2022

Comparing Readability Measures and Computer-assisted Question Evaluation Tools for Self-administered Survey Questions

Rachel Stenger
RTI International, Research Triangle Park, NC, rstenger@rti.org

Kristen Olson
University of Nebraska-Lincoln, kolson5@unl.edu

Jolene Smyth
University of Nebraska-Lincoln, jsmyth2@unl.edu

Comparing Readability Measures and Computer-assisted Question Evaluation Tools for Self-administered Survey Questions

Rachel Stenger,[1] Kristen Olson,[2] and Jolene D. Smyth [2]

1 RTI International, Research Triangle Park, NC, USA

2 University of Nebraska-Lincoln, Lincoln, NE, USA

Corresponding author — Rachel Stenger, RTI International,
3040 E Cornwallis Rd, Research Triangle Park, NC 27709-2194, USA;
email: rstenger@rti.org

ORCID

Rachel Stenger https://orcid.org/0000-0003-0643-198X

Kristen Olson https://orcid.org/0000-0002-8004-0226

Abstract

Questionnaire designers use readability measures to ensure that questions can be understood by the target population. The most common measure is the Flesch-Kincaid Grade level, but other formulas exist. This article compares six different readability measures across 150 questions in a self-administered questionnaire, finding notable variation in calculated readability across measures. Some question formats, including those that are part of a battery, require important decisions that have large effects on the estimated readability of survey items. Other question evaluation tools, such as the Question Understanding Aid (QUAID) and the Survey Quality Predictor (SQP), may identify similar problems in questions, making readability measures less useful. We find little overlap between QUAID, SQP, and the readability measures, and little differentiation in the tools' prediction of item nonresponse rates. Questionnaire designers are encouraged to use multiple question evaluation tools and develop readability measures specifically for survey questions.

Published in *Field Methods* 2022

DOI: 10.1177/1525822X221124469

Introduction

When writing survey questions and consent forms for the general population, a common recommendation is to write questions at a level that can be read and understood by the sampled population, typically around an eighth-grade reading level (e.g. Dillman et al. 2014; Paasche- Orlow et al. 2003; Payne 1951). This guidance seems simple in principle, but in practice, calculating readability requires several decisions. One decision is which readability measure to use. Although the Flesch-Kincaid Grade Level is readily available in Microsoft Word, multiple other readability measures exist. Additionally, researchers must decide what parts of the question to include when calculating readability; this decision is straightforward for some types of questions, but not others (e.g., batteries).

Readability measures are one form of ex ante computer-assisted methods for evaluating survey questions (Caporaso and Presser 2021). Although different readability measures have been shown to vary for the same passages of text (Zhou et al. 2017) and for some survey questions (Lenzner 2014), how different readability measures compare to other computer-assisted question evaluation tools such as the Question Understanding Aid (QUAID, Graesser et al. 2006) and the Survey Quality Predictor (SQP, Saris and Gallhofer 2007) has received surprisingly little attention. Understanding whether different computer-assisted tools identify similar questions as problematic is important so that survey designers can effectively target limited resources. Additionally, whether different readability measures are associated with indicators of response difficulties, such as item nonresponse, is unclear. Thus, in this article, we address the following research questions:

RQ1: Does the readability grade level and whether a survey question meets the eighth-grade level benchmark depend on the measure used to calculate readability?

RQ2: Are difficult-to-read questions also flagged as problematic by QUAID and/or associated with ratings of lower quality in SQP?

RQ3: Is there an association between item nonresponse rates and indicators of question problems measured by readability levels, QUAID, and SQP?

Background

The readability level of a text passage indicates whether a person with a certain level of schooling can read the passage. Readability measures were developed to evaluate average readability of passages containing 100 words or more (Bruce and Rubin 1988) but are also used to evaluate survey questions. Caporaso and Presser (2021) reported about 40% of survey organizations use measures of reading difficulty when pretesting questionnaires.

A rule of thumb is that general population survey questions should be written at an eighth-grade reading level or lower (e.g., Payne 1951), originating from Flesch's (1948) recommendation that a "standard" reading level was eighth to ninth grade. Researchers have used readability statistics to evaluate whether survey questions meet the eighth grade (or other) reading level benchmark (e.g., Betschart et al. 2018; Paz et al. 2009), finding that anywhere from 13% (Betschart et al. 2018) to 100% (Paz et al. 2009) of evaluated survey questions fail to do so. These studies generally use the Flesch-Kincaid Grade Level, but several other measures are available.

