Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty?

Lenzner, Timo

Empfohlene Zitierung / Suggested Citation:
Lenzner, T. (2014). Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty? *Sociological Methods & Research*, *43*(4), 677-698. https://doi.org/10.1177/0049124113513436

Are Readability Formulas Valid Tools for Assessing Survey Question Difficulty?

Timo Lenzner[1]

Abstract

Readability formulas, such as the Flesch Reading Ease formula, the Flesch--Kincaid Grade Level Index, the Gunning Fog Index, and the Dale-Chall formula are often considered to be objective measures of language complexity. Not surprisingly, survey researchers have frequently used readability scores as indicators of question difficulty and it has been repeatedly suggested that the formulas be applied during the questionnaire design phase, to identify problematic items and to assist survey designers in revising flawed questions. At the same time, the formulas have faced severe criticism among reading researchers, particularly because they are predominantly based on only two variables (word length/frequency and sentence length) that may not be appropriate predictors of language difficulty. The present study examines whether the four readability formulas named above correctly identify problematic survey questions. Readability scores were calculated for 71 question pairs, each of which included a problematic (e.g., syntactically complex, vague, etc.) and an improved version of the question. The question pairs came from two sources: (1) existing literature on questionnaire design and (2) the Q-BANK database. The analyses revealed that the readability formulas often favored the problematic over the

[1] GESIS—Leibniz Institute for the Social Sciences, Mannheim, Germany

Corresponding Author:
Timo Lenzner, GESIS—Leibniz Institute for the Social Sciences, P. O. Box 12 21 55,
Mannheim 68072, Germany.
Email: timo.lenzner@gesis.org

improved version. On average, the success rate of the formulas in identifying the difficult questions was below 50 percent and agreement between the various formulas varied considerably. Reasons for this poor performance, as well as implications for the use of readability formulas during questionnaire design and testing, are discussed.

Keywords
survey question difficulty, survey question design, survey question testing, readability formulas, question wording, survey pretesting, readability

Introduction

Survey designers have long been concerned about the language complexity of the questions they ask and its impact on data quality (Cantril 1944; Payne 1951). Questions that are difficult to understand contribute to irrelevant variance and are thus important sources of measurement error (Fowler 1995; Groves 1989). To avoid asking difficult (or less comprehensible) questions, survey designers must somehow determine the language complexity of their questions, identify problematic ones, and modify these appropriately.

One way to assess textual difficulty is to use readability formulas. These formulas generate numerical estimates of the readability of a text, where readability is defined as readers' ‘‘[e]ase of understanding, owing to the style of writing" (Klare 2002:681). Based on this definition, the terms readability, comprehensibility, and (reading) difficulty are used interchangeably in this article to refer to the effort required to understand the meaning of a text. Readability formulas typically involve two measurable aspects of a text, such as word length and sentence length, and a weighted combination of both aspects yields a score for the text, representing either its relative difficulty or the grade level required to understand it (Bruce and Rubin 1988). For example, the Flesch-Kincaid Grade Level index (FKG, Flesch 1979) is based on the average number of words per sentence (sentence length) and the average number of syllables per word (word length). Higher scores mean that texts are harder to understand and require a higher grade level. The FKG score for the English language[1] is computed as follows:

$$\text{FKG score} = 0.39 \times \left(\frac{oof\ words}{oof\ sentences}\right) + 11.8 \times \left(\frac{oof\ syllables}{oof\ words}\right) - 15.59.$$

Because of their quantitative nature and their apparent precision, readability formulas, such as the Flesch Reading Ease formula (FRE, Flesch 1948), the FKG (Flesch 1979), the Gunning Fog index (FOG, Gunning 1952), and the Dale-Chall formula (DC, Dale and Chall 1948) are often considered to be objective measures of text comprehensibility. Not surprisingly, survey researchers have frequently used readability scores as indicators of question difficulty (e.g., Converse 1976; Converse and Schuman 1984; Gafke and Leuthold 1979; Harmon 2001; Holbrook et al. 2007; Kimball and Kropf 2005; Payne 1949; Terris 1949; Velez and Ashworth 2007).

For example, Payne (1949) conducted a split-ballot experiment in which he randomly switched the order of the response options of 16 attitudinal questions. He applied the FRE formula to examine the impact of question difficulty on response order effects. The seven questions in which a significant response order effect (i.e., a recency effect) occurred were also rated by their FRE scores as more difficult than the other nine questions, for which no order effects were observed. Terris (1949) applied the FRE and the DC formulas to the questions in three nationwide surveys and found that over 90 percent of the questions examined were too difficult for more than 10 percent of the U.S. population. Even though Terris duly noted that both formulas had not been designed for short texts, such as survey questions, she argued that they still ‘‘allow us to distinguish various degrees of difficulty among the questions." Terris (1949:315) Finally, Kimball and Kropf (2005) computed Flesch-Kincaid Grade Level scores of ballot instructions used in 250 counties during the 2002 general election and found that the rates of unrecorded votes were higher in counties where the voting instructions were formulated at a higher grade level.

Given that several studies report a clear relationship between readability scores and various indicators of question difficulty, some researchers have suggested applying the formulas during the questionnaire design or pretesting stage. For example, Velez and Ashworth (2007) examined the impact of item readability (measured by FKG) on the amount of midpoint responses in an establishment survey and found a significant positive correlation between FKG scores and the tendency to provide midpoint responses. Hence, the authors suggested applying the formula during the questionnaire design phase to identify problematic items and to assist survey designers in revising flawed questions. While not explicitly computing readability scores, Holbrook et al. (2007) measured question comprehension difficulty by a composite index of(1) the number of sentences in the question (i.e., question length), (2) the number of words per sentence (i.e., sentence length), and (3) the number of letters per word (i.e., word length). Using data from 548 experiments in telephone surveys, they found that questions with longer sentences and longer words were associated with larger response order effects, particularly among less educated respondents. Consequently, the authors argued against the use of long sentences involving many multisyllabic words and advised questionnaire designers to calculate ‘‘the reading difficulty level of a question" (p. 341).

