Protocol for Guiding Future Evaluations of the Readability of the STAAR Assessment

OSP# 201902572-001

January 31, 2020

The Meadows Center
FOR PREVENTING EDUCATIONAL RISK

Purpose

The Meadows Center for Preventing Educational Risk was asked to recommend a methodology that the Texas Education Agency (TEA) and others can use to assess the readability level of the State of Texas Assessments of Academic Readiness (STAAR). This document addresses that task. We provide extensive background on the concept of readability, the ways in which readability has been defined and measured, the distinction between readability and difficulty, and the dangers of relying solely (or even primarily) on readability indices when evaluating high-stakes tests like the STAAR. Following this information, we recommend three components to include when assessing readability and discuss the tools we used to calculate metrics for these components in our evaluation of the readability of the 2019 and 2020 STAAR tests. Within the context of these recommendations, we address previous research regarding the readability of STAAR passages (Lopez & Pilgrim, 2016; Szabo & Sinclair, 2012, 2019).

Background

Historical Developments in Defining and Measuring Readability

Readability is a multifaceted construct that has been operationalized and measured in different ways (Benjamin, 2012). Traditional readability formulas developed in the mid to late 20th century were based on surface features of text such as word length, syllables per word, and words per sentence. Although these text features contribute to readability in important ways, these formulas do not account for other factors that contribute to text complexity, and the different indices can yield widely varying estimates of reading levels for the same text because of differences in the way in which they are calculated. Examples of first-generation measures include the Flesch-Kincaid (FK; Kincaid, Fishburne, Rogers, & Chissom, 1975), SMOG (McLaughlin, 1969), and Fry (Fry, 1968).

Second-generation readability formulas such as Lexile (MetaMetrics), the New Dale-Chall (Chall & Dale, 1995), and Advantage/TASA Open Standard (ATOS; Renaissance Learning; Milone & Biemiller, 2014) built on the traditional formulas by including a metric of vocabulary load (either how common words are or at what age they are typically learned) along with metrics such as word and sentence length (Benjamin, 2012). These formulas improved upon traditional formulas by adding a measure of vocabulary, which is a key element in reading comprehension. In addition, these approaches were validated through research linking them to student performance on reading comprehension items.

Third-generation approaches, such as Coh-Metrix (University of Memphis) and TextEvaluator (Educational Testing Services; Sheehan, Kostin, Napolitano, & Flor, 2014), include the components of earlier approaches but add deeper, semantic features of text and elements of text structure that affect comprehension (Benjamin, 2012; Nelson, Perfetti, Liben, & Liben, 2012). These more comprehensive approaches to assessing text complexity examine variables related to sentence structure, syntax, vocabulary load, text cohesion and coherence, and type and quality of inferences required for comprehension (Graesser et al., 2014; Sheehan et al., 2014). These approaches to determining readability have a strong theoretical basis in cognitive science and reading comprehension research. Recent research also indicates that these tools provide more accurate reading-level estimates across text types, correcting for the genre bias found in earlier formulas (Nelson et al., 2012; Sheehan et al., 2014).

Purpose of and Problems With Assessing Readability

Before describing the recommended methodology for assessing readability, it is important to define readability and related terms as they are currently understood within the research and practice communities. Current thinking views text readability not simply as an abstract feature of the text, but as a concept that describes the interaction of the reader with the text and ideally involves the process of matching a specific reader with a text that is proximal to that reader's comprehension skills (Goldman & Lee, 2014). Thus, readability requires information about both the text (e.g., its features, content) and the reader (e.g., reading ability, reading purpose). These twin aspects of readability are text complexity and text difficulty. Text complexity is an assessment of the text's features that places the text along a continuum of comprehensibility from less complex to more complex. Text difficulty, on the other hand, can only be determined with reference to a particular student and that student's level of reading comprehension skills, prior knowledge, and other cognitive skills that make texts more or less difficult for a particular reader. A text that is very difficult for one fourth-grade student may be much less so for another fourth-grade student.

Moreover, two texts that are comparable in their features (i.e., equally complex) may vary in *difficulty* for a particular reader based on the reader's knowledge level about the topic of one text compared to the other. For example, if a reader is presented with two texts of comparable complexity, one that describes events from 19th century American history and another that describes an episode during the Great Depression era, she may nonetheless find that the two differ in difficulty if she knows more about 19th century American history, for example, than about events from the 1930s. A second reader, one with the same general reading ability as the first reader, may find the passage about an episode during the Great Depression to be less *difficult* than the first reader finds it to be because she possesses greater prior knowledge about the 1930s. The two readers do not differ in reading ability and the passages do not differ in *complexity*. However, the same passage may differ in *difficulty* for the two readers based on the topical knowledge they bring to the text.