Common readability measures use the number of characters, words, syllables, and sentences, and the presence of complex words to calculate a passage's reading grade level. The most common measure is the **Flesch-Kincaid Grade Level** (FKG)

$$FKG = 206.835 - 1.015\left(\frac{\#\ words}{\#\ sentences}\right) - 84.6\left(\frac{\#\ syllables}{\#\ words}\right)$$

(Flesch 1948; Kincaid et al. 1975)

Because counting the number of syllables can be cumbersome, readability formulas beyond FKG ease the calculation by eliminating the number of syllables from the formula **Coleman-Liau Index:**

$$CLI = 0.0588\left(\frac{\#\ letters\ or\ numbers}{\#\ words}\right) - 0.296\left(\frac{\#\ sentences}{\#\ words}\right) - 15.8$$

(Coleman and Liau 1975)

and by using parts of language that are easier to enumerate (e.g., characters including punctuation) **Automated Reading Index:**

$$ARI = 4.71\left(\frac{\#\ characters}{\#\ words}\right) + 0.5\left(\frac{\#\ words}{\#\ sentences}\right) - 21.43$$

(Senter and Smith 1967)

Other measures expanded on the formulas by counting the number of polysyllabic (3 or more syllables) words; **Gunning-Fog Index:**

$$FOG = 0.4\left[\left(\frac{\#\ words}{\#\ sentences}\right) + 100\left(\frac{\#\ polysyllabic\ words}{\#\ words}\right)\right]$$

(Gunning 1952)

or **Simple Measure of Gobbledygook** [SMOG]:

$$SMOG = 1.0430\sqrt{\#\ polysyllablic\ words * \left(\frac{30}{\#\ sentences}\right)} + 3.1291$$

(McLaughlin 1969)

For example, a question such as "*Some cellphones are called 'smartphones' because of certain features they have. Is your cellphone a smartphone or not, or are you not sure?*" contains longer words with many characters (e.g., "cellphones," "smartphone," "features"), but these words are not polysyllabic. The readability grade levels estimated by FKG (4.3), FOG (4.6), and SMOG (3.1) are lower than those estimated by CLI (10.3) and ARI (7.3). Further, some readability measures (FKG, CLI, ARI) can produce negative readability grade levels because they involve subtraction.

Decisions for Calculating Readability for Battery Items

A key decision that researchers make when calculating question readability is what constitutes the question. Distinguishing between the parts of a question (e.g., introduction, question stem [statement of the request eliciting a response], response options, instructions, definitions) is simple for some types of questions and thus of little consequence. For other question types—notably, battery items—identifying what words constitute the question can be considerably more consequential. Battery items contain a common question stem and set of response options that apply to multiple subitems (e.g., Question: "In the past 12 months, how many times did you do each of the following?" Subitem: "You participated in a service organization"). When calculating readability for these items, whether researchers should include the question stem with each subitem or with only the first subitem in the battery is unclear. Because decisions on what constitutes the relevant part of the question in battery items have implications for the resulting readability grade level, we examine battery and non-battery items separately, and evaluate different approaches of defining the question text for battery items.

Other Ex Ante Question Evaluation Methods

It is well established that different question evaluation methods identify different types of problems, called the "complementary methods hypothesis" (Maitland and Presser 2018; Tourangeau et al. 2021). Questionnaire designers are encouraged to use multiple question evaluation methods (including those without additional data collection, called ex ante approaches), guided by the goals of evaluation and a project's available time and resources (Maitland and Presser 2018; Tourangeau et al. 2020). These question evaluation approaches include time- and resource-intensive methods like expert reviews, cognitive interviews, and behavior coding (Tourangeau et al. 2020) and faster, less expensive computer-assisted evaluation tools like readability measures, QUAID (Graesser et al. 2006), and SQP (Saris and Gallhofer 2007). Whether different readability measures themselves form "complementary methods" or yield conclusions about questions that differ from other computer-assisted tools requires further evaluation.

Although computer-assisted tools are less costly than cognitive interviews or behavior coding, calculating each measure still requires researcher time. Furthermore, computer-assisted tools fall into the same "test environment" (Maitland and Presser 2018) for questions, not directly observing the response process. If these computer-assisted tools identify the same questions as problematic, then researcher time could be spent on one of them. Thus, it is useful for questionnaire designers to understand if different readability measures, QUAID, and SQP are indeed complementary or duplicative methods.

