Listening. Learning. Leading.®

Research Report

ETS RR–15-21

# Aligning *TextEvaluator*® Scores With the Accelerated Text Complexity Guidelines Specified in the Common Core State Standards

Kathleen M. Sheehan

December 2015

# ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

# Aligning *TextEvaluator®* Scores With the Accelerated Text Complexity Guidelines Specified in the Common Core State Standards

Kathleen M. Sheehan

Educational Testing Service, Princeton, NJ

The *TextEvaluator®* text analysis tool is a fully automated text complexity evaluation tool designed to help teachers, curriculum specialists, textbook publishers, and test developers select texts that are consistent with the text complexity guidelines specified in the Common Core State Standards. This paper documents the procedure used to align the TextEvaluator reporting scale with the Common Core text complexity scale and provides score ranges for use when placing texts into grade bands. Three evaluations of the proposed score ranges are reported: one implemented with respect to the set of 168 exemplar texts provided in Appendix B of the Common Core State Standards, one implemented with respect to a set of 10 career texts, and one implemented with respect to a set of 59 texts selected from textbooks assigned in first-year, credit-bearing college courses. Results suggest that the proposed ranges can help users determine an appropriate grade band placement for any text that has been evaluated by TextEvaluator, including informational, literary, and mixed texts.

**Keywords** Common Core State Standards; text complexity

doi:10.1002/ets2.12068

The Common Core State Standards (CCSS) accelerate text complexity expectations for students in Grades 1 through 12 in order to close what has been described as a "vast" and "alarming" gap between the complexity levels of the texts that students read in high school (HS) and those that they will be expected to read in college and careers (Common Core State Standards Initiative, 2010, Appendix A, pp. 2–3). For example, Common Core (CC) aligned textbooks targeted at students in Grades 9–10 will now be written at the reading levels previously designated for students in Grades 10–12 (Hiebert & Mesmer, 2013a). Williamson, Fitzgerald, and Stenner (2013) noted that this new trajectory is designed to "close a current-day gap between text complexity levels at high school graduation versus college and workplace," and that it "intentionally target[s] higher levels of text complexity than many, if not most, students currently experience in nearly all grades" (p. 59).

Automated text complexity evaluation tools have been proposed as a way to help teachers, curriculum specialists, textbook publishers, test developers, and other educators select texts that are closely aligned with these new text complexity expectations. Because many different tools are available, research funded by the Bill and Melinda Gates Foundation was conducted to provide "state of the science information regarding the variety of ways text complexity can be measured quantitatively and to encourage the development of text complexity tools that are valid, transparent, user friendly, and reliable" (Council of Chief State School Officers, & the National Governors Association, 2014, p. 1). Analyses implemented as a part of this research are reported in Nelson, Perfetti, Liben, and Liben (2012). The researchers concluded that text complexity classifications generated via each of six different automated text complexity evaluation tools are appropriately structured to "guide curriculum decisions, assist assessment development, and support the efforts of educational publishers to meet complexity guidelines" (Nelson et al., 2012, p. 3).

The set of six tools evaluated in Nelson et al. (2012) included the *SourceRater*$^{\text{sm}}$ engine, a quantitative text complexity evaluation tool developed at ETS. Over the past few years, researchers at ETS have continued to update and refine the SourceRater engine. Recently, the updated, refined engine was renamed the *TextEvaluator®* text analysis tool. The procedures used to develop and validate the enhanced TextEvaluator engine are described in Sheehan (2014a), Sheehan, Flor, and Napolitano (2013), Sheehan, Kostin, and Futagi (2013), and Sheehan, Kostin, Napolitano, and Flor (2014). This