While these findings apparently lend support to the validity of readability formulas as measures of question difficulty, other studies have come to less conclusive results. As a consequence, some researchers have taken a more critical stance toward the usage of readability formulas. For example, Converse (1976) examined a large number of survey questions and did not find a significant correlation of FRE scores with the percentage of no opinion answers (i.e., ‘‘don't know" or ‘‘no answer"), even when he confined his analysis to the respondents with the lowest education level. Thus, he concluded that ‘‘[p]ossibly the Flesch measure itself, standardized on 100-word samples of written words, is not entirely apt for survey questions [...]" (p. 522). Harmon (2001) assessed the difficulty of questions in four data sets by means of three readability formulas (FRE, FKG, and FOG). In two of the four data sets, greater difficulty (as scored by all three readability formulas) was significantly correlated with higher percentages of don't know (DK) responses; in one data set, no significant relationship was found; and in the fourth data set, the FOG scores were even significantly and negatively correlated with the percentage of DK replies. Moreover, in two data sets, he found that endorsement of the midpoint answer category decreased with increasing levels of question difficulty. Holbrook, Cho, and Johnson (2006) used the FKG formula, among other indicators, as a measure of question difficulty and reported a nonlinear relationship between the grade level required to understand the questions and comprehension difficulties (as identified by behavior coding of interactions that were tape-recorded during face-to-face interviews). Although higher grade levels were associated with more comprehension difficulties, this was not the case at the highest grade levels. While the authors did not question the validity of readability formulas, they acknowledged that this result might be due to the fact that more complex questions (as measured by average sentence and word length) may not necessarily increase comprehension difficulty but may, rather, communicate the meaning of the question and its purpose more clearly (and thereby actually decrease question difficulty).

Whereas survey researchers have seemingly embraced readability formulas as readily available tools for assessing question difficulty, the last decades have also witnessed a growing concern about their validity among linguists and reading researchers (e.g., Davison and Green 1988; Templeton, Cain, and Miller 1981; Wheeler and Sherman 1983). In particular, the formulas have encountered severe criticism because most of them are based on only two variables (word length/word frequency and sentence length) that themselves may not even be very good predictors of language difficulty (Anderson and Davison 1988). Moreover, it has been argued that the assumptions underlying the correct use of the formulas are often violated, for example, by applying them to short texts, such as survey questions (Bruce and Rubin 1988; Oakland and Lane 2004). The purpose of this article is twofold. First, it critically reviews the linguistic criteria underlying most readability formulas and highlights potential problems in applying the formulas to survey questions. Second, it reports on an empirical study that examined the validity of four readability formulas (FRE, FKG, FOG, and DC) for identifying difficult survey questions.

Linguistic Criteria Underlying Readability Formulas

Readability scores are usually based on two of the following three linguistic variables: word length, word frequency, and sentence length. Usage of shorter words, commonly used words, and shorter sentences is supposed to make texts easier to understand. However, a closer examination of these variables indicates that they have serious limitations and may not constitute a satisfactory basis for assessing textual difficulty.

Word Length

Most readability formulas presume that word length has a direct effect on the ease with which a text can be read: The longer a word is, the more difficult it is to comprehend. Word length is operationalized in one of two ways, either by the average number of syllables per word (e.g., FRE and FKG) or by the number of polysyllabic words, that is, words consisting of three or more syllables (e.g., FOG). No matter which approach a formula adopts, the central assumption is that words become more difficult to comprehend with an increasing number of syllables. Intuitively, this notion seems plausible, if not obvious. The simple fact that additional syllables require the processing of additional input may account for this increasing difficulty. However, there is some evidence suggesting that this view might be overly simplistic.

Numerous psycholinguistic studies that examined the effect of the number of syllables per word on word recognition found an inhibitory effect for nonwords (or pseudowords) but not for real words (Ans, Carbonnel, and Valdois 1998; Juphard, Carbonnel, and Valdois 2004; Weekes 1997). Other studies found an inhibitory effect for low-frequency words but not for high-frequency words (Ferrand 2000; Ferrand and New 2003; Jared and Seidenberg 1990). These results are compatible with two major theories of visual word recognition (multiple-trace memory model, Ans, Carbonnel, and Valdois 1998 and dual-route cascaded model, Coltheart et al. 2001), which posit that reading relies on two distinct procedures: Words that are familiar to the reader (i.e., words of high frequency) are processed globally, as a unit, whereas unfamiliar words (i.e., words of low frequency and nonwords) are processed sequentially, syllable by syllable. Consequently, the effect of word length on reading difficulty is moderated by word frequency, with words of high and medium frequency having no detrimental effect on comprehension (irrespective of their length). All in all, many research findings suggest that word frequency plays a more fundamental role in word recognition than word length (cf. New et al. 2006).

In addition, a closer look at the vocabulary of English suggests that many short (i.e., mono or bisyllabic) words are more difficult to comprehend than long (i.e., polysyllabic) words (Bailin and Grafstein 2001). This is partly due to the fact that a large number of English words are derivatives and compounds. Derivatives are the result of affixation, that is, the construction of a new word by adding prefixes (e.g., pre-, co-, mis-, un-, anti-) or suffixes (e.g., -er, -ing, -ion, -ness, -ism) to an existing word. Since speakers of a language usually know the function of these affixes, most derivatives are semantically transparent, that is, their word parts give clues to their meaning. Consider, for example, the derivatives unemployment, helplessness, or organization in comparison to the monosyllabic apt, dearth, feint, or queue. Despite their greater length, the polysyllabic derivatives are presumably easier to comprehend than the monosyllabic words. However, readability formulas including syllable counts would favor the shorter, yet less familiar, words.

The formulas also encounter problems when assigning a readability score to a text that includes compounds, that is, words constructed by combining two existing words (e.g., safeguard, overweight, playground). Again, many of these are quite easy to comprehend because people usually know the meaning of the word parts. About three decades ago, Nagy and Anderson (1984) estimated that there were about 240,000 words in printed school English, of which about 182,000 words were semantically transparent derivatives and compounds. Current estimates from the Oxford English Dictionary (http://oxforddictionaries.com) suggest that there are at least 250,000 words in English, so the numbers estimated by Nagy and Anderson may still be quite accurate. This means that, in English, long words are not necessarily, or even usually, difficult to understand (cf. Anderson and Davison 1988).