QUAID is an online tool that identifies five comprehension problems in survey questions (Graesser et al. 2006). QUAID draws on existing lexicons (i.e., lists of words and parts of speech accompanied by linguistic measures; Coltheart 1981) to identify individual problematic words in the question. QUAID flags words used infrequently as *unfamiliar technical terms*; common adjectives and adverbs (e.g., many, few) as *vague or imprecise predicate or relative terms*; and words with high levels of abstraction (e.g., "vehicle" vs. "Honda"), higher polysemy values (the number of different meanings the word can have), and lower concreteness values as *vague or ambiguous noun phrases*. Because these flags are based on single words rather than sentence structure, we expect little agreement between the questions identified as having these three QUAID problems and failing to meet the eighth-grade readability benchmark.

QUAID also assigns flags based on other elements of sentence structure. QUAID uses measures of the number of words (i.e., number of words before the main verb of the main sentence clause; number of noun modifiers) to identify *complex syntax*. Likewise, QUAID flags a survey question as having a problem with *working memory overload* if the question exceeds an undisclosed threshold of conjunctions (e.g., "if," "or," "and"), words that may lengthen survey questions. Thus, we expect more agreement between the questions identified as having these QUAID problems and failing to meet the eighth-grade readability benchmark.

Like QUAID, SQP is a web-based tool for evaluating survey questions (Saris and Gollhofer 2007). SQP provides estimates of reliability, validity, and quality (quality=reliability*validity) from a meta-analysis of multitrait–multimethod experiments. SQP users enter codes for many question characteristics, including linguistic features of the question (e.g., number of words, abstract nouns) and information about response options (e.g., number of response options, unipolar vs. bipolar scale). SQP incorporates more information than the readability measures, although some of the inputs are similar (e.g., number of words, number of sentences). Nevertheless, we expect a negative association between question reading levels (higher = more difficult to read) and SQP quality scores (higher = better quality).

Evaluation Methods and Data Quality

Researchers who evaluate their survey questions often anticipate that identified problems will result in field difficulties, such as item nonresponse. Previous work examining computer-assisted tools inconsistently predict item missingness or other data quality indicators (e.g., Dykema et al. 2020; Maitland and Presser 2018; Tourangeau, et al. 2021). If these tools successfully detect comprehension problems, we expect questions that are more difficult to comprehend have higher readability grade levels, more likely to have QUAID-identified problems, and lower SQP quality and thus will have higher item nonresponse rates.

Data

Our data are question-level measures for each question in the Community Values and Opinions in Nebraska Survey (CVONS), an English-language mail and web survey conducted by the Bureau of Sociological Research in spring 2017 (*N* = 2,705; AAPOR RR2 = 28.1%). CVONS contained 60 nonbattery questions and nine battery questions with 91 subitems about respondents' community, crime victimization, and demographics, among other topics (see Appendix A). The questions represent a mix of items taken from existing national surveys (e.g., General Social Survey; American National Election Survey) and items written by two of the authors modeled after commonly asked survey questions. Sampled households in CVONS were randomly assigned to one of two questionnaires containing different question design and format experiments. We evaluated each version separately if the question stem wording differed across versions.

We used readable.com to calculate the five readability measures for each question. Given the prevalence of FKG in Microsoft Word, we also calculate FKG using MS Word 2016 (Microsoft Word caps the minimum FKG readability value at zero; readable.com permits negative readability values). For the 60 non-battery items, this yields 360 readability calculations. For the nine battery items, we calculate each of the six readability measures two ways: (1) with the question stem only included with the first subitem (the stem-with-first approach) and (2) with the question stem included with every subitem (the stem-with-all approach), yielding

546 readability measures. In all cases, our readability calculations focused on the question stem; response options were included only when they were required to finish the question (e.g., question 50, Appendix A). To evaluate QUAID-identified problems, the question stem was entered into http://quaid.cohmetrix.com/ . We used the same question text in QUAID as used for the readability measures, excluding the response options unless needed to finish the question. All QUAID codes thus identified QUAID-identified problems in the question stem only (see Appendix B). For each question stem, an indicator variable was generated for each of the five QUAID problems (0 = problem not present; 1 = problem present). Sizeable percentages of questions were flagged by QUAID for each of unfamiliar technical terms (32–85%; Appendix E), vague or imprecise relative terms (16–60%), and vague or ambiguous noun-phrases (19–41%). No questions were flagged for complex syntax and very few were flagged for working memory overload (0–11.7%).