*Corresponding author*: K. M. Sheehan, E-mail: ksheehan@ets.org

**Table 1** Three Approaches for Establishing an Alignment Between the Scores Generated via an Automated Text Complexity Evaluation Tool and Human-Assigned Text Complexity Scores Designed To Be Consistent With Common Core Text Complexity Expectations

| Element | Approach 1 | Approach 2 | Approach 3 |
|---|---|---|---|
| Model Lexile scores as a function of GL classifications assigned by human experts.[a] | ✓ | | |
| Model GL classifications assigned by human experts as a function of TextEvaluator scores | | ✓ | ✓ |
| Quantify expected variation within a GB via the interquartile range (difference between 75th and 25th quantiles) | ✓ | | |
| Quantify expected variation within a GB via a $(1 - \alpha)$ confidence interval based on a $t$ distribution with $n - p$ degrees of freedom where $n$ is the number of texts and $p$ is the number of estimated parameters | | ✓ | ✓ |

*Note*. GL = grade level, GB = grade band.

[a] When implementing this approach for tools other than the Lexile tool, also use an equipercentile equating function to translate score ranges expressed on the Lexile scale into score ranges expressed on any other scale (e.g., the Flesch–Kincaid scale, the ATOS scale, the DRP scale and/or the Reading Maturity scale).

paper documents the analyses implemented to establish an alignment between the TextEvaluator reporting scale and the accelerated text complexity exposure trajectory referenced in the CCSS.

This paper is organized as follows. First, the approach used to define the accelerated text complexity exposure trajectory referenced in the CCSS is reviewed. Second, three approaches for establishing an alignment between that trajectory and the scores generated via an automated text complexity evaluation tool are evaluated. Table 1 highlights similarities and differences among the three alignment methodologies studied. Approach 1 is the approach that was used to define the score ranges provided in Appendix C of Nelson et al. (2012). These score ranges are also provided in the Supplement to Appendix A published by the CCSSO and NGA (2014). Approaches 2 and 3 differ from Approach 1 with respect to a fundamental element of the estimation approach: the criterion variable selected for use when defining the needed alignment function. As is indicated in Table 1, Approach 1 is implemented with Lexile scores (Stenner, Burdick, Sanford, & Burdick, 2007) as the criterion variable. By contrast, both Approach 2 and Approach 3 are implemented with grade level (GL) classifications assigned by human experts as the criterion variable. Table 1 also highlights a second distinguishing feature of the proposed approaches: the confidence probability adopted for use when constructing approximate confidence bands. Additional information about these three alignment methodologies is summarized below. Because each alignment methodology is designed to yield text classifications that are consistent with the accelerated text complexity exposure trajectory specified in the CCSS, the paper begins with a review of the method used to specify that trajectory.

## The Procedure Used to Specify the Accelerated Text Complexity Exposure Trajectory Presented in the Common Core State Standards

Williamson, Fitzgerald, and Stenner (2013) noted that "The CCSS set a controversial aspirational, quantitative trajectory for text complexity exposure for readers throughout the grades, aiming for high school graduates to be able to independently read complex college and workplace texts" (p. 59). This section provides a detailed description of the method used to establish that trajectory.

### Overview

The Common Core (CC) text complexity exposure trajectory was developed in four steps. First, a text complexity exposure trajectory intended to represent current-day elementary and secondary reading demands was estimated. Second, an "end-of-high-school target for text complexity exposure" (Williamson et al., 2013, p. 59) was estimated. Third, the difference between the Grade 12 point on the current-day text complexity exposure trajectory and the end-of-high-school target for text complexity exposure was calculated. Fourth, the resulting gap was distributed across the GLs so that the Grade

12 point on the accelerated text complexity exposure trajectory was equal to the proposed HS graduation target. The following paragraphs provide additional information about the data, methods, and assumptions employed at each step.