Word Frequency

Some readability formulas, such as DC and Spache (1953), rely on word frequency lists to assess the semantic difficulty of a text. These word lists contain a certain number of frequently used words and words that do not appear on these lists are flagged as ‘‘difficult" or ‘‘rare." The number of rare words is then divided by the total number of words to yield an estimate of the semantic difficulty of the text. This approach is based on the assumption that words that occur less often in written or spoken language are more difficult to understand than words that are used more frequently.

Indeed, ample empirical evidence shows that low-frequency words are more difficult to comprehend than high-frequency words (e.g., Inhoff and Rayner 1986; Williams and Morris 2004). This word frequency effect has not only been identified in virtually every measure of word recognition, but also in survey question comprehension (Lenzner, Kaczmirek, and Galesic 2011). Nevertheless, there are two major problems with the ways in which readability formulas employ the concept of word frequency.

First, vocabulary tends to change rapidly, with new words entering the language and other words going out of use. This means that word lists, which were compiled at a specific point in time and thus representative of the language used in that particular era, become outdated relatively quickly. The original DC word list, which was compiled in 1948, consisted of 763 words and was updated and extended to roughly 3,000 words in 1995 (Chall and Dale 1995). However, the more recent list still includes words such as hairpin, maypole, cobbler, and washtub, which today are likely to be unfamiliar to younger readers because they do not relate to their current realities (Bailin and Grafstein 2001). At the same time, frequently used words that have entered the language during the last decade, such as Internet, download, or ringtone, do not appear on the list and are thus treated as ‘‘hard words" by the DC formula. It is important to note that simply updating word frequency lists on a regular basis by removing outdated (old-fashioned) words and including new ones would not eliminate this problem of vocabulary change. The old-fashioned words would still be as familiar to older generations as the new words would be to youngsters. Hence, a fundamental problem of these word frequency lists is that they ignore the fact that different sociocultural groups can have quite different vocabularies.

Second, because the DC and Spache word lists consist of relatively few words (2,946 and 1,040, respectively), they do not pay appropriate attention to the fact that many English words are derivatives and compounds. As argued above, these are often semantically transparent, because their constituents provide clues to their meaning. Hence, most derivatives and compounds are as easy (or difficult) to comprehend as the word stems they are derived or constructed from. In a few instances, the lists do account for this fact. For example, the DC word list includes the stem ache as well as the derivative aching and the two simple words bath and room as well as the compound bathroom. However, given that there are approximately 182,000 derivatives and compounds in printed school English (see above), it is clear that the vast majority of these words are missing in a list of roughly 3,000 words. Thus, while satisfactory, major, ear, and ring are treated as familiar words by the DC formula, satisfied, majority, and earring are considered to be unfamiliar and rare.

All in all, readability formulas treat word frequency as an indicator of the absolute comprehensibility of a word that can be determined by the presence or absence of the word on a given list. Word frequency, however, might better be conceived as an indicator of the relative comprehensibility of a word that can only be assessed in relation to other words. For example, it is reasonable to improve the comprehensibility of a text by consulting linguistic thesauruses and replacing low-frequency words with higher frequency synonyms (cf. Lenzner 2011). If higher frequency synonyms exist, these are almost certainly easier to understand than the low-frequency words. The presence or absence of a word from a relatively short word frequency list, however, is a considerably weaker and more error-prone indicator of word difficulty.

Sentence Length

Virtually all of the popular readability formulas use average sentence length (number of words) as an indicator of the syntactic difficulty of a text. Similar to the supposed effect of word length on semantic difficulty, usage of this indicator rests on the assumption that there is a strong correlation between the length of a sentence and its syntactic complexity, with longer sentences being more complex and difficult to comprehend than shorter ones. Again, this notion seems intuitively plausible, given that longer sentences require readers to process more input than shorter sentences. However, previous research suggests that sentence length by itself is not a source of comprehension difficulty (Davison and Green 1988). The ways in which words are combined to form a sentence (i.e., the syntactic structure of sentences) seem to be more important than the sheer number of words.

For example, sentences with left-branching syntax are more difficult to comprehend than sentences with right-branching syntax (Fodor, Bever, and Garrett 1974; Lenzner, Kaczmirek, and Galesic 2011). Left-branching syntax means that readers have to process many clauses and qualifiers before they encounter the predicate of the main clause. These structures require readers to remember information about the main clause while they process the embedded clauses. In contrast, sentences with right-branching syntax are easier to process because they first present the main clause and subsequently add clauses and phrases that qualify it. In the following example, question (1) has left-branching syntax whereas question (2) asks the same question with right-branching syntax:

(1): How likely is it that if a law was considered by parliament that you believed to be unjust or harmful, you, acting alone or together with others, would try to do something against it?

(2): How likely is it that you, acting alone or together with others, would try to do something against a law that was considered by parliament and that you believed to be unjust or harmful?

In question (1), respondents have to process 28 words and retain information from five propositions[2] before encountering the main predicate of the main clause (try to do). In contrast, question (2) requires respondents to process only 13 words and to retain only two propositions[3] before they encounter the main predicate. Hence, question (2) is much easier to comprehend than question (1). Nonetheless, given that both questions consist of the same number of words, readability formulas treat them as equally easy/difficult to understand.

Problems With Applying the Formulas to Survey Questions

The formulas' ability to correctly assess survey question difficulty may be limited by the fact that most of them have not been designed for estimating the readability of short texts, such as survey questions. Usually, the formulas require at least a 100-word passage for proper implementation (Homan, Hewitt, and Linder 1994) and, given that survey questions are rarely of that length, it is likely that inaccuracies in the ratings of some questions occur simply because the formulas are used in a way that was not intended by their originators. Of course, it is possible to group several survey questions together to obtain a text of sufficient length to apply readability formulas. However, this practice does not yield information about the readability of individual questions and is thus of limited value (cf. Oakland and Lane 2004).