One author coded and entered each survey question into SQP (at http://sqp.upf.edu/). Because SQP requires information about response options, questions were entered separately for each questionnaire version, even if the question stem was the same (e.g., response format of numeric open-end in version 1 vs. ordinal scale in version 2). The SQP coding instructions specify that the question stem in a battery be included with the first subitem but not with subsequent subitems (the stem-with-first approach); we followed this instruction and thus do not have stem-with-all measurements from SQP. The two open-ended narrative questions cannot be assessed with SQP. Therefore, SQP measures were obtained for 62 non-battery questions and 98 battery questions. Following previous work (Maitland and Presser 2018; Tourangeau et al. 2021), we focus on the SQP quality score (non-battery: M= 0.548, SD = 0.055; battery: M = 0.546, SD = 0.039).

Last, we calculate question-level item nonresponse rates from the survey data (*N* = 2,705, non-battery: M = 5.57%, SD = 5.25; battery: M = 5.50%, SD = 3.34). Check-all-that-apply questions and questions that followed a filter question in a skip pattern were excluded, for an analytic data set of 47 nonbattery and 91 battery items. Some battery items experimentally varied use of check-all-that-apply versus forced-choice (yes/no grid) formats. Item nonresponse rates were only calculated for the forced-choice version of the questions (*N* = 1,307 respondents; three batteries).

Analysis Plan

We first compare mean readability grade levels across the five readability measures using dependent *t*-tests (RQ1). Multiple comparisons are accounted for within each battery and non-battery group using a Bonferroni correction (p<0.0033 indicates significant differences). We classify each question according to whether (1) or not (0) it meets the eighth-grade benchmark using each readability measure and compare the proportion of questions that meet the benchmark across the readability measures with dependent *t*-tests. Because the FKG level in Microsoft Word (FKG-Word) is a commonly used readability tool, we compare this tool against each of the other measures using kappa statistics (see Appendix C for other agreement measures). Many rules of thumb for interpreting kappa exist (Landis and Koch 1977). We use kappa values above 0.75 to indicating excellent agreement, 0.60–0.74 to indicate good agreement, 0.40–0.59 as fair agreement, and below 0.40 to indicate poor agreement.

We then examine the rate of agreement that questions are problematic or not between the readability measures and QUAID, as well as the association between the readability measures, QUAID, and SQP (RQ2). Questions that fail to meet the eighth-grade reading benchmark and those that receive a QUAID problem flag are considered problematic. For this analysis, we examine agreement using kappa (Appendix D contains other agreement measures); we exclude working memory capacity for the battery questions because it did not occur for these items. Complex syntax did not occur in any survey question. Rather than producing a question problem flag, SQP predicts “quality scores.” We estimate the Pearson correlation between the continuous readability grade levels, QUAID indicators, and the SQP quality scores.

Finally, we examine whether the various question evaluation methods predict data quality as measured by item nonresponse (RQ3).We estimate the Pearson correlation between the continuous six readability measures, QUAID, and SQP scores with the question-level item nonresponse rates.

All analyses are stratified by non-battery and battery items, and separately examine the stem-with-all and stem-with-first approaches for the battery items.

Results

Comparing Readability Measures

We start by examining the readability measures themselves. For the non-battery questions, FKG-Word, FKG-Readable, and ARI produce statistically similar mean readability estimates across the non-battery items (Table 1, top panel, $p > 0.0033$). Similarly, the percentage meeting the eighth-grade benchmark did not significantly differ between FKG-Word, FKG-Readable and ARI ($p > 0.0033$); they were all between 50% and 58%. CLI, FOG, and SMOG yield significantly higher average readability ($p < 0.0033$) than FKG and ARI, and significantly lower percentages that meet the eighth-grade benchmark ($p = 0.0033$), which ranged from 17% to 37% (but did not significantly differ among each other). These findings are mirrored when looking at kappa, which found highest agreement of FKG-Word with FKG-Readable (excellent agreement); all others showed lower agreement, including ARI (fair), FOG (fair), CLI (fair), and SMOG (poor). Thus, researchers using different readability measures risk drawing very different conclusions about reading levels for their non-battery questions. A researcher using FKG-Word would conclude that many of the questions (58.3%) meet the benchmark while a researcher using SMOG would conclude that very few (16.7%) do.

For battery questions, we start by examining reading levels for the stem-with-first item approach. In this calculation, the average readability level is similar across the different readability measures; only FKG-Readable (M = 8.94) and CLI (M = 6.28) significantly differ from each other (Table 1, middle panel, $p<0.0033$). Despite these similarities in the mean reading levels, there is substantial variation in the estimated reading levels across items, with standard deviations ranging from sixth- to ninth-grade levels. Additionally, all of the measures except for SMOG have similar estimates of the percent of questions meeting the eighth-grade reading level benchmark (51.6%–62.6%). The overall proportion masks variation in which questions are identified as meeting the eighth-grade benchmark; agreement with FKG-Word is highest for FKG-Readable (excellent agreement), and lower for ARI (good), FOG (good), SMOG (fair), and CLI (poor).