Further, a reader may find a text more or less difficult to comprehend depending on his or her reasons for engaging with that text. For example, a passage from a middle school English/language arts textbook will pose different degrees of difficulty depending on whether the reader is trying to determine its main idea or gain content knowledge about a literary time period. It is the same reader and the same passage; however, depending on the reader's assigned task, the passage's difficulty may differ.

By assessing readability in a way that places both the student's skills and the complexity of the text on the same scale, or metric, a teacher can provide a student with texts that are at the student's level and that are just beyond the student's level to foster growth in the student's reading skills. This approach to readability is implemented in tools such as ATOS and Lexile. Based on an ATOS or Lexile score from an assessment of a student's reading skills, teachers can match students to texts that are within reach given the student's ability. This approach makes readability an instructional tool to help students in their development of reading comprehension by challenging them with texts that are neither too complex nor too easy for them.

The recent application of readability indices for purposes other than matching readers with appropriate texts, such as to evaluate high-stakes assessments, has resulted in greater scrutiny of the validity of readability metrics when used to evaluate texts in isolation. One of the primary criticisms of readability measures is that the tools are often calibrated and evaluated against a criterion measure other than student reading comprehension but are then used to make determinations about the likelihood of a student being able to comprehend a particular text. Although recent research has examined the validity of some readability formulas against several criterion variables including expert ratings, student comprehension, and ratings of other tools (e.g., Nelson et al., 2012), some researchers argue against the va-

lidity of assessing readability in ways that fail to account for differences between readers and between reading tasks (Cunningham & Mesmer, 2014). In short, the interpretation and use of readability results is limited by the quality and type of criterion against which the tools were developed and validated.

In light of these compelling arguments, we recommend that any protocol for assessing text complexity and its suitability for assessing the reading comprehension skills of readers at a particular grade be used with caution. This document's recommended protocol represents one source of useful information, but it should be used in combination with other data sources when evaluating the appropriateness of high-stakes tests such as the STAAR. Additional sources of information may include an evaluation of content within the tested curriculum and item-level psychometric data. The utility of the high-stakes assessment in predicting outcomes of interest to students, parents, educators, and society, such as success in future grades, in postsecondary education, or in a career field, also should be considered.

The following protocol provides information on text features that influence text complexity. Remember that text complexity and text difficulty differ, as previously described, such that complexity depends on the text alone, whereas text difficulty depends on the characteristics of the text, the skills of the reader, and the purpose(s) for reading the text. The protocol can be used as one piece of evidence in determining whether a passage is likely a good choice for assessing the reading comprehension of a typical student in the target grade band because it is not excessively complex or too simple to be useful for measuring the curriculum content standards with which it is aligned. However, we strongly discourage the use of readability indices in isolation when evaluating the appropriateness of reading passages for high-stakes tests like the STAAR. Text complexity metrics on their own provide limited information about the suitability of a passage for measuring the reading comprehension abilities of students in a specific grade. Whether a specific text will provide valid and reliable information about students' reading comprehension ability in a particular grade is not readily determined by any measure of text readability. Ultimately, the utility of any text passage for inclusion on a high-stakes test that will be used for making important decisions about students, teachers, schools, and funding is not a function of the text complexity of the passage. Although it is certainly possible that a text can be too easy or too difficult for the majority of readers at a given grade, the utility of a given passage for use in assessment will depend on many factors other than the complexity of the passage's text.

Text complexity metrics have limited utility because the ease or difficulty of a passage of text for a particular student in a given grade is a complex interplay of factors (Kulesz, Francis, Barnes, & Fletcher, 2016), only some of which are captured by quantitative readability tools (Goldman & Lee, 2014). When assessing reading skills, the purpose of the assessment and how scores will be used must be considered (Messick, 1995). If we wish to scale the reading abilities of all students in a grade, we need one kind of test; if we wish to ascertain whether a student's ability exceeds some threshold of ability, a different kind of test may be needed. A given passage may be more or less useful for one purpose compared to the other. Readability metrics provide little guidance for determining the suitability of a text for each testing purpose.

Item Readability

The existing research on readability pertains primarily to passages of text. There is very little guidance and even less research on evaluating the readability of test items, other than a widespread recognition of the measurement challenges involved in assessing the readability of small segments of text. Due to the absence of prior research or established best practice in determining the readability of test items, we decided to evaluate the overall reliability of our protocol for assessing the readability of STAAR passages when we applied it to STAAR test items. To do so, we prepared the item text for analysis in several different formats, none of which should alter its readability (defined as text complexity). For example,

we formatted the text with line breaks included and excluded between the question and answer choices, we formatted it to include only the correct answer choice and all answer choices, we analyzed items individually and all items together as a test unit, and we tried more variations that did not change the attributes of the text itself. These kinds of differences should not result in different estimates of readability for the same item because they do not represent the features of text that are thought to influence its complexity. In other words, differences in the formatting of text should not create substantive differences in the readability metrics of the items. In all of these analyses, we used the same indices to determine the complexity of the text. In theory, similar results across different formats of the same text would support the relative "durability," or stability, of readability data when applied to test items.