The only formula that has been specifically designed to estimate the readability of individual sentences is the Homan-Hewitt readability formula (Homan, Hewitt, and Linder 1994). This formula includes three predictor variables: (1) number of unfamiliar words, measured by familiarity ratings listed in The Living Word Vocabulary (Dale and O'Rourke 1981); (2) number of long words, measured by determining how many words have more than six letters; and (3) sentence complexity, measured by the average number of words per Hunt's T-Unit. Hunt's (1965:141) T-Unit refers to the smallest word group that could be considered a grammatical sentence and is defined as ‘‘one main clause plus all the subordinate clauses attached to or embedded in it." Even though Homan, Hewitt, and Linder (1994) published validation results for their formula, it has, to date, been rarely adopted by test developers or questionnaire designers. Moreover, it has been argued that the Homan-Hewitt formula was designed for and validated with reading material adequate for elementary school students and may thus be inappropriate for estimating the readability of adult-level texts (Badgett 2010). Finally, the Homan-Hewitt formula has not yet been incorporated into commercially available computer software, so that its application is both time consuming and prone to human error. On these grounds, the Homan-Hewitt formula has been excluded from the present study's methodology.

Data and Method

The purpose of the current study was to examine whether four of the most popular readability formulas (FRE, FKG, FOG, and DC) correctly identify problematic survey questions, and hence whether they are valid tools for assessing question difficulty. Readability scores were calculated for a set of question pairs, each of which contained a problematic (e.g., syntactically complex) and an improved version of the question. The study examined whether the formulas assigned different readability scores to the two versions and correctly classified them as being more or less difficult.

The decision to examine question pairs was based on two considerations. First, readability scores should either favor one version over the other or indicate that both versions are equally difficult. Hence, provided that both versions did indeed differ in their difficulty (as was suggested by the sources from which they were taken, see next paragraph), this strategy allowed us to assess the performance of the readability formulas directly (without having to rely on proxy indicators of question difficulty, such as item nonresponse or number of midpoint responses). Second, using two versions of the same question may also mimic the situation in which questionnaire designers try to improve their draft questions during the questionnaire construction phase. Given that some authors have proposed using readability formulas for exactly this purpose (e.g., Holbrook et al. 2007; Velez and Ashworth 2007), it is important to check whether the formulas really help to write better questions.

The questions examined in this study came from two sources. First, a literature search was conducted to identify journal articles and textbooks on questionnaire design that report examples of problematic survey questions, together with recommendations for rewording and improving these questions. In order to be included in the analysis, these question pairs needed to consist of a problematic and an improved question version that were explicitly tagged as such, and the problematic aspects of the questions had to relate to the question itself, not to the response options. Moreover, only question pairs were included in which the improved version consisted of one question (i.e., question pairs in which the improved version recommended dividing one question into several questions were excluded). This search resulted in 15 question pairs reported in five publications (Bassili and Scott 1996; Fowler 1992, 1995, 2004; Fowler and Cosenza 2008). Although not comprehensive, this literature search identified several question pairs that may generally serve as prototypical examples of problematic and revised survey questions.

Second, the Q-Bank database (www.cdc.gov/qbank) was searched for question pairs fulfilling the same criteria as the ones identified in the literature search. Q-Bank is an online database that houses pretested survey questions, together with the original evaluation reports that contain specific findings about the questions. Questions can be searched by question topic (e.g., demographics, health), survey title, testing agency, and response error (e.g., problematic terms, ambiguous concepts, biased/sensitive). A response error search (wwwn.cdc.gov/qbank/RespError.aspx) among establishment as well as interviewer—and self-administered population surveys using the key words ‘‘problematic terms" and ‘‘overly complex" returned 169 questions, of which 56 questions fulfilled the inclusion criteria (DeMaio, Landreth, and Hughes 2000; Hughes and Hunter 2003; Hunter 2005; Hunter-Childs et al. 2006; Kerwin 2003; Loomis and Rothgeb 1999; Maitland, Beatty, and Choi 2006; Miller and Beatty 2000; Miller and DeMaio 2006; Miller and Schoua-Glusberg 2006; Miller and Willson 2004; Rho 2009; Willson 2004; Willson 2006; Willson 2007; Wood, Forsyth, and Levin 2006). These two types of response errors were chosen because they were assumed to be closely related to the two variables involved in readability formulas (i.e., word length/frequency and sentence length).

In total, 71 question pairs entered the readability analysis. The text analysis software TextQuest 4.0 was used to calculate four separate readability scores for every question pair: the FRE, the FKG, the FOG, and the DC. In the FRE formula, higher scores indicate texts that are easier to understand; in the FKG, FOG, and DC formulas, higher scores indicate texts that are more difficult to understand.[4]

Results

The results of the readability analysis for all question pairs are displayed in the online appendix (which can be found at http://smr.sagepub.com/supple-mental/). The formulas' validity in identifying problematic survey questions was assessed by calculating success rates for each formula. These success rates compare the number of correct classifications (e.g., the formulas predict that readability is lower for the problematic question version than for the improved one) with the total number of questions in the analysis. Binomial tests were used to determine if the classification accuracies of the formulas were better than expected by chance alone (p values in parentheses).