In contrast, when we examine the stem-with-all approach, most of the measures differ from one another on average. ARI yields the lowest

Table 1. Mean and Standard Deviation of Reading Level, Proportion of Questions that Meet the Eighth-grade Benchmark, and Kappa Between Flesch-Kincaid in Microsoft Word and Each of the Readability Measures for Non-battery and Battery Items.

Readability measure	*Mean*	*SD*	*Proportion Meeting 8th Grade Benchmark*	*Kappa versus FKG-Word*
Non-battery items (*n* = 60)				
FKG-Word$_a$	8.20$_{c,d,e}$	3.92	0.583$_{c,d,e}$	
FKG-Readable$_b$	8.21$_{c,d,e}$	4.12	0.567$_{c,d,e}$	0.9659****
FOG$_c$	10.77$_{a,b,d,f}$	4.50	0.300$_{a,b,f}$	0.4063***
CLI$_d$	9.37$_{a,b,c,e,f}$	4.22	0.367$_{a,b,f}$	0.4574****
SMOG$_e$	10.57$_{a,b,d,f}$	3.77	0.167$_{a,b,f}$	0.2500**
ARI$_f$	7.84$_{c,d,e}$	4.33	0.500$_{c,d,e}$	0.5667****
Battery items – Stem-with-first (*n* = 91)				
FKG-Word$_a$	8.68	6.16	0.626$_e$	
FKG-Readable$_b$	8.94$_d$	7.31	0.615$_e$	0.9767****
FOG$_c$	9.38	9.63	0.516$_e$	0.4238****
CLI$_d$	6.28$_b$	9.37	0.593$_e$	0.2849**
SMOG$_e$	7.61	3.39	0.330$_{a,b,c,d,f}$	0.4132****
ARI$_f$	8.67	6.99	0.571$_e$	0.5668****
Battery items – Stem-with-all (*n* = 91)				
FKG-Word$_a$	6.67$_{c,d,e,f}$	1.46	0.923$_{c,d,e}$	
FKG-Readable$_b$	6.86$_{c,d,e,f}$	1.56	0.846$_{c,d,e}$	0.6286****
FOG$_c$	10.76$_{a,b,d,f}$	2.12	0.121$_{a,b,d,e,f}$	0.0227
CLI$_d$	8.53$_{a,b,c,e,f}$	2.23	0.560$_{a,b,c,e,f}$	0.1921***
SMOG$_e$	10.76$_{a,b,d,f}$	1.29	0.011$_{a,b,c,d,f}$	0.0019
ARI$_f$	5.92$_{a,b,c,d,e}$	1.90	0.912$_{c,d,e}$	0.6369****

Readability was calculated using readable.com unless otherwise indicated. Differences in proportions were calculated using dependent t-tests. Subscripts denote significant differences at the $p < 0.0033$ level, using the Bonferroni correction.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$; **** $p < 0.0001$

mean readability grade level, followed by the two FKG measures, CLI, and FOG and SMOG as the highest grade levels (Table 1, bottom panel). Furthermore, variation across the items within each measure is substantially reduced compared to the stem-with-first approach. The proportion of items that meet the eighth-grade benchmark also vary dramatically, with the two FKG measures and ARI producing the highest rates (all >85% and not significantly different from each other), FOG and SMOG producing the lowest rates (12.1% and 1.1% respectively), and CLI in the middle (56%). Likewise, the agreement between FKG-Word and the other measures are starkly different with FKG and ARI producing kappa values around 0.63 (good agreement), and SMOG, FOG, and CLI producing kappa values below 0.20 (poor agreement).

In sum, readability evaluations depend heavily on how the question stem and subitems for the battery items are treated. In the stem-with-first approach, conclusions about the average readability level across the items are similar across measures, but there is substantial variation across items. In contrast, the stem-with-all approach produces considerably different conclusions across the readability measures but washes out within-battery differences in reading levels.

Comparing Readability Measures, QUAID, and SQP

Now we examine agreement between FKG-Word and QUAID problem flags. As expected (Figure 1), agreement between FKG-Word and QUAID is weak, suggesting that readability measures and QUAID are complementary methods. Most of the kappa values hover around zero or are negative, all falling into the poor agreement range (<0.4) between the readability measures and QUAID. Overall, the readability measures and the QUAID flags come to different conclusions about whether questions are problematic, suggesting that they are detecting different types of problems. For battery items, this is true regardless of how the question stem was treated in the readability calculations. Thus, readability assessments and QUAID are not substitutes for one another in question evaluation.