However, our results showed the opposite pattern. When we compared the text complexity metrics of the same item formatted differently, the three indices that compose our protocol shifted substantially. The FK (a measure of text at the word and sentence levels) and narrativity (the index of vocabulary load) were particularly unstable; the index of the complexity of the syntax in the items was somewhat more reliable or consistent across differently formatted items. This finding of an inconsistent pattern of results aligns with the advice of others who work on questions of readability (Oakland & Lane, 2004); namely, a brief segment of text is insufficient as a sample for purposes of evaluating text complexity. Unless and until additional research provides clear guidance and evidence of a reliable way to evaluate item readability, we cannot recommend the analysis of the grade-based text complexity of test items. Therefore, the protocol presented below does not provide information on the complexity of text used in test items.

Additionally, it is important to note that item and test readability are distinct from item and test difficulty. Difficulty is a well-defined and widely researched property of tests and test items. In simple terms, item difficulty is a metric that reflects the level of ability required to successfully answer an item. Test difficulty, on the other hand, provides information on the range of abilities that can be measured on a test, and it is sometimes indexed by the ability level at which test scores are most precise. In more sophisticated terms, difficulty represents the amount of knowledge or ability a student must have in the domain being tested to have a prespecified probability of answering a particular item correctly or performing at a proficient level on a test. The complexity of text passages on a reading test is one component of item and test difficulty, but it has not been shown to be central; the link between item difficulty and item readability is even more tenuous. Research on accommodations for students with disabilities has shown that reading test items to students without disabilities (instead of having students read the items on their own) does not affect test performance (Fletcher et al., 2006). These findings suggest that the text complexity of items is not a significant factor in item difficulty for students without disabilities. Unless an item's readability is so far beyond a student's reading ability that the item is incomprehensible, measurement experts would expect that a student's mastery of the content standard being tested would be the primary factor in the likelihood of answering an item correctly.

Importantly for this protocol and our analysis of the 2019 and 2020 STAAR tests, an analysis of item and test difficulty requires a different approach than an analysis of readability. Test and item difficulty must be evaluated using specific methodologies that are well beyond the scope and purpose of this protocol. Analyses aimed at investigating STAAR item and test difficulty were not within the scope of work for this project. Therefore, results from the application of the protocol described below should not be used to draw conclusions about STAAR test difficulty or about likely student or school performance on the STAAR.

Protocol Components

The following protocol outlines a process for evaluating readability of STAAR passages based on the prevailing best practice for measuring text complexity. The protocol components were selected based on the strength of the research behind each and to maximize the benefits of each generation of readability formulas while incorporating the latest developments in the field that have addressed limitations.

The protocol outlines three text characteristics to evaluate when estimating text complexity. Grade-band information on each text characteristic can be obtained from a variety of tools; the protocol includes guidance for selecting among the available tools.

Text Characteristic 1: A Measure of Text Complexity Based on Word and Sentence Length

Despite their seeming simplicity, certain first-generation readability formulas that incorporate word and sentence length have been shown to correlate strongly with other approaches to determining text complexity and with reading comprehension scores (Graesser et al., 2014). Therefore, including at least one such measure in future assessments of the complexity of text on STAAR tests is recommended. However, the selection and use of first-generation readability formulas requires careful considerations of several key issues.

In earlier research on the readability of the STAAR, Lopez and Pilgrim (2016) and Szabo and Sinclair (2012, 2019) used as many as eight different readability formulas to evaluate passages from STAAR tests. This approach typically produces very different estimates of the likely grade level for the same passage of text, which is the pattern observed in these earlier studies. To the extent that complexity of a given text is immutable, it would make sense that the same passage would have a constant text complexity level, regardless of the method used for estimation; a fourth-grade passage is a fourth-grade passage. At most, one might expect minor variations across different readability indices for a passage. However, this expected result does not describe the earlier cited work, nor does it characterize the work of others who have used multiple formulas to estimate the text complexity of non-STAAR-related text (Gallagher, Fazio, & Gunning, 2012).

This variability in the measurement of text readability arises because different formulas include different dimensions of text complexity and involve different methods of calculation. For example, three commonly used formulas, the FK, SMOG, and Fry, compute readability in three different ways. The FK formula uses the average word and sentence length of a passage. The SMOG formula uses the number of words with three or more syllables and the total number of sentences in the passage. The Fry approach computes the average number of sentences and average number of syllables across three 100-word samples from the passage. Thus, the three readability metrics can yield widely varying results. Additionally, these three formulas established grade-level norms using varying corpora of texts as their criterion. For example, the Fry formula was initially calibrated using the grade level from a collection of books indicated by the publisher. The SMOG was calibrated using the grade levels indicated on the McCall-Crabbs test lessons in reading, a collection of reading passages with corresponding comprehension questions. Finally, some readability formulas were developed for a specific grade range (e.g., Spache Readability formula; Spache, 1953) or a specific purpose (e.g., the Linsear Write Formula was developed by the U.S. Air Force to gauge readability of technical manuals) and may yield diverging results when used outside the bounds of the developers' intent.