Across all 71 question pairs, the success rates of the four formulas were 51 percent (p = 1.0) for FRE, 49 percent (p = 1.0) for FKG, 39 percent (p = .10) for FOG, and 38 percent (p = .06) for DC. These rates are disappointingly low and the binomial tests revealed that none of the formulas performed significantly better than expected by chance. On the contrary, the classification performance of the FKG, FOG, and DC formulas was even worse than expected for random guessing, which would result in a success rate of 50 percent. We then looked separately at the classification accuracy for the questions identified via the literature search and via Q-Bank. When we restricted the analysis to those question pairs reported in the existing literature (n = 15), the success rates dropped to 27 percent (p = .12) for each of the four formulas. Classification accuracy was considerably better for the Q-Bank questions (n = 56) with success rates of 57 percent (p = .35) for FRE, 55 percent (p = .50) for FKG, 43 percent (p = .35) for FOG, and 41 percent (p = .23) for DC. Finally, we looked separately at the Q-Bank questions tagged as including a ‘‘problematic term" (n = 34) and those tagged as being ‘‘overly complex" (n = 22). For questions including a problematic term, the success rates were 53 percent (p = .86) for FRE, 50 percent (p = 1.0) for FKG, 44 percent (p = .61) for FOG, and 44 percent (p = .61) for DC. For those tagged as ‘‘overly complex," the rates were 64 percent (p = .29) for FRE, 64 percent (p = .29) for FKG, 41 percent (p = .52) for FOG, and 36 percent (p = .29) for DC. All in all, the FRE and FKG formulas outperformed the FOG and the DC formulas while still showing unacceptably low classification accuracy on average. In addition, none of the formulas identified problematic questions significantly better than expected for random guessing.

Table 1 lists examples of question pairs, together with their respective readability scores. These examples illustrate several limitations of the formulas and may make their poor performance in identifying problematic survey questions more understandable. First, consider question pair Q18. Cognitive testing revealed that many respondents experienced difficulties in determining what the term ‘‘health organization" refers to. Thus, the recommendation was to define this vague term more clearly, for example, by rewording it to ‘‘government health organization" (Kerwin 2003). The readability formulas, however, disagree with this recommendation because, according to their underlying assumptions, adding a multisyllable word to the question should make it less readable, regardless of whether the additional word clarifies an unclear term or not. Similarly, according to the formulas, the additional information provided in the improved question version of Q23 reduces the readability of the question, whereas cognitive testing revealed that the question is, in fact, easier to answer if a reference period is provided (Maitland, Beatty, and Choi 2006). Question pair Q30 illustrates the formulas' neglect of the syntactic structure of a text. The negatively worded question version is certainly more difficult to comprehend than the positively worded version (cf. Akiyama, Brewer, and Shoben 1979; Clark and Chase 1972). However, the FRE and FKG formulas both favor the negatively worded question, because adding two very short words to the question reduces the number of syllables per word. Question pair Q56 demonstrates that the formulas sometimes produce absurd results when they are applied to short texts. For example, the FRE score of the improved question version is 112.1, even though the FRE formula was designed to rate texts on a 100-point scale. Similarly, the grade level of 0.1 assigned by the FKG formula certainly underestimates the reading level required to comprehend this question. Finally, consider question pair Q58. Given that both questions consist of the same number of letters, syllables, and words, all four formulas assign the same readability score to the problematic and the improved question version. Thus, they are oblivious to the fact that the term ‘‘expert" in the problematic version is less precise than the term ‘‘doctor" in the revised version. Cognitive testing revealed that the problematic version can be misinterpreted as

Table 1. Examples of Question Pairs With Respective Readability Scores.

Question Pairs	Flesch Reading Ease (FRE)	Flesch–Kincaid Grade Level (FKG)	Gunning Fog (FOG)	Dale–Chall (DC)	Source of Question Pairs
Q18 (Problematic, P): How much would you trust information about physical activity or nutrition from a health organization? Would you say a lot, some, a little, or not at all?	61.5	8.0	5.5	4.4	Kerwin (2003)
Q18 (Improved, I): How much would you trust information about physical activity or nutrition from a government health organization? Would you say a lot, some, a little, or not at all?	56.7	8.8	5.7	4.4	
Q23 (P): How many times have you had bronchitis?	91.0	2.3	2.9	4.0	Maitland, Beatty, and Choi (2006)
Q23 (I): How many times have you had bronchitis during the past five years?	88.9	3.8	4.8	4.3	
Q30 (P): Policies that do not safeguard the environment are bad	56.7	7.6	3.7*	4.2*	Bassili and Scott (1996)
Q30 (I): Policies that safeguard the environment are good	42.6	9.1	2.9*	4.1*	
Q56 (P): Have you ever had a Pap test to check for cervical cancer?	88.9*	3.8*	4.8*	4.3*	Willson (2007)
Q56 (I): Have you ever had a Pap smear or Pap test?	112.1*	0.1*	4.0*	4.2*	
Q58 (P): Has a doctor ever told you that some experts recommend the PSA test and others do not?	75.1	7.0	6.8	4.5	Wood, Forsyth, and Levin (2006)
Q58 (I): Has a doctor ever told you that some doctors recommend the PSA test and others do not?	75.1	7.0	6.8	4.5	

Note: PSA = prostrate-specific antigen; P = problematic question version; I = improved question version.

In the FRE formula, *higher* scores indicate higher readability; in the FKG, FOG, and DC formulas, *lower* scores indicate higher readability.

*Indicates that the readability formula correctly diagnoses the problematic and improved question versions according to the pretesting findings documented in Q-Bank and the question wording examples identified in the literature on questionnaire design.

Table 2. Agreement Between Readability Formulas in Diagnosing Problematic and Improved Question Versions.

Two-way Combinations of Formulas	Agreement Percentage	k (SE)
FRE + FKG	90.1	.80 (.07)
FRE + FOG	57.7	.16 (.11)
FRE + DC	59.2	.19 (.11)
FKG + FOG	67.6	.35 (.11)
FKG + DC	69.0	.38 (.11)
FOG + DC	95.8	.91 (.05)

Note: FRE = Flesch Reading Ease formula; FKG = Flesch-Kincaid Grade Level index; FOG = Gunning Fog index; DC = Dale-Chall formula.

being about ‘‘why the doctor ordered the test" instead of ‘‘whether the doctor mentioned two competing opinions about the test" (Wood, Forsyth, and Levin 2006).

