Clarification Regarding STAAR & Lexile   2/22/2019

Context
- **STAAR Measures Texas Standards:** STAAR is an end-of-year test given to students in Texas in several subjects, including reading, starting in 3rd grade. It is designed to determine whether students have mastered the Texas State Board of Education-adopted grade level-specific knowledge & skills.
- **Lexile & Readability Algorithms:** Lexile is a widely-referenced computer-based algorithm to assess "readability". These computer-based algorithms generally define readability based on the number of syllables per word, the number of words per sentence, and the number of sentences per paragraph. Most books in publication have been scored using the Lexile algorithm. For this reason, TEA uses Lexile to help parents find books to support their children's reading growth.
- **Lexile Concerns:** Based on results from readability algorithms like Lexile applied to STAAR reading passages, some have voiced concerns that STAAR Reading tests are above grade level for students, and are therefore too difficult for many students to pass.
- **Student Performance Concerns:** Some have gone so far as to assert that many students who have performed poorly on STAAR Reading have been improperly evaluated by STAAR, and that elementary reading performance in Texas is far better than is shown with STAAR results.

Key Points
1. **The Real Problem: Texas Reading Performance is Declining.** Even if we ignore results from STAAR, NAEP reading scores show that 4th grade Texas elementary reading performance has been flat or declining since 2007.
   a. The National Assessment of Educational Progress (NAEP) is a test administered in all US states every other year used to compare performance across states over time.
   b. The decline in Texas NAEP reading performance has occurred while the nation has improved. See Figure 1.
2. **STAAR & Prior Texas Tests Show Similar Results.** Reading scores on TAKS (the pre-STAAR Texas test) were flat or moved by 1 point, starting in 2007, similar to NAEP results. STAAR results continue to look similar to NAEP. See Figure 2.
3. **Texas Districts Focused on Reading Improvement Are Getting Results.** While the state as a whole has been flat or declined in early reading performance, multiple districts have proven that STAAR reading score improvements are possible, regardless of district size. Some examples from 2015 to 2018 include:
   a. La Pryor ISD, with 468 students, 80% low income, grew 3rd & 4th grade reading STAAR scores by 42% & 19% for students at or above the Approaches grade level standard.
   b. Laredo ISD, with 24,000 students, 96% low income, grew 3rd & 4th grade reading STAAR scores by 12% & 16% for students at or above the Approaches grade level standard.
   c. Dallas ISD, with 157,000 students, 87% low income, grew 3rd & 4th grade reading STAAR scores by 9% & 5% for students at or above the Approaches grade level standard.
4. **Lexile is Inappropriate to Use as the Measure of Being on Grade-Level.** Readability algorithms like Lexile are useful for a variety of reasons, but not for evaluating grade-level appropriateness of passages used on a test of Texas standards.
   a. From Metametrics, owners of Lexile: "Lexile measures are not generated from grade-level norms and do not presume a specific grade-level interpretation…. Rather, students

have an independent Lexile measure and can select appropriately difficult books in their Lexile range."

b. Computer-based readability algorithms like Lexile produce results that are sometimes inconsistent with human judgment of grade-level appropriateness. *The Grapes of Wrath* has a Lexile of 680, while Dr Seuss's *If I Ran the Zoo* has a Lexile of 870.

c. As with the STAAR, other state's standardized tests include passages that span a wide range of Lexile readability levels. This is because these tests are built to measure their state's grade level standards, not to test Lexile or other readability algorithms.

5. **Texas Teachers Confirm Texas Tests Have Grade-Level Appropriate Passages.** Like other Texas tests before it (TAKS, TAAS, etc.), STAAR test questions are created using a detailed process that includes not only content experts but also approval by current practicing Texas teachers.

a. The Texas teacher teams ensure every passage/item meets a four-part test, with one being that they are grade-level appropriate.

b. This process was not materially different for STAAR than it was for any prior Texas test, and is a standard best-practice used for state tests across the nation.

c. As an example, in 2018, 58 3rd grade teachers from all regions of Texas reviewed 3rd grade passages and assessment items.