Another complicating factor in using many first-generation tools that measure word and sentence length is that they are in the public domain and not maintained by a specific entity. In particular, online

readability calculators that purport to use the same formula may not agree on the grade level of a passage due to differences in how the programmer implemented the formula in the calculator. In addition, first-generation tools are sensitive to small changes in text due to the nature of the features analyzed by the formula. A misplaced period or line break can alter the reported grade level of a text considerably. In particular, for brief text samples such as items or for genres with atypical structures such as poetry and drama, results can vary considerably depending on the method employed when preparing the text for analysis.

Further, the approach to combining the differing formulas described by Lopez and Pilgrim (2016) and Szabo and Sinclair (2012, 2019) is problematic. Calculating an *average* grade level for a given passage by averaging grade-level estimates based on different formulas assumes that the different indices can be combined as a mean. The same concern applies to the computation of average grade levels across several different passages on a specific STAAR test. First-generation formulas typically report the grade level for a passage as a grade or a grade equivalent. For example, the SMOG formula might rate a passage as 4.3, meaning that it is appropriate for students in the third month of fourth grade. The FK might rate the same passage as 5.1, meaning it is appropriate for students in the first month of fifth grade. A third index, the Fry, might report readability without the month and rate the same passage as 4, meaning it is appropriate for students in fourth grade. Averaging these ratings to get a readability level of 4.5 ignores two critical factors about the types of numbers that these formulas produce.

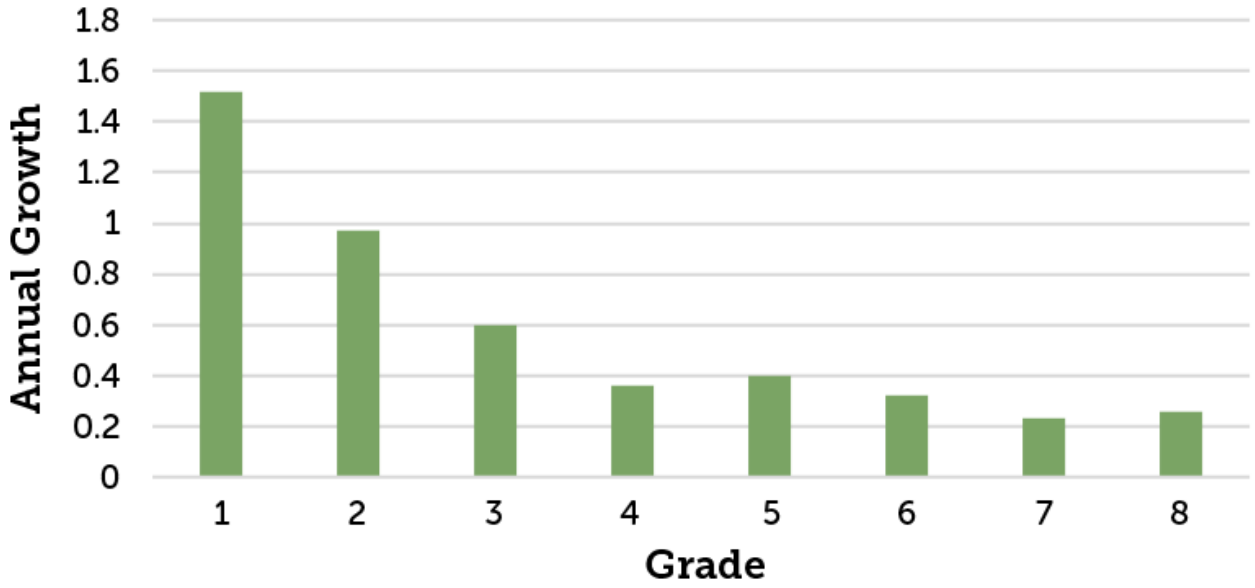
First, a grade-level readability result (with or without the month) is an ordinal number. Ordinal numbers communicate the rank order of values (i.e., first grade comes before fourth grade, which comes before seventh grade). However, ordinal numbers do not presume that the distance between each number is the same (i.e., texts at a 6th-grade level are not twice as hard to read as texts at a 3rd-grade level; the readability of 4th-grade passages is not one-third the readability of passages at the 12th-grade level). Grade levels assigned to passages by different readability formulas cannot be meaningfully added, multiplied, or divided in these ways. The reason is that all assigned grade levels represent ordinal numbers and do not reflect equal intervals on a scale of text complexity, such that equal differences between numbers (6 vs. 4 as compared to 7 vs. 5) imply equal differences in readability. All that these grade-level scores communicate about readability is that the passages with higher numbers are rated as more complex than passages with lower numbers. As a result, calculating an average readability by adding up the results of multiple readability formulas and then dividing that number by the number of formulas used is not an appropriate or meaningful way to determine the grade level of a passage.