Finally, we examined the degree of agreement between the four formulas in classifying the questions as problematic or improved. For all two-way combinations of the formulas, we coded whether the same or different question versions were identified as problematic and improved. Agreement statistics differed substantially between the various two-way combinations, with k values ranging from .16 to .91 (Table 2). Good agreement was found between the FRE and FKG formulas (k = .80) and between the FOG and DC formulas (k = .91). Very poor agreement was found between the FRE and FOG formulas (k = .16) and between the FRE and DC formulas (k = .19). These findings show that the choice of formula can have a strong impact on the results of a readability analysis. For example, in more than 40 percent of the question pairs tested in this study, the FRE and DC formulas came to completely opposite conclusions about which question version is better, in terms of readability. Hence, survey designers applying the FRE formula would favor one question version, while survey designers applying the DC formula would favor the other version. Again, this is a troublesome finding that raises the question of whether readability formulas are valid tools for assessing survey question difficulty.

Discussion

This study examined whether readability formulas are good predictors of survey question difficulty. Readability scores were calculated for a set of question pairs, each including a problematic and an improved question version, and the analyses revealed that the formulas often misdiagnosed the question versions by attesting the problematic version higher readability than the improved question version. Across the four formulas, overall classification accuracy was below 50 percent and none of the formulas identified the problematic questions better than expected by chance. Moreover, agreement between the formulas varied considerably, depending on the particular formulas that were compared. This finding suggests that the choice of formula can have a strong influence on the conclusions drawn from readability analyses, with different formulas favoring different question versions. All in all, our results indicate that the formulas' judgments are often misleading and that they are not very useful tools for assessing question difficulty.

The formulas' ability to correctly diagnose survey question difficulty seems to be limited by several factors. First, the formulas are not appropriate for estimating the readability of short texts. Usually, they require samples of at least 100 words for estimating readability and, given that survey questions are rarely of that length, the formulas are likely to produce inaccurate or even absurd results when applied to survey questions. Second, they only rely on two aspects of a text, namely word length or word frequency and sentence length, and neglect many other (and potentially more important) variables determining text comprehensibility. For example, problems with survey question comprehension can occur because of vague or ambiguous terms, complex syntactic structures, misleading or incorrect presuppositions, and unclear question purposes (cf. Graesser et al. 2006), all of which are variables that are likely to have a greater impact on question difficulty than those included in readability formulas. As mentioned previously, the formulas' underlying assumptions that word and sentence length are directly related to comprehension are overly simplistic and are not compatible with current linguistic analyses. Moreover, readability formulas rely completely on formal properties of a text and thereby neglect the semantic, pragmatic, psycho- and sociolinguistic aspects of language.

Some previous studies found significant (though relatively weak) correlations between readability scores and question difficulty indicators, such as response order effects, DK responses, and midpoint responses. However, it is important to note that correlation does not entail causation, and hence that the variables included in readability formulas do not necessarily cause comprehension difficulties. Again, there are many other variables contributing to question difficulty and our findings indicate that those incorporated in readability formulas are not among the most influential ones. Survey designers who ignore the difference between correlation and causation might be tempted to revise their questions by shortening words and sentences and by substituting seemingly ‘‘hard to understand" words by ‘‘more common" words. In doing so, they may ignore other and more detrimental question characteristics and may, in fact, make their questions more difficult to understand. With regard to the practical implications of the study, our findings suggest that readability formulas do not help to write better questions because their judgments are often misleading. Hence, we conclude that they are only of limited use in diagnosing question difficulty and recommend that they should not be used for testing and revising draft questions.

This study is limited by the fact that its analytical design rests on the assumption that the question pairs identified via the literature search and via Q-Bank actually consist of a problematic and an improved question version and that the improved version is indeed less difficult. Even though the recommendations and findings from both sources have high face validity, the possibility remains that the judgments of the textbook authors and the cognitive interviewers are faulty, so that they are not the perfect criteria for evaluating the performance of readability formulas. As far as we know, these question versions have not been tested against each other in splitballot experiments, so we cannot be sure that the improved questions produce more reliable and more valid responses than the problematic ones. Additional research is needed to examine the reliability and validity of the question versions, because these are the ultimate quality criteria. However, in the absence of this information, it seems reasonable to place more trust in the judgments of expert survey methodologists and on empirical findings from questionnaire pretests than on the suggestions made by readability formulas.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Even though most readability formulas were developed for rating English texts, some have also been modified or designed for use with other languages. This article is exclusively concerned with the formulas developed for English, and all studies reported here have applied the English versions of the formulas.

2. Consider (parliament, law), believe (you, law, unjust), believe (you, law, harmful), act (you alone), act (you together with others).
3. Act (you alone), act (you together with others).
4. The Flesch-Kincaid Grade Level index (FKG) formula is shown earlier. The formulas for computing the other three readability scores for English texts are:

$$FRE = 206.835 - \frac{words}{sentences} \times 1.015 - \frac{syllables}{words} \times 84.6.$$

$$FOG = 0.4 \times \left(\frac{words}{sentences} + \frac{3syllables}{words}\right).$$

$$DC = 0.1579 \times \frac{rarewords}{words} + 0.0496 \times \frac{words}{sentences} + 3.6365.$$

References

Akiyama, M. Michael, William F. Brewer, and Edward J. Shoben. 1979. ‘‘The Yes-No Question Answering System and Statement Verification." Journal of Verbal Memory and Verbal Behavior 18:365-80.

Anderson, Richard C. and Alice Davison. 1988‘‘Conceptual and Empirical Bases of Readability Formulas."Pp. 23-53 in Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered, edited by Alice Davison and Georgia M. Green. Hillsdale, NJ: Lawrence Erlbaum.

Ans, Bernard, Serge Carbonnel, and Sylviane Valdois. 1998. ‘‘A Connectionist Multiple-trace Memory Model for Polysyllabic Word Reading." Psychological Review 105:678-723.

Badgett, Barbara A. 2010. ‘‘Toward the Development of a Model to Estimate the Readability of Credentialing-examination Materials." UNLV Theses/Dissertations/Professional Papers/Capstones. Paper 185. University of Nevada, Las Vegas, NV.

Bailin, Alan and Ann Grafstein. 2001. ‘‘The Linguistic Assumptions Underlying Readability Formulae: A Critique." Language & Communication 21:285-301. Bassili, John N. and B. Stacey Scott. 1996. ‘‘Response Latency as a Signal to Question Problems in Survey Research." Public Opinion Quarterly 60:390-99.