6. **STAAR Validity Has Been Verified, Including Being Grade-Level Appropriate.** The legislature required a third-party evaluation of STAAR in 2016 to ensure that it was valid. The evaluation included confirming the "test bears a strong association with on-grade curriculum requirements". See https://tea.texas.gov/student.assessment/reports/

a. Additionally, STAAR has been shown to correlate with important life outcomes. Students who Master Grade Level on STAAR have a 75% chance of passing freshman college classes. See the Standard Setting Technical Report: https://tea.texas.gov/student.assessment/staar/performance-standards/

7. **TEA Improved the Test-Development Process Based on Previously-Raised Concerns.** The agency heard concerns in 2016 about passage readability based on Lexile scores. Although Lexile and other readability measurements are not aligned to Texas standards and don't determine validity, Lexile is still useful and widely referenced. Given this, the agency modified the test-development process in response to these concerns to increase public confidence in the STAAR.

a. The test-development process was updated in 2017 with the goal of ensuring all passages on STAAR are within Lexile-specific grade level bands. See Memo 1.

b. The 2018 STAAR reading tests reflect the first phase of this update, with nearly all reading passages across grades 3-8 falling within the stretch band.

c. STAAR results remained largely flat, even with this change.

8. **TEA Provides Lexile Resources to Help Parents Find Good Books.** TEA has updated STAAR report cards for parents to include resources to use Lexile to help find a good book to support a love of reading in their children. This Lexile information should not be misinterpreted as a characteristic of STAAR, or for any other use than its intended purpose.

a. To ensure the Lexile advice given to parents was appropriate, TEA conducted a Lexile Linking Study in 2017 testing a sample of students in Texas on a Lexile-aligned assessment. A mathematical model was used to link the scores of those students on the Lexile exam back to their prior STAAR scores. See https://tea.texas.gov/student.assessment/reports/

Figure 1

NAEP - Reading Grade 4 - Average Scale Score

Figure 2

Percent of Students at Minimum Standard in 4th Grade Reading in Texas
on TAAS/TAKS/STAAR and NAEP

From 1994 to 2007, there was consistent growth in reading performance at the minimum passing standard on state assessments (TAAS & TAKS), and this largely mirrored the trend from NAEP.