### Step 1: Quantify Current-Day Elementary and Secondary Reading Demands

A collection of 487 texts was assembled for use in quantifying current-day elementary and secondary reading demands (Williamson, Koons, Sandvik, & Sanford-Moore, 2012). All texts were selected from the textbook adoption lists published in Florida, Georgia, Indiana, North Carolina, Oregon, Texas, and Virginia. Each GL sample consisted of approximately 41 textbooks, although actual sample sizes ranged from 27 to 61. To ensure adequate coverage of the full range of texts likely to be encountered by students in Grades 1 through 12, each grade-specific sample was constrained to include texts selected from each of the following content areas: health, language arts, literature, mathematics, science, and social studies (Stenner, Koons, & Swartz, 2009, p. 4). Note that this constraint implies that both informational and literary texts were included at each GL.

Next, the Lexile tool (Stenner et al., 2007) was used to generate an estimated text complexity score for each text, the median text complexity score at each GL was determined, and the resulting set of 12 grade-specific medians was modeled as a polynomial function of GL. Williamson et al. (2013) reported that the resulting set of 12 medians was "very well modeled ($R^2 = .99$) by a fifth-degree polynomial" (p. 63).[1] The resulting smoothed curve placed the current-day reading demands of students in Grade 1 at a Lexile score of 310 and those of students in Grade 12 at a Lexile score of 1130.

### Step 2. Select an End-of-High-School Target for Text Complexity Exposure

Analyses designed to select an end-of-HS target for text complexity exposure were reported in Stenner, Sanford-Moore, and Williamson (2012), Williamson (2008), Williamson et al. (2012), and Williamson et al. (2013). A total of 2,990 texts were evaluated, including texts selected from each of the following sources: workplace documents ($n = 1{,}401$), citizenship documents ($n = 54$), military documents ($n = 22$), textbooks assigned in first-year credit-bearing courses at technical colleges ($n = 81$), textbooks assigned in first-year credit-bearing courses at community colleges ($n = 161$), textbooks assigned in first-year credit-bearing courses at universities ($n = 294$), articles extracted from Wikipedia ($n = 945$), and articles extracted from international newspaper ($n = 32$).

Stenner et al. (2012) reported that this diverse sample yielded a median Lexile score of 1300 and an interquartile range that extended from about 1200 to 1380 and argued that these summary statistics "represent the level of reading ability required for college and career" (p. 2). As is illustrated on page 3 of Stenner et al.'s (2012) report, however, this description is misleading because the upper tail of the distribution is almost entirely determined from two types of noncollege, noncareer texts: articles extracted from Wikipedia ($n = 945$) and articles extracted from international newspapers ($n = 32$). Analyses designed to quantify the increases in college and career text complexity resulting from the decision to include Wikipedia articles and articles from international newspapers in the college and career sample are documented in the appendix.

### Step 3. Quantify the High School/Postsecondary Text Complexity Gap

The HS/postsecondary text complexity gap was then estimated as follows:

$$\text{Gap} = 1300 - 1130 = 170, \qquad (1)$$

where 1300 is the median Lexile score estimated from the combined sample of college, workplace, newspaper, and Wikipedia text assembled at Step 2, and 1130 is the Grade 12 point on the current-day reading demand curve estimated at Step 1. Sanford-Moore (2013) noted that this estimate (i.e., 170 Lexile points) can be interpreted as the amount by which text complexity expectations for students must be stretched in order to ensure that students are adequately prepared for the advanced reading demands of college and careers. Note, however, that less stretching would have been needed if the postsecondary sample had not been expanded to include Wikipedia articles and articles from international newspapers. The additional estimation errors introduced as a result of this expansion are investigated in the appendix.

### *Step 4. Stretch the Current-Day Reading Demand Curve So That the Gap Is Eliminated*

Alternative approaches for stretching the current-day reading demand curve have been proposed. Two approaches are described below: the approach proposed in Williamson et al. (2013) and the approach proposed in Sanford-Moore and Williamson (2012). Both approaches are described because the former is more well-known, yet the latter is the approach that was actually adopted by the CCSSO and the NGA for use in the CCSS.