A second problem with averaging results from multiple formulas is the differences in the formulas. The formulas are not linked or equated in such a way as to give the user confidence that a result of 4.3 on the SMOG and the FK, for example, mean the same thing. Indeed, results of the earlier cited work show that they do not. Furthermore, some formulas report the grade level without a month, leaving unclear whether a passage rated a 4, for example, is readable by students at the beginning of fourth grade or the end of fourth grade. Given the amount of reading growth that happens within early grades, passages rated as appropriate for grade 4 could vary considerably from one another on many aspects of text complexity.

To explain further why the averaging approach is problematic, it is important to understand that the amount of growth in reading that students typically gain in each grade declines as students advance through elementary school, middle school, and high school. Using data from seven nationally normed standardized tests, Bloom and other researchers documented that on average, students experience much more growth in reading skills in grades 1–3 than in later grades (Bloom, Hill, Black, & Lipsey, 2008). The following figure quantifies reading growth as the difference between reading scores at the end of one grade and the end of the following grade. It shows that the difference between a passage

rated as readable for students at the beginning vs. the end of third grade is much greater than the difference between a passage rated as readable for students at the beginning vs. the end of seventh grade. The interval between 3.0 and 3.12 is much larger than the interval between 7.0 and 7.12.

Figure 1. Amount of Reading Growth in Grades 1–8



Because the various readability formulas are not equated with each other and the intervals that represent the amount of reading growth in each grade are not the same, grade-level results from readability formulas should not be combined by averaging. This may seem like statistical nitpicking, but the consequences are real. As an example of the problems that arise when ordinal data is treated as if it has equal intervals and averaged, think of examining the high school class rankings of students who apply to a university. If one were to average the class rank across all students admitted to a university, one possible result might be 12.5. Clearly, this result is uninterpretable, as there is no such thing as a 12th and a half class rank. To extend this example, in looking at individual class rankings, we may know the order of students according to their academic performance (e.g., the student ranked 5th in the class had a higher grade-point average than the student ranked 10th), but unless we know the actual differences in their grade-point averages, we cannot draw conclusions about the amount of difference between the performance of the 5th-ranked and the 10th-ranked students. It is unlikely that the 5th-ranked student performed twice as well as the 10th-ranked student. We also cannot assume that the performance difference between the 5th- and 10th-ranked students is comparable to the difference between the 6th- and 11th-ranked students or even to the 5th- and 10th-ranked students at a different high school. The pairs differ by five when rank ordered. However, that difference provides no information on the difference in grade-point averages beyond their relative position in their class, and it may be that the difference between the 5th- and 10th-ranked students at one school is a fraction of the difference between the 5th- and 10th-ranked students at another school.

Having laid out these cautions for the ways in which first-generation readability formulas should and should not be used to assess text complexity, our recommended approach is to select a single formula to measure the word- and sentence-level characteristics of a passage. One of the most widely used first-generation tools is the FK grade-level estimate of readability. The formula has been shown to correlate highly with other measures such as Degrees of Reading Power and Lexile and with reading comprehension assessed at the sentence level using a cloze task (McNamara, Graesser, McCarthy, & Cai, 2014). This type of multiple-choice reading comprehension assessment requires students to select

the correct word to fill a missing word in a sentence. The FK also has been validated using traditional reading comprehension items on the Gates-MacGinitie reading test (Cunningham & Mesmer, 2014). As a result, the FK has demonstrated validity through a direct connection with student reading comprehension performance. Therefore, we selected the measure of text complexity based on word and sentence length for evaluating the text of the 2019 and 2020 STAAR tests.

Text Characteristic 2: A Measure of Vocabulary Load

A measure of word and sentence length does not capture fully the elements that make a text easier or more difficult to comprehend. For example, an informational passage may include one or more long words that are familiar to students because they are terms that are taught as part of grade-level content standards, such as *perpendicular*, which is included in the fourth-grade math content standards. The use of technical terms that students were taught explicitly has been shown to improve the readability of texts (Kachchaf et al., 2016). Additionally, students acquire some long words, such as *tomorrow* and *mountain*, relatively early (both are considered grade 3 or below vocabulary words). Despite students' familiarity with these words, their presence in a sentence typically will raise the grade-level readability on a first-generation formula compared to a sentence that consists of the same number of shorter words, even if the shorter words are more advanced, such as *theory* or *acre* (which are considered grades 6–8 vocabulary words).

