level
by Survey Question

	Flesch-Kincaid Grade Level	
Questions	Survey Form 1	Survey Form 2
Item 1	8.1	3.6
Item 2	10.4	5.2
Item 3	8.2	0.7
Item 4	7.6	0.7
Item 5	7.3	3.6
Item 6	13.0	5.8
Item 7	15.4	11.3
Item 8	14.2	9.0
Item 9	9.0	6.2
Item 10	10.9	9.0
Item 11	2.2	2.2
Item 12	11.7	5.4
Average	9.83	5.22

The results show a better readability of the survey after the revision following the Flesch-Kincaid test. The version 1 has a rating mean of 4.43 with a standard deviation of 1.4. The version 2 received a lower mean equal 3.60 with a standard deviation of 1.3.

Table 2
Descriptive Statistics for Survey Rating

	N	*M*	*SD*	Min	Max
Survey Form 1	30	4.43	1.36	2	7
Survey Form 2	30	3.60	1.33	2	7

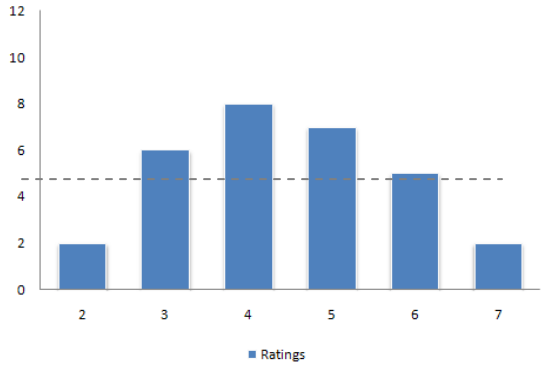

Figure 1. Distribution of Rating of Survey Form 1

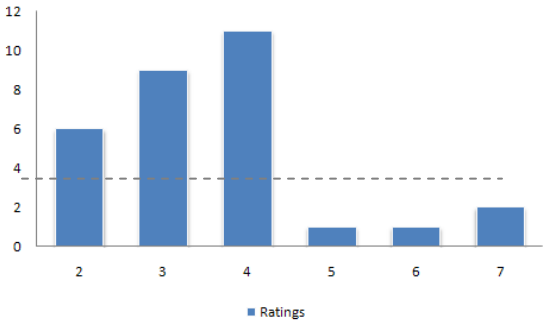

Figure 2. Distribution of Rating of Survey Form 2

Although the Flesch-Kincaid test shows that the GL is higher than the target population, the rating of the participants shows that the survey could be understood by the users. The following may be the principal reason justifying the difference between the two results. The questions are tested individually; therefore, the software cannot relate them to each other. While the readability test provides a high GL, the participants may not have difficulty understanding because they take the context of the writing into consideration. However the test was useful as it helped revising the survey for easy reading.

Critiques of Readability Formulae

There is no doubt that readability testing has always been at the center of controversy. In "The Principles of Readability," DuBay (2004) listed numerous papers that criticized readability testing. These papers, as Dubay wrote, have titles such as "Readability: A Postscript" (Manzo 1970), "Readability Formulas: What's the Use?" (Duffy 1985) and "Last Rites for Readability Formulas in Technical Communication" (Connaster 1999). Still others suggest the idea of usability as an alternative for readability. However usability testing is not able to provide an objective prediction of text difficulty (Dubay, 2004). Before other reasonable alternatives are invented, readability testing for text reading level prediction remains essential. This point is well illustrated in my study mentioned above.

Simple Skill, Often Forgotten

In Scriven's (2007) Key Evaluation Checklist (KEC) he urges professional evaluators not to ignore readability testing. Not only do we have the commitment to provide our customers with accurate information, but we also have the obligation to present those findings in a report that is easily understandable.

In the summer of 2006, I was part of a team of 10 evaluators to metaevaluate five evaluation reports by other evaluators, who had at least 2 years of experience. The project was initiated by the Department of Educational Studies of my school and supervised by an expert in evaluation. The objective was to provide opportunities for evaluators to further develop evaluation skills (i.e., build evaluation capacity) that will improve evaluation reports and the practice of evaluation. The metaevaluators judged the evaluation reports using criteria for sound evaluations according to The Program Evaluation Standards (Joint Committee on Standards for Educational Evaluation, 1994). The metaevaluation focused on the extent to which the five reports individually and collectively met the requirements for utility, feasibility, propriety, and accuracy. For the purpose of this paper only a part of the U5 standard in the Stufflebeam's (1999) Metaevaluation Checklist is presented.