The approach described in Williamson et al. (2013) starts with the current-day reading demand curve estimated in Step 1 expressed as a polynomial function of GL as follows:

$$\hat{y} = 9.0909 + 364.76x - 67.61x^2 + 6.9919x^3 - .348x^4 + .0065x^5, \tag{2}$$

where $\hat{y}$ is the expected Lexile score of a text with a known GL classification of $x$. This curve is then stretched so that it passes through a Lexile score of 1385 at Grade 12. As noted in Williamson et al. (2013), 1385 is the 75th quantile of the complexity distribution obtained when the definition of a postsecondary text is expanded to include Wikipedia articles and articles from international newspapers. Because the smoothed curve has six estimated coefficients (including the intercept term), there are six ways that it can be altered to pass through a Lexile score of 1385 when $x = 12$. Each altered curve is obtained by fixing five of the six estimated coefficients at the values shown in Equation 2, setting $\hat{y} = 1385$ when $x = 12$, and then solving for the sixth coefficient. Williamson et al. (2013) argued that any of the resulting set of six curves can be used to define a text complexity exposure trajectory that culminates in the reading ability needed to comprehend college and career texts.

A seventh approach for stretching the current-day reading demand curve is presented in an earlier paper by Sanford-Moore and Williamson (2012). This alternative approach can be summarized as follows. First, the total growth in reading ability needed to be adequately prepared for college and careers is estimated as follows:

$$\text{Total College and Career Growth} = 1300 - 310 = 990, \tag{3}$$

where 1300 is the median Lexile score estimated from the combined sample of college, workplace, citizenship, newspaper, and Wikipedia text estimated at Step 2, and 310 is the Grade 1 point on the current-day reading demand curve estimated at Step 1. Next, a proportional allocation approach is used to distribute this growth across Grades 2 through 12. The proportional allocation approach is designed to preserve current-day estimated grade-by-grade growth trends to the extent possible. For example, the current-day reading demand curve estimated at Step 1 suggested that 25% of the total growth from Grades 1 to 12 occurred in the interval between Grades 1 and 2, while just 5% of that growth occurred in the interval between Grades 10 and 11. Consequently, 25% of the total growth estimated in Equation 3 was allocated to the interval between Grades 1 and 2, and just 5% of that growth was allocated to the interval between Grades 10 and 11. Both Sanford-Moore and Williamson (2012) and Stenner et al. (2012) argued that the resulting curve culminates in the reading ability needed to comprehend college and career texts while preserving current-day grade-by-grade growth trends.

The accelerated text complexity exposure trajectory obtained via the above process is illustrated in the figure that appears on page 64 of Williamson et al. (2013). Two curves are shown. The lower curve is the smoothed current-day reading demand trajectory estimated at Step 1, and the upper curve is the accelerated text complexity exposure trajectory that was adopted for use in the CCSS by the CCSSO and the NGA. Dotted lines indicate score ranges for use when placing texts into grade bands (GBs). These are the interquartile ranges estimated from the collection of 487 texts described in Step 1, after applying the stretching algorithm described in Step 4, and then collapsing across GLs to define the CC grade bands (GBs) of 2-3, 4-5, 6-8, 9-10, and 11-CCR, where CCR means college and career ready.

## Alternative Approaches for Establishing an Alignment Between the Scores Generated via an Automated Text Complexity Evaluation Tool and the Common Core Text Complexity Scale

The analyses summarized above were used to define the accelerated text complexity exposure trajectory referenced in the CCSS. This section documents three approaches for establishing an alignment between the scores generated by an automated text complexity evaluation tool and that proposed trajectory. Approach 1 is the approach that was used to generate the score ranges presented in Appendix C of Nelson et al. (2012) and in the Supplement to Appendix A published by the CCSSO and NGA. Approaches 2 and 3 are alternative approaches designed to be more effective at distinguishing texts likely to scale at lower and higher levels on the CC text complexity scale. Similarities and differences in these three approaches are summarized below. Score ranges generated via each approach are evaluated in a subsequent section.

