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

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 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

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 diff 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).

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

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.

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*