We now examine the correlation between the readability measures, QUAID flags, and the SQP quality scores (Appendix F). For the nonbattery items, none of the readability-SQP associations differ from zero (corr (readability, SQP) ranges from –0.193 to –0.00, $p > 0.14$ for all correlations), although they are in the right direction. For the QUAID and SQP associations, questions with vague or imprecise relative terms have lower quality scores (corr (vague/imprecise, SQP) = –0.286, $p < 0.05$); the other QUAID measurements are not associated with SQP quality scores ($p > 0.49$). Thus, there is no clear association between these questionnaire evaluation methods for non-battery items.

A different pattern exists for battery items. Here, the correlations between FKG-Word ($\text{corr}_{\text{stem-with-first}}$ (FKG-Word, SQP) = 0.219, $p = 0.03$) and ARI ($\text{corr}_{\text{stem-with-first}}$ (ARI, SQP) = 0.375, $p = 0.0001$) and the SQP quality scores are positive and significant in the stem-with-first approach

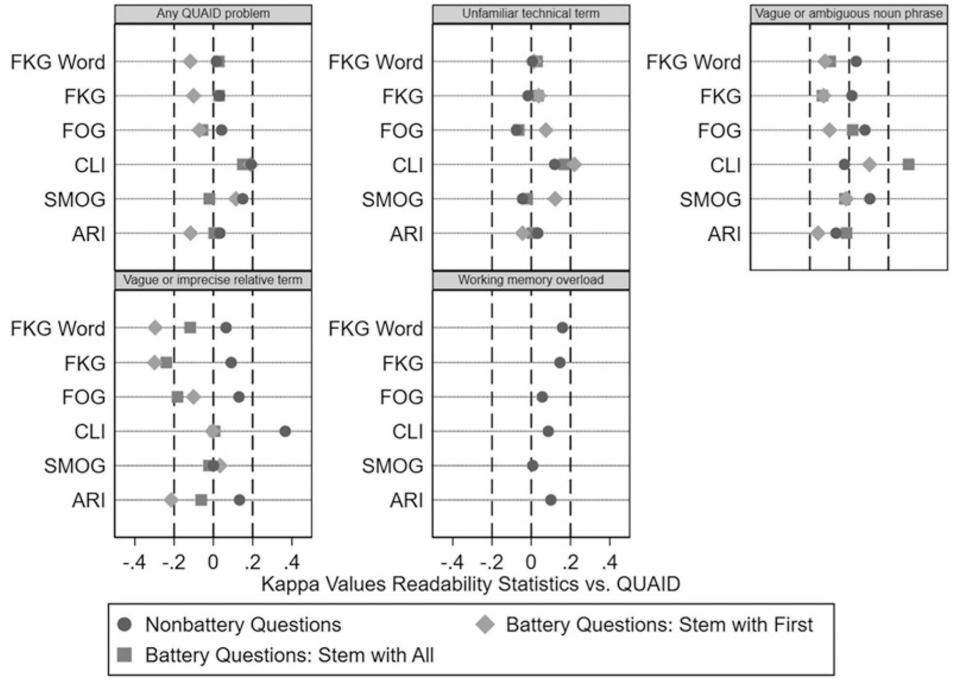

Figure 1. Kappa values for QUAID identified problems and indicators of question meeting the eighth-grade benchmark, battery and non-battery items.

(which matches the SQP instructions), counter our hypothesis. In contrast, SQP correlations with CLI and SMOG are in the expected negative direction (corr$_{\text{stem-with-first}}$ (CLI, SQP) = –0.23, p = 0.02; corr$_{\text{stem-with-first}}$ (SMOG, SQP) = –0.27. p = 0.007) and significant in the stem-with-first approach, but positive and significant (corr$_{\text{stem-with-all}}$ (CLI, SQP) = 0.37, p = 0.0002; corr$_{\text{stem-with-all}}$ (SMOG, SQP) = 0.23, p = 0.02) when the readability calculations use the stem-with-all approach. In the stem-with-first approach, the QUAID flags have the expected negative significant association with SQP for imprecise relative terms (corr$_{\text{stem-with-first}}$ (imprecise relative terms, SQP) = –0.33, p = 0.0008) and vague noun-phrases (corr$_{\text{stem-with-first}}$ (vague noun phrases, SQP = –0.28, p = 0.005), but not for unfamiliar technical terms (p = 0.27). These SQP associations with QUAID are attenuated in the stem-with-all approach. These results suggest that the stem should be included only with the first subitem in the battery.