To address this limitation, our protocol incorporates a measure of the vocabulary load of a text. Vocabulary load can be determined in a number of ways. These include the frequency with which the content[1] words in a passage appear in a corpus of texts commonly used in the K–12 curriculum, the age or grade when content words in the passage are acquired in oral and written language, and the frequency of use of different parts of speech in a text that make it more or less readable to students in a particular grade. Second-generation readability tools (e.g., Lexile) include vocabulary load in their formula. Third-generation tools typically include a measure of vocabulary both as part of the formula and reported as a separate index (e.g., ATOS, Coh-Metrix, and TextEvaluator). Vocabulary load is likely to be higher on content area texts, where domain-specific vocabulary is more prevalent, despite the fact that students have been taught these terms. As a result, evaluators should select appropriate subject area norms for grade bands that reflect the type of text involved.

In analyzing the 2019 and 2020 STAAR tests, we used the narrativity index in Coh-Metrix to measure vocabulary load because it describes the extent to which a text is "likely to contain more familiar oral language that is easier to understand" (McNamara et al., 2014, p. 85), which closely aligns with the notion of vocabulary load. The index is labeled "narrativity" because narrative (story-like) passages are characterized by frequent use of words that are acquired earlier in the development of language comprehension. Researchers have found that although the average narrativity score is higher for language arts texts than it is for social studies and science texts within each grade band, narrativity scores decrease (i.e., text becomes less narrative) as a function of grade level regardless of the subject area (Graesser, McNamara, & Kulikowich, 2011). Therefore, the narrativity index is an appropriate measure of vocabulary load for different text types and provides a robust estimate of the network of attributes that contribute to a text's vocabulary load. It encompasses commonly used metrics of vocabulary, such as age of acquisition and word frequency, and has been shown to correlate with grade level (McNamara et al., 2014).

1 The content words in a passage are the nouns, verbs, and adjectives.

Text Characteristic 3: A Measure of Syntax

An additional limitation of the first-generation readability formulas that is not addressed in the second-generation formulas is the complexity of the syntax, or structure, of a text. Although many tools use sentence length as a proxy measure of syntax, recently developed tools (e.g., Coh-Metrix, TextEvaluator) use a more sophisticated approach to measuring syntactic structure and report an index of syntactic complexity. Syntactically simpler text tends to comprise shorter sentences. However, text also is simpler when it contains fewer clauses and prepositional phrases. As a result, text written with simpler syntax requires less use of working memory and other cognitive resources, making its complexity fall at a lower grade level than the text written using more complex syntax (McNamara et al., 2014). Thus, syntax influences text comprehension (Graesser et al., 2014). Indeed, a measure of syntax can be used to rank texts in order of complexity (Graesser, McNamara, & Kulikowich, 2011). Texts that are suitable for earlier grades tend to have simpler syntax that is typical of narrative (story-like) passages. More complex syntax, typically found in informational texts, becomes more common in later grades. As a result, our protocol includes a measure of syntactic complexity. Such measures can be found in third-generation tools such as Coh-Metrix and TextEvaluator.

Tools for Evaluating Text Complexity

We selected Coh-Metrix (McNamara et al., 2014), a third-generation text analysis tool that provides more than 100 indices of text features, to evaluate word- and sentence-level text complexity, narrativity, and syntactic simplicity. Coh-Metrix is used throughout the measurement and evaluation communities for a variety of text analysis purposes. Thorough documentation on the development, validation, extant research, meaning, and interpretation of the indices is available in McNamara et al. (2014). This publication guided our selection of the three text characteristics to use in evaluating the text complexity of the STAAR.

Of the more than 100 indices included in Coh-Metrix, some are purely descriptive (e.g., mean and standard deviation of the number of words per sentence, mean and standard deviation of the number of syllables per word). Many others are included to inform research in fields such as linguistics, cognitive science, and literacy and have limited applicability to an evaluation of grade-level text complexity or insufficient research bases to allow them to be used in this way. According to McNamara et al. (2014), Coh-Metrix includes five indices specifically for use by educators interested in evaluating text complexity for instructional purposes. Of these five, three (narrativity, syntactic simplicity, and referential cohesion) were shown to correlate most strongly with grade level and with student performance on reading comprehension assessments when evaluated by a group of researchers who were not involved in the development of Coh-Metrix (Nelson et al., 2012). We initially considered referential cohesion as an additional text characteristic in our evaluation of STAAR text but decided to eliminate it for the sake of parsimony, as research indicates that it is not as strongly related to grade level as the other two indices (Graesser, McNamara, & Kulikowich, 2011) and of the three, syntax and vocabulary load are more established components of text complexity in the field of readability.

Although other tools are available that provide estimates for word- and sentence-level text features, vocabulary load, and syntactic simplicity, we selected Coh-Metrix in part because these tools did not meet the security requirements for this project and/or would not produce results that were able to be replicated. In the spirit of transparency, we used a process that, when replicated, should produce the results summarized in this report. Coh-Metrix software is available from its developers at The University of Memphis (**http://cohmetrix.com**). Other commercial vendors provide third-generation tools (e.g., TextEvaluator by ETS) and may be options to consider in future evaluations of STAAR text complexity.