## Approach 1

Nelson et al. (2012) noted that "a common scale, based on this study and including the metrics examined here, has been published and is included as Appendix C" (p. 49). Appendix C provides a table titled "Common Scale for Band Level Text Difficulty Ranges" (p. 55). The table has five rows and six columns. Each row is labeled with one of the five CC GBs; each column is labeled with one of the six text complexity tools evaluated in the report. According to the title of the table, entries in the table are "band level text difficulty ranges" (p. 55).

One limitation of this common scale is that no information about how it was developed or validated is provided, either in the report, or in any of the publications included in the report's reference list. Because no published information is available, the information below is from M. Liben (personal communication, August 13, 2011). This communication noted that the specified score ranges were developed as follows. First, text complexity scores generated via each of six different quantitative text complexity evaluation tools were obtained for a collection of texts. Second, the Lexile score ranges presented in Williamson et al. (2013) were used to define a lower and upper endpoint for each GB. Third, an equipercentile equating program was used to align the scales generated via each of the other tools to the Lexile scale. Fourth, the estimated alignment was used to generate a corresponding set of score ranges for each of the other tools.

Several technical issues related to the above procedure should be noted. First, the approach assumes that the Lexile score ranges listed in Nelson et al. (2012) are appropriately structured for use by teachers, curriculum specialists, textbook publishers, and test developers when determining an appropriate GB placement for a text. As noted in Stenner et al. (2012), Williamson et al. (2012), and Williamson et al. (2013), however, these ranges were originally developed for *an entirely different purpose*: to quantify the levels of text complexity that students at different GLs should be exposed to in order to ensure that all students remain on track to acquire the advanced reading skills needed for success in college and careers. Note that, in this alternative estimation problem, GL classifications enter into the analyses as known constants, not as random variables. Williamson et al. (2013) correctly addressed this fundamentally different estimation problem by first grouping textbooks by their known GL classifications, next summarizing the Lexile scores obtained for the textbooks in each specified GL group, and then using those results to characterize the range of Lexile scores that students can expect to experience at each of 12 known GLs. In other words, variation in the median Lexile scores obtained for a sample of textbooks with known GL classifications was modeled as a function of those known GL classifications. When selecting texts for use in instruction and assessment, however, teachers, curriculum specialists, textbook publishers, and test developers are faced with a fundamentally different estimation problem. In this alternative problem, what is known, or can always be generated, is the Lexile score of a text, and what is not known, so must be predicted, is the particular GL to which any given text should be assigned. The solution proposed in Williamson et al. (2013) is not optimally structured for use in this alternative estimation problem because regression techniques are not symmetric. That is, a model that predicts Lexile scores as a function of a text's known GL classification (i.e., the estimation problem addressed in Williamson [2008], Williamson et al. [2012], and Williamson et al. [2013]) will not necessarily also be effective when applied to the fundamentally different problem of predicting the unknown GL of a text, conditional on its known Lexile score (i.e., the alternative estimation problem addressed by teachers, curriculum specialists, textbook publishers, and test developers when selecting texts for use in instruction and assessment).

A second technical issue concerns the finding reported in Nelson et al. (2012) that the Lexile tool was successful at distinguishing gradations of text complexity at the lowest GLs, but tended to "flatten out" at the highest GLs (p. 46). This finding suggests that the proportional allocation strategy incorporated within Approach 1 may tend to overestimate expected growth at the lowest GLs, while simultaneously underestimating expected growth at the highest GLs. In other words, score ranges generated via Approach 1 may require too much growth at the lowest GLs, and not enough growth at the highest GLs.

A third technical issue concerns the decision to characterize the expected range of variation within each CC GB via the interquartile range (Williamson et al., 2012). By definition, the interquartile range is expected to cover just 50% of the total variation. Although 50% is a relatively low coverage rate, no justification for this choice is presented.