Comparing Readability Measures and Item Nonresponse Rates

We now examine the correlation between the readability measures and item nonresponse rates (Appendix F). For non-battery items, although the associations are generally in the expected direction, there is no statistically significant association ($p > 0.16$) between item nonresponse and *any* of the computer-assisted measures, a pattern that also generally arises for battery items (with only two exceptions across the 20 correlations). Thus, none of the computer-assisted methods consistently predict item nonresponse rates, similar to previous studies (e.g., Maitland and Presser 2018).

Discussion

Survey designers need cost-efficient methods to evaluate survey questions for potential problems prior to data collection. Readability measures are one such tool that have received insufficient empirical attention. Across our analyses, FKG and ARI yield lower reading levels and identify more questions as meeting the eighth-grade benchmark, FOG and SMOG consistently yield higher reading levels and identify fewer questions as meeting the eighth-grade benchmark, and CLI landed in the middle. Thus, we recommend that survey designers use at least two readability measures when evaluating the reading level of their questionnaires—one from FKG (via any calculation method) and ARI and one from FOG, CLI, or SMOG. These two groups of readability measures will allow a survey designer to evaluate the range of the risk of questions not being able to be read by their target population. Additionally, the lack of agreement across the readability measures evaluated here may suggest that these measures are problematic. Future development of readability measures specifically for evaluating survey questions is warranted.

Different question evaluation tools often identify different types of problems; we replicate this finding here. None of the readability measures for these survey items consistently aligned with the QUAID-identified problems or SQP quality measure. Most of the QUAID problems are based on single words whereas the readability measures are based on multiple words (Appendix E). Similarly, readability and SQP are not interchangeable question evaluation tools; SQP draws on a wide range of question characteristics beyond sentence structure. That these tools do not flag the same problems should not be seen as troublesome for any indicator; rather, they are tapping into different aspects of the same question and are complementary methods. Unfortunately, we were unable to assess agreement between readability and the QUAID-identified problems that draw more strongly on elements of sentence structure (working memory overload and complex syntax) because these occurred too infrequently in our survey. Future research should do so.

Our results show that none of the computer-assisted tools consistently predict question-level item nonresponse rates. Other question characteristics may be associated with both the evaluations from these tools and item nonresponse rates; future research could examine item nonresponse while accounting for these characteristics. Additionally, item nonresponse is only one field outcome; other outcomes such as question reliability, concurrent and predictive validity, or other indicators of respondent problems (e.g., answer changes) may be more closely associated with these measures. Unfortunately, our data do not support direct evaluations of other field outcomes (Appendix G contains page-level response time analyses). More research is needed to know for what types of data quality and measurement problems each of these tools is most efficacious. Future research also should replicate these evaluations using interviewer or respondent behaviors in interviewer-administered surveys.

Human decisions in question evaluation can affect inferences from computer-assisted tools. Our results suggest that differences in calculating readability may partially explain heterogeneity in the efficacy of reading levels for identifying “problematic” interview behaviors across studies (Dykema et al. 2020; Olson et al. 2020). For example, we find substantial variation in estimated reading levels in battery items across the stem-with-first and stem-with-all approaches of calculating readability. We recommend that researchers use the stem-with-first approach, but this analysis should be replicated on other studies. Furthermore, the codes for question characteristics in SQP require human judgment and are prone to coder error (Bais 2021). QUAID provides text output that requires researcher decisions about how to operationalize the various identified problems. More work is needed to assess how different researchers use these tools and the implications for understanding question quality. Additionally, the conclusions made here are limited to surveys written in English. Future research should examine how these findings translate to multilingual surveys.

One interpretation of the variation in reading levels across readability measures is that readability measures are not useful for evaluating survey questions. However, because previous question evaluation tool studies (Maitland and Presser 2018; Tourangeau et al. 2021) have not consistently examined readability measures, and organizations are using readability as one evaluation measure for questions (Caporaso and Presser 2021), we urge the survey methodology field to continue research on readability measures, including whether a survey question-specific measure could be developed. We examined three question evaluation tools in a survey designer's toolkit. To save time and money, survey researchers may be tempted to rely on one computer-assisted question evaluation tool alone. This analysis suggests that designers should continue to use multiple question evaluation tools.