Using Grade Bands to Evaluate Text Complexity

For each of the three text characteristic metrics, our methodology involves determining whether results fall within or below a grade band, defined as the tested grade and the two adjacent grades (i.e., +/- 1 grade). Grade bands are the most commonly used unit for evaluating readability because a text may not "uniquely represent one specific grade" (Nelson et al., 2012). In other words, a text may be appropriate for assessing the reading abilities of students in a range of grades depending on the specific purpose for which the passage is to be used on the reading assessment.

Some measures provide a grade-level estimate of text complexity. For example, a tool may estimate that a passage from a fifth-grade assessment is written at the fourth-grade level of complexity. In this example, because the grade band for a fifth-grade test is fourth–sixth grades, the estimate provided by the tool falls within the grade band for the tested grade. Other tools provide a range of grades for which the text would be appropriate. Here again, if the range encompasses the tested grade or the two adjacent grades, the estimate could indicate that the text is appropriate. Still other formulas provide an index score that is then compared to grade-level norms. In this instance, the upper and lower limits of the grade band would be defined as the mean score for the index of text complexity in the two adjacent grades (+/- 1) to the planned grade-level usage of the text. In this case, the index is on an equal-interval scale (unlike a grade equivalent), and calculating mean scores is appropriate.

Qualitative Judgment and Alignment to Content Standards

There is widespread agreement among experts that the limitations of quantitative readability measures warrant using a qualitative analysis that considers the reading task and text content to supplement quantitative results (Cunningham & Mesmer, 2014). For a criterion-referenced test such as the STAAR, additional considerations would be passage content and its alignment with the curriculum being tested. Prior to appearing on a STAAR assessment, test content undergoes a process that provides a qualitative judgment about content appropriateness for the tested grade and subject area. As part of the process, an item review committee comprising Texas teachers, administrators, curriculum specialists, and regional education service center staff members review items in terms of alignment, appropriateness for the grade level, any potential bias, and more (see TEA, 2018, for additional details). Presently, a qualitative review that attends to content and task factors takes place during STAAR test development. However, we recommend that this qualitative review be extended to include the process that we used to examine the alignment of passages used on the STAAR to the Texas Essential Knowledge and Skills content standards. Requiring an explicit examination of alignment for reading passages to the content standards would serve to highlight the importance of alignment as a qualitative judgment of the appropriateness of the text included in STAAR tests.

In the alignment study of the 2019 and 2020 tests, we defined items as the question, answer choices, and any accompanying passages, maps, graphs, charts, or figures. Alignment was defined as agreement between the knowledge and skills assessed by the item and those encompassed in the precoded content standard. When items are aligned with the content standard, students who have mastered the knowledge and skills in the corresponding student expectation would be expected to answer the item correctly. Aligned items may address only a portion of the precoded standard. For example, an item aligned with standard 4.11B ("Students are expected to distinguish fact from opinion in a text and explain how to verify what is a fact") may address only the first skill listed (distinguishing fact from opinion) and still be aligned. In addition, we used the TEA guidelines (2015) to explain to reviewers that examples following the terms "such as" and "including" do not represent the only examples that may provide the basis for an item. Items considered not aligned assess knowledge and skills that are not associated with the precoded student expectation.

Our procedure for evaluating item alignment with the precoded content standards involved convening a panel of MCPER staff and affiliated faculty members with content expertise and research and evaluation experience. In each subject area (reading, mathematics, social studies, science, and writing), two panelists independently coded each item as either aligned or not aligned. When panelists disagreed, a third panelist independently reviewed the item in question and made a final determination. We selected reviewers with leadership roles on research studies or professional development projects to serve as the third reviewer. Third reviewers were able to render an unbiased, expert judgment because they had not previously rated the items in question. When a rating of not aligned was assigned, the reviewer indicated the reason(s) for the rating and provided an alternative student expectation that more closely aligned with the knowledge and skills addressed in the item, if one existed. To evaluate the alignment of tests as a whole to the curriculum standards, if the alternative student expectation provided by the third reviewer was from the tested grade or any grade below, the item was considered aligned to the curriculum, even though it was judged as not being aligned to the precoded student expectation.

Using a Preponderance of Evidence to Evaluate Readability

Our protocol for evaluating STAAR passages considers the profile of results when evaluating the readability of an item or passage. Because of the previously described issues that occur when averaging results from text complexity metrics, we recommend making decisions based on the preponderance of evidence. Using multiple sources of information can provide a more complete profile of text characteristics that contribute to complexity. A passage would be deemed "readable" at the designated grade level if results from the majority of measures of text complexity recommended in this protocol (a word- and sentence-level metric, a metric of vocabulary load, a metric of syntactic complexity, and a qualitative determination of alignment to content standards) fall within or below the grade band that encompasses the test's grade level. The preponderance-of-evidence standard provides a way of synthesizing results across metrics of text complexity without computing averages.