From 2007 to 2017, there was largely flat or declining performance at the minimum passing standard on state assessments (TAKS & STAAR), and this largely mirrored the trend from NAEP.
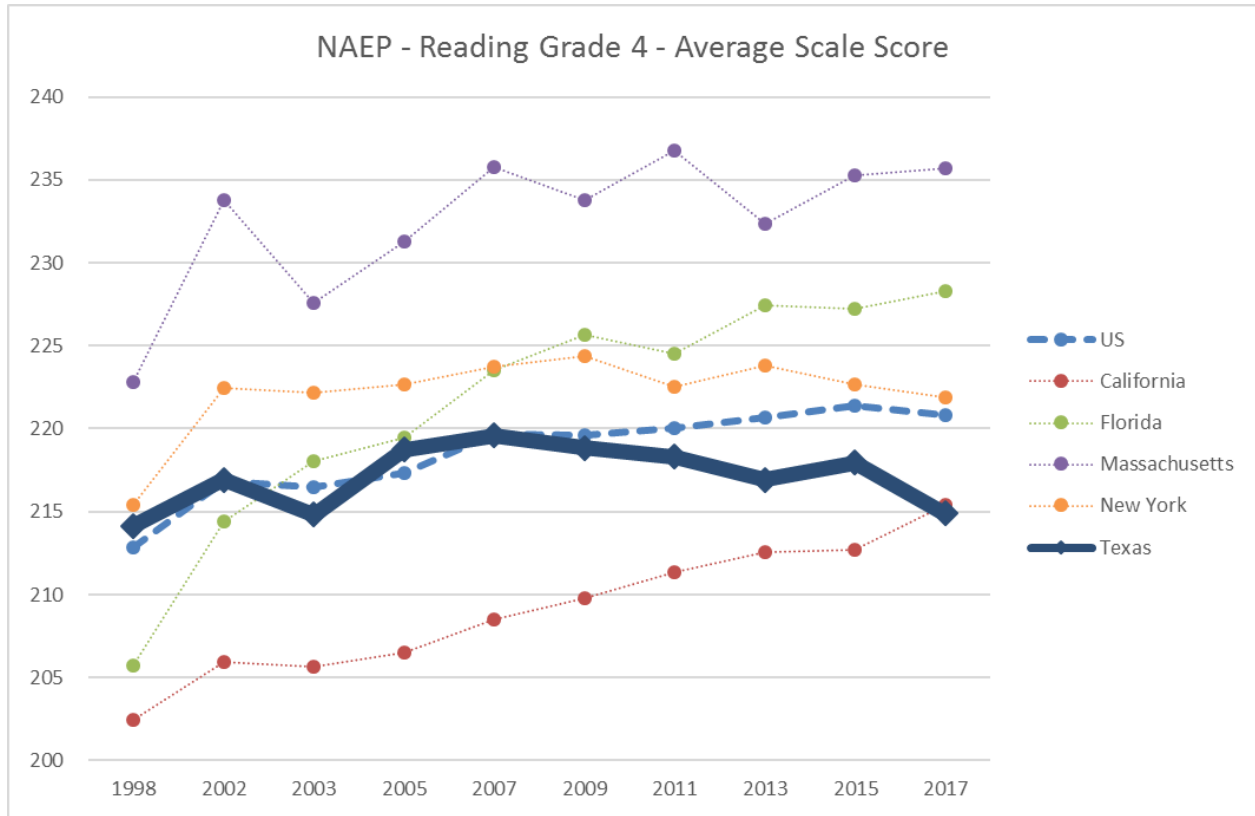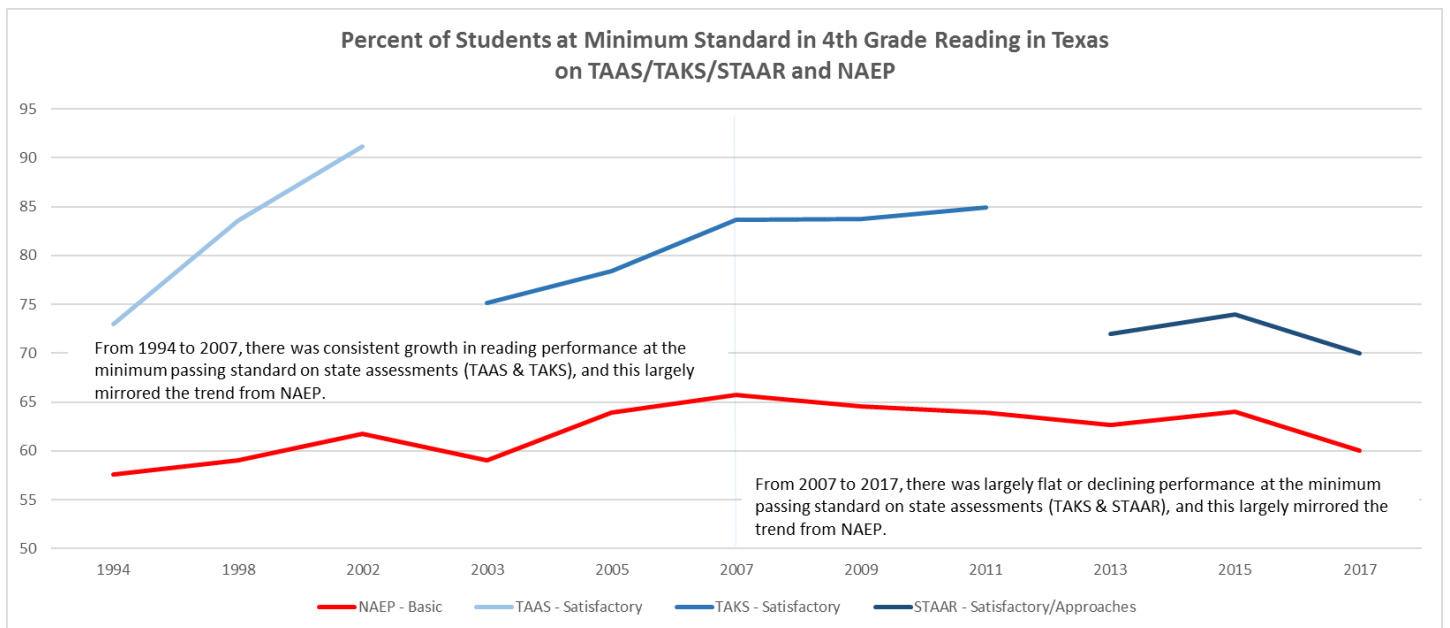
Memo 1: Lexile Grade Bands and Test Specifications: 2018–2020

In March of 2017 TEA adopted the Lexile grade-band framework below and established the requirement that all passage Lexile scores fall within their prescribed grade-band ranges. Two grade-band ranges are represented in Table 1 below. The "Current Lexile Band" was established by MetaMetrics in 2009; in 2012, MetaMetrics established the "Stretch Lexile Band" in response to the national movement toward increased rigor and emphasis on college and career readiness.

Table 1: The Lexile Grade-Band Framework

Grade Band	Current Lexile Band	Stretch Lexile Band
K–1	N/A	N/A
2–3	450L–730L	420L–820L
4–5	640L–850L	740L–1010L
6–8	860L–1010L	925L–1185L
9-10	960L–1120L	1050L–1335L
11–CCR	1070L–1220L	1185L–1385L

During a transitional period, TEA has identified target proportions for these bands with operational and field test administrations. These targets are desired, but not absolutes and will depend upon the robustness of the passage pool and feasible efforts to supplement this pool swiftly. In this transition, TEA will define a grade band as the score range that is marked by the bottom of the current band and the top of the stretch band. (For example, the grades 4–5 band extends from 640L to 1010L.) However, to facilitate a transition to full alignment to the Current Lexile Band over the next three years, TEA will phase in incremental targets for the Current band and will tighten limits on use of the Stretch band in operational forms and field-test developments.