As a result of this metaevaluation, we found that only one evaluation met the following two criteria of the comparison analysis: 1. *Provides definition of terms used* and 2. *Uses audience-appropriate language, tables, and graphs.* However none of the evaluators provided information about testing the readability level of their documents (research tools and reports). The judgment made by the metaevaluators was based not only on their own ability to read and understand the reports, but also by the implementation of readability testing to account for the skills of the target population. Although the report entitled *Improving Asset Utilization* was difficult to understand, the evaluator not only provided definitions of difficult expressions but also clearly disclaimed that language used in the report was familiar to his client. Despite the fact that this assessment was done using only 5 evaluation reports, it clearly illustrates that the readability testing is not always used by evaluators.

Selecting an Appropriate Formula

Recently I was talking to a friend about my interest in readability testing of evaluation documents. He agreed that this is a "must be done" act for any researcher. Then my friend continued the discussion by saying that all of the formulas do not provide the same reading level. The most important question he asked during our conversation was the following: "How should we select an appropriate test (formula)?" The next section will attempt to answer this important question.

Indeed more than 200 readability test formulas were invented since the 1940s. However, only a few of these are currently being used. The recent study led by Calderón (2006), presented in Table 3, is the result of a literature search on survey readability. The Literature search was conducted in 2005 by examining major medical publications: Medline (1966-2003), CINAHL (1982-2003), ClinPSYC (1993-2003), and PsychInfo (2003-2005). The 17 articles reported focused on readability and presented methods for estimating readability scores. Table 3 shows that Flesch-Kincaid and the Flesch Reading Ease are the most common formulas used to assess readability (Calderón et al., 2006). These formulas are the most widely used because they are the most reliable formulas. Especially respected is the Flesch Reading Ease formula because it is the most tested and the most reliable (Chall 1958, Klare 1963). In addition, the formula is incorporated in the Microsoft Word software which favors its ease-of-use factor.

Table 3
Publications on Survey Readability: Methods, Application to Test, and Scores

Author, Year	Survey Theme	Readability Test	Application	Scores
Berndt, 1983	Depression inventories–5	FRE, Gunning	Not indicated	5th-12th grade
Price, 1985	Obesity	SMOG	Response options	9th grade
Jensen, 1987	Marital surveys–9	Forbes-Cottle	Instructions and questions	6th-college
Devins, 1990	Kidney knowledge	Multiple methods	Not indicated	9th grade
Macey, 1991	Critical care family needs	Gunning-Fox	Not indicated	9th grade
Paolo, 1993	Dissociative experiences	F-K	Questions	10th grade
Beckman, 1997	Clinical outcomes surveys–5	FRE	Instructions and questions	8th-9th grade
Edlund, 1997	Health plan satisfaction	F-K	Not indicated	6th grade
MacDiarmid, 1997	Urological symptoms	Dale-Chall	Survey as a whole	6th grade
Eaden, 1999	Colitis knowledge	FRE	Not indicated	74.3
Pande, 2000	Osteoporosis	FRE	Not indicated	74.3
Heyland, 2001	Intensive care unit satisfaction	F-K	Not indicated	6th grade
Kimble, 2001	Cardiac	F-K	Not indicated	4th grade
Rowan, 2001	Foot pain	FRE	Not indicated	82.4
Otley, 2002	Pediatric inflammatory bowel disease	F-K	Not indicated	5th grade
Dolovich, 2004	Foot pain	FRE	Not indicated	82.4
Travess, 2004	Pediatric inflammatory bowel disease	F-K	Not indicated	5th grade

FRE = Flesch Reading Ease; SMOG = Simple Measure of Gobbledygook; F-K = Flesch-Kincaid Formula.

However, good research practice suggests that we use several methods for testing because error is inevitable. Therefore my suggestion is that we use the combination of more than one formula to assess the reading level of our evaluation documents. Using more than one test provides greater insight into the document. Be reminded that any measurement is susceptible to error.

Indeed errors are the essence of the field of measurement. Some of the readability formulas tend to predict higher scores than others. This is the case for the SMOG and the Fog formulas. Users of the readability formulas find discrepancy between the formulas because each of them are constructed with a specific objective in mind. Therefore, "Different uses of a text require different levels of difficulty" (DuBay, 2004). For example, while FORCAST provides a good prediction for non running narrative (Questionnaire, Form), FOG is widely used for running text in the health care and general insurance industries for general business publications. To illustrate the discrepancy among readability tests, I tested the present paragraph using Flesch and SMOG. The result is shown below. In addition, I provide in Table 4 a list of frequently used readability test formulas with what they test the best.