**Table 2** Passages Selected for Use in Establishing an Alignment Between the TextEvaluator Reporting Scale and the Accelerated Text Complexity Scale Proposed in the Common Core State Standards

| Source | Target area | *N* |
|---|---|---|
| Passages selected from Common Core compliant Grade 1 textbooks[a] | Lower anchor | 50 |
| Exemplar passages from Chall, Bissex, Conrad and Harris-Sharples (1996) with adjusted GL classifications | Middle grades | 52 |
| Passages selected from textbooks targeted at students in first-year, credit-bearing college courses[b] | Upper anchor | 59 |
| Total | | 161 |

[a] Selected from Grade 1 textbooks published by Scott Foresman in 2013.

[b] Includes 19 texts selected from the T2KSWAL Corpus (Biber et al., 2004) and 40 texts provided by staff at Student Achievement Partners.

## Approach 2

This alternative approach can be summarized in terms of three steps. First, a corpus of texts with GL classifications expressed on the accelerated text complexity scale specified in the CCSS is assembled. Second, a TextEvaluator score is generated for each text and a regression analysis is conducted. In contrast to the reverse regression presented in Williamson et al. (2013), this alternative regression is designed to predict the unknown CC GL classification of a text conditional on its known TextEvaluator score. Finally, score ranges for use when placing texts within GBs are generated. Additional information about each step is summarized below.

Table 2 summarizes the collection of texts assembled for use in modeling the relationship between TextEvaluator scores and the location of a text on the accelerated text complexity scale referenced in the CCSS. The column labeled *Target area* indicates the portion of the CC scale that each group of texts was selected to address.

As is indicated in Table 2, the lower end of the scale is represented by a collection of 50 texts selected from a textbook series published by Scott Foresman in 2013 (Afflerbach et al., 2013). This series was published after the introduction of the CCSS and has been described as consistent with CC text complexity recommendations. Consequently, each text was classified as having a CC GL classification of Grade 1.

Table 2 also lists 59 texts selected to characterize variation at the upper end of the scale. These were selected from textbooks targeted at students in first-year credit-bearing college courses. Consistent with the method employed in Nelson et al. (2012), these texts were classified as having a CC GL classification of Grade 12.

Table 2 also lists a collection of 52 passages that is frequently used to illustrate the increases in text complexity that students can expect to experience as they progress from beginning reader to proficient, college graduate (Chall et al., 1996). A quantitative text complexity score assigned by Chall et al. (1996) is included for each passage. These are expressed on a numeric scale that ranges from Level 1 (suitable for students who have successfully completed first grade) to Level 16 (suitable for students who have successfully completed 4 or more years of college). Chall et al. reported that the following aspects of text variation were considered by the human experts who generated these scores:

- *Language*—This aspect was evaluated by considering the proportion of words viewed as being “unfamiliar, abstract, polysyllabic, and/or technical” (p. 16).
- *Sentence complexity*—This aspect was evaluated by considering the proportion of sentences that were “longer, more complex, less direct, with greater embedding of ideas” (p. 5).
- *Conceptual difficulty*—This aspect was evaluated by considering “the conceptual understanding required to comprehend the text, e.g.: the degree of abstractness, the amount of prior knowledge needed to understand the text” (p. 16).
- *Cognitive difficulty*—This aspect was evaluated by considering the amount of “thought, reasoning, analysis, and critical abilities [needed] to fully understand [the text]” (p. 6).

Because the aspects of text variation summarized above are closely aligned with the text complexity model outlined in Appendix A of the CCSS, a strategy of adding these passages to the alignment dataset may facilitate the goal of estimating a more stable alignment function.