Table 2: Targets and Limits on the Current and Stretch Bands

Year	Forms	Current	Stretch
2018	Operational	50%	50%
	Field Test	50%	50%
2019	Operational	50% (or higher)	50% (Max)
	Field Test	80%	20%
2020	Operational	80% (or higher)	20% (Max)
	Field Test	100%	0%

TEA has also established the guideline that text excerpts that are out of Lexile range for the intended grade may still be appropriate and eligible for use in that grade if the Lexile score for the extended work is within range. It is important to emphasize the process within which this special consideration is applied:

- Content experts review a grade-appropriate work of literature to identify excerpts that are appropriate for assessment of the TEKS. Readabilities are run for the excerpt.

- If the Lexile score for the excerpt is within grade-band range, then the excerpt remains eligible. If the Lexile score is out of range but the extended work is within range, then the passage may remain eligible if, upon further review, TEA content experts still deem the passage to be grade appropriate and useful for assessment of the TEKS.

To limit the extent to which this consideration is applied, we recommend that it remain applicable only for grade-appropriate fiction and in rare instances for literary nonfiction. Developers should use discretion in its application during each development, and an operational form should never include more than one of these cases. Additionally, developers must ensure that the number of passages that meet this criteria in a given development year does not result in a lack of sufficient passages available for use.

In addition, MetaMetrics has identified in a peer reviewed journal* an acceptable confidence interval that can be applied to shorter excerpts of text. This confidence interval is +/- 64 for a particular score. This confidence interval can be used as additional information in assigning a passage to a grade level. For example, a passage that scores 625 may still be considered in the grade 4-5 band since the score falls within a range of 64 below 640.

To limit the extent to which this exception occurs, we recommend that only one passage on an operational form may fall within either the confidence interval or the extended work exception; only one exception per form. Here, too, developers should use discretion in its application during each development.

*Stenner, A.J., Burdick, H., Sanford, E., & Burdick, D. (2006). "How accurate are Lexile text measures?" *Journal of Applied Measurement 7*(3), pp. 307-322.

APPENDIX

A justification for using fiction excerpts that are below Lexile range

Readability formulae have not yet reached a point where they can be adjusted to account for differences in genre (fiction/nonfiction). To date, readability formulae have been a "one size fits all" measure, even though there are substantial differences between fiction and nonfiction writing. A large corpus of research related to readabilities exists, but much of this work addresses informational text. This is due to the relative importance of informational text in adult life: materials such as legal and medical documents have serious consequences for adults and thus have been the focus of considerable attention. Fiction has been studied much less. One study that does address fiction concluded that fiction is, indeed, quantitatively different from informational text. The authors suggest that a separate readability index be created for fiction (literature), but note that the amount of data needed to complete this task has been an impediment.*

Readability formulas (including Lexiles) use surface-level attributes to produce a score: number of syllables per word, number of one-syllable words per sentence, number of characters per word, etc. These characteristics are important to reading comprehension, but they do not address deeper meaning within text or all factors that contribute to text complexity.

Aspects of literature not measured by readability statistics:

1) Engagement
2) Mood or tone
3) Multiple-meaning words ("bear" is a mammal, but also a complex verb)
4) Topic
5) Narration vs introspection and internal exploration
6) Quality
7) Novelty of concepts
8) Intention of author (example: Malcolm Gladwell intentionally writes about complex topics at a more accessible level)
9) Sophisticated aspects of style and/or structure (irony, sarcasm, word play)
10) Knowledge demand/required background knowledge

One means of addressing these aspects of literature is to supplement the readability level with a text complexity rubric that asks various stakeholders (content experts from ETS and TEA, or Educators) to make an expert judgment about an excerpt. If the Lexile acts as the backbone of the grade-level assignment, the text complexity rubric can act as an appendage. The two pieces of information could work in concert to give a fuller, more nuanced representation of the text. This approach could mitigate any effects of how readabilities do not account for all elements of complexity in literary texts.

* Dell'Orletta, F., Venturi, G., & Montemagni, S. (2012). Genre-oriented Readability Assessment: A Case Study. Proceedings of the Workshop on Speech and Language Processing Tools in Education, 91-98. Retrieved May 30, 2018.