Bruce, Bertram and Andee Rubin. 1988. ‘‘Readability Formulas: Matching Tool and Task."Pp. 5-22 in Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered, edited by Alice Davison and Georgia M. Green. Hillsdale, NJ: Lawrence Erlbaum.

Cantril, Hadley. 1944. Gauging Public Opinion. Princeton, NJ: Princeton University Press.

Chall, Jeanne S. and Edgar Dale. 1995. Readability Revisited: The New Dale-Chall Readability Formula. Cambridge, MA: Brookline Books.

Clark, Herbert H. and William G. Chase. 1972. ‘‘On the Process of Comparing Sentences against Pictures." Cognitive Psychology 3:472-517.

Coltheart, Max, Kathleen Rastle, Conrad Perry, Robyn Langdon, and Johannes Ziegler. 2001. ‘‘DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud." Psychological Review 108:204-56.

Converse, Jean M. 1976. ‘‘Predicting No Opinion in the Polls." Public Opinion Quarterly 40:515-30.

Converse, Jean M. and Howard Schuman. 1984. ‘‘The Manner of Inquiry: An Analysis of Survey Question Form across Organizations and Over Time." Pp. 283-316 in Surveying Subjective Phenomena, edited by Charles F. Turner and Elizabeth Martin. New York, NY: Russell Sage.

Dale, Edgar and Jeanne S. Chall. 1948. ‘‘A Formula for Predicting Readability." Educational Research Bulletin 27:11-20, 37-54.

Dale, Edgar and Joseph O'Rourke. 1981. The Living Word Vocabulary. Chicago, IL: World Book International.

Davison, Alice and Georgia M. Green. 1988. Linguistic Complexity and Text Comprehension: Readability Issues Reconsidered. Hillsdale, NJ: Lawrence Erlbaum.

DeMaio, Theresa, Ashley Landreth, and Kristen Hughes. 2000. Report of Cognitive Research on the School Crime Supplement for the 2001 National Crime Victimization Survey. Washington, DC: U.S. Bureau of the Census.

Ferrand, Ludovic. 2000. ‘‘Reading Aloud Polysyllabic Words and Nonwords: The Syllabic Length Effect Reexamined." Psychonomic Bulletin & Review 7: 142-48.

Ferrand, Ludovic and Boris New. 2003. ‘‘Syllabic Length Effects in Visual Word Recognition and Naming." Acta Psychologica 113:167-83.

Flesch, Rudolf F. 1948. ‘‘A New Readability Yardstick." Journal of Applied Psychology 32:221-33.

Flesch, Rudolf F. 1979. How to Write in Plain English: A Book for Lawyers and Consumers. New York, NY: Harper.

Fodor, Jerry A., Thomas G. Bever, and Merrill F. Garrett. 1974. The Psychology of Language: An Introduction to Psycholinguistics and Generative Grammar. New York, NY: McGraw-Hill.

Fowler, Floyd J. 1992. ‘‘How Unclear Terms Affect Survey Data." Public Opinion Quarterly 56:218-31.

Fowler, Floyd J. 1995. Improving Survey Questions. Thousand Oaks, CA: Sage.

Fowler, Floyd J. 2004. ‘‘The Case for More Split-sample Experiments in Developing Survey Instruments." Pp. 173-88 in Methods for Testing and Evaluating Survey Questionnaires, edited by Stanley Presser, Jennifer M. Rothgeb, Mick P. Couper, Judith L. Lessler, Elizabeth Martin, Jean Martin, and Eleanor Singer. New York, NY: John Wiley.

Fowler, Floyd J. and Carol Cosenza. 2008. ‘‘Writing Effective Questionnaires." Pp. 136-60 in International Handbook of Survey Methodology, edited by Edith de Leeuw, Joop Hox, and Don A. Dillman. New York, NY: Lawrence Erlbaum.

Gafke, Roger and David Leuthold. 1979. ‘‘The Effect on Voters of Misleading, Confusing, and Difficult Ballot Titles." Public Opinion Quarterly 43:394-401.

Graesser, Arthur C., Zhiqiang Cai, Max M. Louwerse, and Frances Daniel. 2006. ‘‘Question Understanding Aid (QUAID). A Web Facility That Tests Question Comprehensibility." Public Opinion Quarterly 70:3-22.

Groves, Robert M. 1989. Survey Errors and Survey Costs. New York, NY: John Wiley.

Gunning, Robert. 1952. The Technique of Clear Writing. New York, NY: McGraw-Hill.

Harmon, Mark D. 2001. ‘‘Poll Question Readability and ‘Don't Know' Replies." International Journal of Public Opinion Research 13:72-79.

Holbrook, Allyson L., Young Ik Cho, and Timothy Johnson. 2006. ‘‘The Impact of Question and Respondent Characteristics on Comprehension and Mapping Difficulties." Public Opinion Quarterly 70:565-95.

Holbrook, Allyson L., Jon A. Krosnick, David Moore, and Roger Tourangeau. 2007. ‘‘Response Order Effects in Dichotomous Categorical Questions Presented Orally. The Impact of Question and Respondent Attributes." Public Opinion Quarterly 71:325-48.

Homan, Susan, Margaret Hewitt, and Jean Linder. 1994. ‘‘The Development and Validation of a Formula for Measuring Single-sentence Test Item Readability." Journal of Educational Measurement 31:349-58.

Hughes, Kristen and Jennifer Hunter. 2003. Results & Recommendations from the Cognitive Pretesting of the 2003 SIPP Recipiency History Module. Washington, DC: U.S. Bureau of the Census.

Hunt, Kellogg W. 1965. Grammatical Structures Written at Three Grade Levels (NCTE Research Report No. 3). Urbana, IL: National Council of Teachers of English.

Hunter, Jennifer. 2005. Cognitive Test of the 2006 NRFU. Washington, DC: U.S. Bureau of the Census.

Hunter-Childs, Jennifer, Eleanor Gerber, George Carter, and Jennifer Beck. 2006. Cognitive Test of the 2006 NRFU: Round 2. Washington, DC: U.S. Bureau of the Census.