Recommendations

Based on the research and best practices described above, we provide the following recommendations:

1. The application of readability metrics to items or passages on high-stakes tests such as the STAAR should be approached with great caution, if at all. The best use of readability information is in instructional practices that match students with texts that provide an appropriate degree of challenge to their reading comprehension skills. Uses of readability metrics that reflect text complexity in isolation of a particular student's reading skills are of limited value, particularly in the context of determining the appropriateness of an assessment.
2. Indices that measure text complexity do not provide information on the difficulty of items or passages. Item and test difficulty are distinct constructs from readability; assessing difficulty requires the application of specific, well-established methodologies that are quite different from those used to assess text complexity. Our results and this protocol should not be interpreted as providing information on the difficulty of the STAAR tests. To the extent that determining difficulty is of interest in future evaluations of the STAAR tests, we recommend that readability not be considered as an aspect of test or item difficulty.
3. If future investigations of the complexity of texts used in the STAAR tests are conducted, we recommend evaluating three text characteristics: word and sentence length, vocabulary load, and syntactic complexity. Further, we recommend selecting one measure of each characteristic, assessing text complexity relative to grade bands rather than a single grade level, and conducting a qualitative review of alignment with Texas curriculum content standards. The recommended criteria for reaching a conclusion about text complexity is one of preponderance of the evidence of these metrics.

References

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of texts. *Education Psychology Review, 24*, 63–88.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*(4), 289–328.

Chall, J., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline.

Cunningham, J. W., & Mesmer, H. A. (2014). Quantitative measurement of text difficulty. *The Elementary School Journal, 115*(2), 255–269.

Fletcher, J. M., Francis, D. J., Boudosquie, A., Copeland, K., Young, V., Kalinkowski, S., & Vaughn, S. (2006). Effects of accommodations on high-stakes testing for students with reading disabilities. *Exceptional Children, 72*, 136–150.

Fry, E. (1968). A readability formula that saves time. *Journal of Reading, 11*(7), 513–578.

Gallagher, T. L., Fazio, X., & Gunning, T. (2012). Varying readability of science-based text in elementary readers: Challenges for teachers. *Reading Improvement, 49*(2), 93–112.

Goldman, S. R., & Lee, C. D. (2014). Text complexity: State of the art and the conundrums it raises. *The Elementary School Journal, 115*(2), 290–300.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*, 210–22.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223–234.

Kachchaf, R., Noble, T., Rosebery, A., O'Connor, M. C., Warren, B., & Wang, Y. (2016) A closer look at linguistic complexity: Pinpointing individual linguistic features of science multiple-choice items associated with English language learner performance. *Bilingual Research Journal, 39*(2), 152–166.

Kincaid, P., Fishburne, R., Jr., Rogers, R., & Chissom, B. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Retrieved from https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary

Kulesz, P. A., Francis, D. J., Barnes, M. A., & Fletcher, J. M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology, 108*, 1078–1097.

Lopez, M., & Pilgrim, J. (2016). Text complexity: A study of STAAR readability. In E. Martinez & J. Pilgrim (Eds.), *Literacy summit yearbook* (pp. 87–93). Belton, TX: Texas Association for Literacy Education.

McLaughlin, G. H. (1969). SMOG grading—A new readability formula. *Journal of Reading, 12*(8), 639–646.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749. doi:10.1037//0003-066X.50.9.741

Milone, M., & Biemiller, A. (2014). *Development of the ATOS readability formula*. Wisconsin Rapids, WI: Renaissance Learning.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. Retrieved from https://achievethecore.org/page/1196/measures-of-text-difficulty-testing-their-predictive-value-for-grade-levels-and-student-performance

Oakland, T., & Lane, H. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4*, 239–252.

Sheehan, K. M., Kostin, I., Napolitano, D., & Flor, M. (2014). The TextEvaluator tool: Helping teachers and test developers select texts for use in instruction and assessment. *The Elementary School Journal, 115*(2), 184–209.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal, 53*(7), 410–413.

Szabo, S., & Sinclair, B. (2012). STAAR reading passages: The readability is too high. *Schooling, 3*(1), 1–14.

Szabo, S., & Sinclair, B. (2019). Readability of the STAAR Test is still misaligned. *Schooling, 10*(1), 1–12.

Texas Education Agency. (2015). *An explanation of the terms such as and including on STAAR*. Retrieved from https://tea.texas.gov/sites/default/files/STAAR%20Such%20As_Including%20Policy.pdf

Texas Education Agency. (2018). *Technical digest 2017–2018: Chapter 2: Building a high-quality assessment system*. Retrieved from https://tea.texas.gov/sites/default/files/TechDigest_2017_2018_Chapter2_r4-for%20tagged.pdf

Suggested Citation

The Meadows Center for Preventing Educational Risk. (January 2020). *Protocol for guiding future evaluations of the reliability of the STAAR assessment*. Austin, TX: Author.