* * * * *

Authors' Note This research was conducted when R.S. was at University of Nebraska-Lincoln.

Conflicting Interests The authors have no potential conflicts of interest with respect to the research, authorship, and publication of this article.

Funding Data collection for this study was funded by the UNL Office of Research and Economic Development. The analysis was partially funded by Cooperative Agreement USDA-NASS 58-AEU-5-0023, supported by the National Science Foundation National Center for Science and Engineering Statistics.

Supplemental Material for this article is attached to the archive record for this document.

References

Bais, F. 2021. Constructing behaviour profiles for answer behaviour across surveys. Ph.D. Dissertation. Utrecht University, The Netherlands.

Betschart, P., D. Abt, H.-P. Schmid, P. Viktorin, J. Langenauer, and V. Zumstein. 2018. Readability assessment of commonly used urological questionnaires. *Investigative and Clinical Urology* 59:297–304.

Bruce, B., and A. Rubin. 1988. Readability formulas: Matching tool and task. In *Linguistic complexity and text comprehension*, 5–22. Hillsdale, NJ: Erlbaum Associates.

Caporaso, A., and S. Presser. 2021. Cognitive interviewing and survey pretesting: State of the art. Paper presented at the AAPOR Annual Meeting, May 2021. Virtual.

Coleman, M., and T. L. Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60:283–84.

Coltheart, M. 1981. The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A* 33:497–505.

Dillman, D. A., J. D. Smyth, and L. M. Christian. 2014. *Internet, phone, mail, and mixed-mode surveys*. Hoboken, NJ: Wiley.

Dykema, J., N. C. Schaeffer, D. Garbarski, and M. Hout. 2020. The role of question characteristics in designing and evaluating survey questions. In *Advances in questionnaire design, development, evaluation, and testing*, eds. P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis, and A. Wilmot, 117–52. Hoboken, NJ: Wiley.

Flesch, R. 1948. A new readability yardstick. *Journal of Applied Psychology* 32: 221–33.

Graesser, A. C., Z. Cai, M. M. Louwerse, and F. Daniel. 2006. Question Understanding Aid (QUAID): A web facility that tests question comprehensibility. *Public Opinion Quarterly* 70:3–22.

Gunning, R. 1952. *The technique of clear writing*. New York: McGraw-Hill.

Kincaid, J. P., R. P. Fishburne Jr., R. L. Rogers, and B. S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. Research Branch Report 8–75. Millington, TN: Naval Air Station Memphis.

Landis, J. R., and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33:159–74.

Lenzner, T. 2014. Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods & Research* 43:677–98.

Maitland, A., and S. Presser. 2018. How do question evaluation methods compare in predicting problems observed in typical survey conditions? *Journal of Survey Statistics and Methodology* 6:465–90.

McLaughlin, G. H. 1969. SMOG grading—A new readability formula. *Journal of Reading* 12:639–46.

Olson, K., J. D. Smyth, and A. Kirchner. 2020. The effect of question characteristics on question reading behaviors in telephone surveys. *Journal of Survey Statistics and Methodology* 8:636–66.

Paasche-Orlow, M. K., H. A. Taylor, and F. L. Brancati. 2003. Readability standards for informed-consent forms as compared with actual readability. *New England Journal of Medicine* 348:721–26.

Payne, S. L. B. 1951. *The art of asking questions*. Princeton, NJ: Princeton University Press.

Paz, S. H., H. Liu, M. N. Fongwa, L. S. Morales, and R. D. Hays. 2009. Readability estimates for commonly used health-related quality of life surveys. *Quality of Life Research* 18:889–900.

Saris, W. E., and I. N. Gallhofer. 2007. *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, NJ: John Wiley and Sons.

Senter, R. J., and E. A. Smith. 1967. *Automated Readability Index*. Dayton, OH: Wright-Patterson Air Force Base.

Tourangeau, R., A. Maitland, D. Steiger, and T. Yan. 2020. A framework for making decisions about question evaluation methods. In *Advances in questionnaire design, development, evaluation, and testing*, eds. P. C. Beatty, D. Collins, L. Kaye, J.-L. Padilla, G. B. Willis, and A. Wilmot, 47–73. Hoboken, NJ: Wiley.

Tourangeau, R., H. Sun, and T. Yan. 2021. Comparing methods for assessing reliability. *Journal of Survey Statistics and Methodology* 9:651–73.

Zhou, S., H. Jeong, and P. S. Green. 2017. How consistent are the best-known readability equations in estimating the readability of design standards? *IEEE Transactions on Professional Communication* 60:97–111.