Inhoff, Albrecht W. and Keith Rayner. 1986. ‘‘Parafoveal Word Processing during Eye Fixations in Reading: Effects of Word Frequency." Perception & Psychophysics 40:431-39.

Jared, Debra and Mark S. Seidenberg. 1990. ‘‘Naming Multisyllabic Words." Journal of Experimental Psychology: Human Perception & Performance 16:92-105.

Juphard, Alexandra, Serge Carbonnel, and Sylviane Valdois. 2004. ‘‘Length Effect in Reading and Lexical Decision: Evidence from Skilled Readers and a Developmental Dyslexic Participant." Brain and Cognition 55:332-40.

Kerwin, Jeffrey. 2003. HINTS Mod. #5 Interviews: Round 2 Summary of Findings and Recommendations. Rockville, MD: Westat.

Kimball, David C. and Martha Kropf. 2005. ''Ballot Design and Unrecorded Votes on Paper-based Ballots." Public Opinion Quarterly 69:508-29.

Klare, George R. 2002. ''Readability." Pp. 681-744 in Handbook of Reading Research, edited by P. David Pearson. Mahwah, NJ: Lawrence Erlbaum.

Lenzner, Timo. 2011. A Psycholinguistic Look at Survey Question Design and Response Quality. Mannheim, Germany: University of Mannheim, MADOC.

Lenzner, Timo, Lars Kaczmirek, and Mirta Galesic. 2011. ''Seeing through the Eyes of the Respondent: An Eye-tracking Study on Survey Question Comprehension." International Journal of Public Opinion Research 23:361-73.

Loomis, Laura and Jennifer Rothgeb. 1999. Final Report on Cognitive Interview Research Results and Revisions to the Welfare Reform Benefits Questions for the March 2000 Income Supplement to the CPS. Washington, DC: U.S. Bureau of the Census.

Maitland, Aaron, Paul Beatty, and Colleen Choi. 2006. Adult Core Respiratory Disease Questionnaire Including Occupational Components Cognitive Testing Results. Washington, DC: National Center for Health Statistics.

Miller, Kristen and Paul Beatty. 2000. Cognitive Testing of the Strengths and Difficulties Questionnaire (SDQ). Hyattsville, MD: National Center for Health Statistics.

Miller, Kristen and Theresa J. DeMaio. 2006. Report of Cognitive Research on Proposed American Community Survey Disability Questions. Hyattsville, MD: National Center for Health Statistics.

Miller, Kristen and Alisu Schoua-Glusberg. 2006. Cognitive Testing Results of Oral Health Questions. Hyattsville, MD: National Center for Health Statistics.

Miller, Kristen and Stephanie Willson. 2004. Cognitive Interviewing Evaluation of the Survey on Emerging Traumatic Events: Surveillance (Tier I) Questionnaire. Hyattsville, MD: National Center for Health Statistics.

Nagy, William E. and Richard C. Anderson. 1984. ''How Many Words Are There in Printed School English?" Reading Research Quarterly 19:304-30.

New, Boris, Ludovic Ferrand, Christophe Pallier, and Marc Brysbaert. 2006. ''Reexamining the Word Length Effect in Visual Word Recognition: New Evidence from the English Lexicon Project." Psychological Bulletin & Review 13:45-52.

Oakland, Thomas and Holly B. Lane. 2004. ''Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests." International Journal of Testing 4:239-52.

Payne, Stanley L. 1949. ''Case Study in Question Complexity." Public Opinion Quarterly 13:653-58.

Payne, Stanley L. 1951. The Art of Asking Questions. Princeton, NJ: Princeton University Press.

Rho, Christine. 2009. Cognitive Tests of Veterans Supplement Questions. Washington, DC: U.S. Bureau of Labor Statistics.

Spache, George. 1953. ‘‘A New Readability Formula for Primary-grade Reading Materials." Elementary School Journal 53:410-13.

Templeton, Shane, Carolynt T. Cain, and James O. Miller. 1981. ‘‘Reconceptualizing Readability: The Relationship between Surface and Underlying Structure Analyses in Predicting the Difficulty of Basal Reader Stories." Journal of Educational Research 74:382-87.

Terris, Fay. 1949. ‘‘Are Poll Questions Too Difficult?" Public Opinion Quarterly 13: 314-19.

Velez, Pauline and Steven D. Ashworth. 2007. ‘‘The Impact of Item Readability on the Endorsement of the Midpoint Response in Surveys." Survey Research Methods 1:69-74.

Weekes, Brendan S. 1997. ‘‘Differential Effects of Number of Letters on Word and Nonword Naming Latency." Quarterly Journal of Experimental Psychology 50: 439-56.

Wheeler, George and Thomas F. Sherman. 1983. ‘‘Readability Formulas Revisited." Science and Children 20:38-40.

Williams, Rihana S. and Robin K. Morris. 2004. ‘‘Eye Movements, Word Familiarity, and Vocabulary Acquisition." European Journal of Cognitive Psychology 16: 312-39.

Willson, Stephanie. 2004. Cognitive Interviewing Evaluation of the Survey on Emerging Traumatic Events: Post-event (Tier II) Questionnaire. Hyattsville, MD: National Center for Health Statistics.

Willson, Stephanie. 2006. Cognitive Interviewing Evaluation of the 2007 Complementary and Alternative Medicine Module for the National Health Interview Survey. Hyattsville, MD: National Center for Health Statistics.

Willson, Stephanie. 2007. Cognitive Interview Evaluation of the 2008 National Health Interview Survey Supplement on Immunizations & Cancer Screenings: Results of Interviews Conducted June-August, 2007. Hyattsville, MD: National Center for Health Statistics.

Wood, Elizabeth, Barbara Forsyth, and Kerry Levin. 2006. PSA Test Decision Making: Results from Cognitive Testing: Round 2. Rockville, MD: Westat.

Author Biography

Timo Lenzner is a senior researcher at the survey pretesting laboratory at GESIS - Leibniz Institute for the Social Sciences, Germany. His research focuses on questionnaire design and evaluation, Web surveys, and usability.