Table 4
Suggested Usage of Common Readability Formulas

Test Name	Usage	Formula
Flesch Grade Level	Most reliable when used with upper elementary and secondary materials.	$GL = (0.39 * ASL) + (11.8 * ASW) - 15.59$
Flesch Reading Ease	Most reliable when used with upper elementary and secondary materials.	$RE = 206.835 - (1.015 * ASL) - (84.6 * ASW)$
FORCAST	Focuses on functional literacy. Used to assess non- running narrative, e.g. questionnaires, forms, tests ...	Grade level = 20 – (N ÷ 10) Where N = number of single-syllable words in a 150-word sample.
Fry Graph	Used over a wide grade range of materials, from elementary through college and beyond.	
Gunning FOG	Widely used in the health care and general insurance industries for general business publications, the Navy.	$GL = \left(\frac{EasyWords + 3(HardWords)}{Sentences} - 3 \right) \div 2$
New Dale-Chall	The "new" (1995) version of the Dale-Chall formula. A vocabulary-based formula normally used to assess upper elementary through secondary materials.	$RawScore = (0.1579 * PDW) + (0.0496 * ASL) + 3.6365$
Powers-Sumner- Kearl	Used in assessing primary through early elementary level materials.	$GL = (0.0778 * ASL) + (0.0455 * NS) - 2.2029$
SMOG	Unlike any of the other formulas, SMOG predicts the grade level required for 100% comprehension.	$SMOG = 3 + \sqrt{PolysyllableCount}$

As can be seen from the formulas in Table 4, manual calculation of reading level, can be boring, complex and sometimes time consuming because words, sentences, and paragraphs must be counted. Fortunately, many of these formulas are incorporated in software applications to make them easy to use. Unfortunately, not all the applications provide reliable results.

Here is a list of concerns you should have in mind when you thinking about using such tools:

1. There are many free tools to consider. However, sometimes free is also cheap in value.
2. Use of more than one testing tools will provide you with a significant knowledge of your document

3. Consider also a visual display of the result of your test. This will help you to know what to focus your revision on.
4. The tool should help to locate the best test for your audience.
5. Establish the credibility of the author (s) of the tool.
6. Test the accuracy of the tool by testing the reading level of sentences such as "The students saw Mrs. Kate during the recess." Some software may consider this as 2 sentences. If this happens, the software may not reconsider you choice.

Readability testing should be part of all evaluators' projects, even when those individuals are internal evaluators and think they know the common language used in the institution. The final report of an evaluator can be disseminated to the public at large, not only to an internal constituency. Therefore the evaluator should not only have the client in mind while working for her/him but should also think of the audience (Scriven, 1991). Scriven has even suggested that reports be field tested to suit the target. Whether you are writing a proposal, the first question of a research tool or a report, it is a valuable habit to pretest for readability.

You should not look at this just in terms of KEC, as Scriven advises, but should consider that cultivating this habit will also save considerable time in revision and even make you a better evaluator.

The KEC put a focus on testing frequent consultation with stakeholders and audience before, during, and after we have developed any evaluation document. Assessing both the reading ability of the audience and the readability of the text will greatly facilitate this process. The field test suggested by Scriven will allow you to find out if your document suits the target. If it does not, you have to review and test again. But by subjecting your documents to readability testing, you will predict (know) the reading level and save one or two stages of field testing.

Readability testing will become more necessary than ever before because of the multiple layers of reading capability within our diverse society. As with any tool, it can only do best what it is designed to do. Despite the limits of readability formulas, they remain a unique way to predict the extent to which documents can be comprehended by their intended target. However researchers have demonstrated that although readability testing is relatively simple, it is often forgotten.

References

Association of Medical Directors. (2004). Comprehension and reading level. Retrieved February 20, 2008, from http://www.informatics-review.com/FAQ/reading.html

Calderón, J. L., Morales, L. S., Liu, H., & Hays, R. D. (2006). Variation in the readability of items within surveys. *American Journal of Medical Quality, 21*(1), 49-56.

Chall, J. S. (1958). *Readability: An appraisal of research and application*. Columbus, OH: Ohio State University Press.

DuBay, W. (2004). *The principles of readability*. Retrieved July 24, 2008, from http://www.impact-information.com.

FreadabilityFormulas.com. *Can YOU read me now?* Retrieved July25, 2008, from http://www.readabilityformulas.com/free-ebooks.php.

Kirsh, I., Jungeblut, A., Jenkins, L., & Kolstad, A. (1993). *Adult literacy in America: A first look at the results of*

the National Adult Literacy Survey. Washington, DC: US Department of Health, Education and Welfare.

Powers, R. D., Sumner, W. A., & Kearl, B. E. (1958). A recalculation of four adult readability formulas. *Journal of Educational Psychology*, *49*(2), 99-105.

Scriven, M. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage..

Scriven, M. (2007). *Key evaluation checklist*. Kalamazoo, MI: The Evaluation Center, Western Michigan University.

Stufflebeam, D. L. (1999). *Program evaluation models metaevaluation checklist*. Kalamazoo, MI: The Evaluation Center, Western Michigan University.