The decision to include the Chall passages in the alignment dataset is also supported by the multiple validity analyses reported in Chall et al. (1996). These analyses compared the text complexity scores assigned by Chall and her colleagues

**Table 3** A Fourth-Degree Polynomial Constructed to Predict Text Grade Level as a Function of the TextEvaluator Measure of a Text

| Source | Estimated coefficient | Standard error | *t* statistic | *p*(> *t*) |
|---|---|---|---|---|
| *TE* | −1.4619 | .6139 | −2.3814 | .0185 |
| *TE*^2 | .5463 | .1440 | 3.7930 | .0002 |
| *TE*^3 | −.0411 | .0122 | −3.3563 | .0010 |
| *TE*^4 | .0019 | .0003 | 2.8026 | .0057 |

*Note*. GL = grade level, *TE*^*n* = TextEvaluator score raised to the power of *n*. The model yielded a standard error of 1.542 GLs and an $R^2$ of 0.92.

to four types of reference scores: (a) difficulty rankings provided by groups of teachers and school administrators, (b) difficulty rankings provided by students, (c) cloze comprehension scores obtained by administering the passages to groups of students, and (d) text complexity scores obtained via readability formulas. Conclusions about the validity of the proposed classifications were summarized as follows: “Validity was generally high when the qualitative reading levels were compared with student comprehension, with readability formula scores, and with independent estimates of difficulty by different judges — teachers, administrators, and students” (Chall et al., 1996, p. 81). These results provide further support for the claim that the text complexity scores provided by Chall and her colleagues for this set of 52 passages may facilitate efforts to estimate an accurate TextEvaluator/CC alignment function.

One limitation of the complexity classifications reported in Chall et al. (1996) is that the classifications are not expressed on the accelerated text complexity scale proposed in the CCSS. An approximate rescaling methodology designed to address this limitation was developed by considering differences in the GL classifications provided by Chall and her colleagues, and by the reading experts involved in the development of the CCSS, for texts with similar TextEvaluator scores. This analysis suggested that, on average, the GL classifications assigned by Chall and her colleagues were about 1.5 GLs higher than those referenced in the CCSS. Consequently, the classifications assigned by Chall and her colleagues were each decreased by 1.5 GLs. For example, all passages with a human-assigned GL classification of Grade 4 were reclassified as exhibiting a CC GL classification of Grade 2.5. Note that, while this approximate approach is consistent with the general principle specified in Sanford-Moore and Williamson (2012) that alignment methodologies should be designed to preserve current grade-by-grade growth trends to the extent possible, it is less subject to the flattening phenomenon reported in Nelson et al. (2012) because TextEvaluator scores tend to exhibit notably less flattening at the highest GLs (Nelson et al., 2012, p. 14).

Variation in the GL classifications obtained for the 161 texts summarized in Table 2 was then modeled as a function of the TextEvaluator scores generated for each text. Analyses suggested that a fourth-degree polynomial constructed to predict the CC GL classification of a text conditional on its known TextEvaluator score provided an adequate fit to the data. The model yielded an estimated standard error of $s = 1.542$ GLs and a multiple $R^2$ of 0.92.[2] Estimated coefficients are listed in Table 3. Note that all of the coefficients are significant at an alpha level of 0.05.

When interpreting the magnitude of the coefficients in Table 3, one should note that raw scores obtained via the TextEvaluator engine are expressed on a GL scale that predates the accelerated text complexity scale proposed in the CCSS. In particular, because all of the GL classifications in the training portion of the TextEvaluator corpus were collected prior to the introduction of the CCSS, raw scores generated via the TextEvaluator engine represent pre-CC GL expectations, not post-CC GL expectations. To avoid confusion resulting from differing GL expectations, raw scores generated via the TextEvaluator engine are reexpressed on an alternative numeric scale prior to reporting. This is accomplished by multiplying each raw score by 100, thereby creating a reporting scale that ranges from 0 to 2000. This new scale appropriately addresses the problem of changing GL expectations because users are unlikely to assign spontaneously an incorrect GL interpretation to a scale that does not look like a GL scale. Note that a similar rescaling option is employed in Appendix C of Nelson et al. (2012). That is, changing GL expectations are accommodated not by changing the score assigned to a text, but rather, by changing the score ranges provided in a corresponding table of realigned score ranges.

Predicted scores obtained via the polynomial regression model estimated above, after reexpression on the TextEvaluator reporting scale metric, are plotted in Figure 1. The dashed lines provide an approximate confidence interval about the estimated regression curve. These were obtained by first determining the 25th quantile of a *t* distribution with 161 − 5 = 256 degrees of freedom, then multiplying the estimated standard error of prediction, *s*, by that value, and then defining a

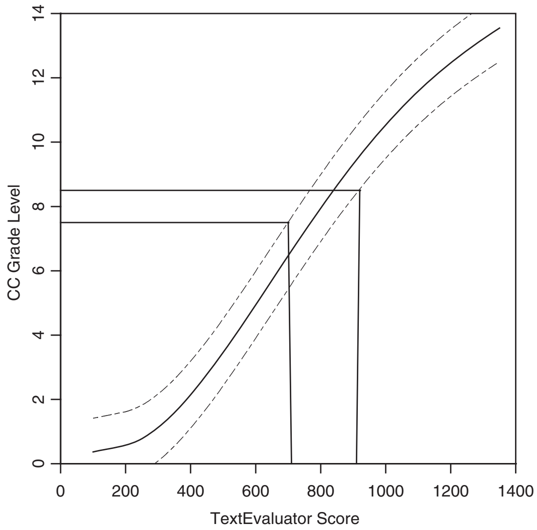

**Figure 1** A fourth-degree polynomial constructed to predict Common Core (CC) grade level (GL) classifications as a function of TextEvaluator scores with approximate confidence bands plotted at $\pm t_{(.25,256)}s$, where $t_{(.25,256)}$ is the 25th quantile of a t distribution with 256 degrees of freedom, and $s$ is the estimated standard error of prediction. Vertical lines show the range of TextEvaluator scores that are expected to yield a CC GL prediction that rounds to Grade 8.

($\text{Lo}_i$, $\text{Hi}_i$) interval about each predicted CC score, $\hat{y}_\imath$, as follows:

$$\text{Lo}_i = \hat{y}_\imath - t_{(0.25,\ 256)} \times s = \hat{y}_\imath - 0.6754 \times 1.542 = \hat{y}_\imath - 1.0414,$$
$$\text{Hi}_i = \hat{y}_\imath + t_{(0.25,\ 256)} \times s = \hat{y}_\imath + 0.6754 \times 1.542 = \hat{y}_\imath + 1.0414. \quad (4)$$

Like the interquartile ranges reported in Stenner et al. (2012), Williamson (2008), and Williamson et al. (2013), these intervals are expected to contain the population value $y_i$ about 50% of the time in repeated samples.

Although the model summarized in Table 3 and Figure 1 is designed to predict the unknown CC GL classification of a text conditional on its known TextEvaluator score, the model can also be adapted for use in other types of prediction problems. For example, the vertical lines in Figure 1 show the range of TextEvaluator scores that are likely to yield a predicted CC GL classification that rounds to Grade 8. Because score ranges like the one shown in Figure 1 were specifically requested by staff at Student Achievement Partners, this ad hoc procedure was repeated for each of the grades in the range from Grade 2 to Grade 12, and the resulting score ranges were then collapsed to provide a set of TextEvaluator score ranges for use when placing texts into GBs. Because theoretically based claims about the performance of the resulting score ranges are not possible, multiple empirical evaluations were conducted. Key results are summarized in a subsequent section.

## Approach 3

The approximate confidence intervals described above are designed to include the targeted population parameter 50% of the time in repeated samples. Because this low coverage rate may not be appropriate for use in some applications, score ranges based on a slightly larger confidence probability were also estimated. This was accomplished by defining a slightly larger ($\text{Lo}_i$, $\text{Hi}_i$) interval about each predicted CC score, $\hat{y}_\imath$, as follows:

$$\text{Lo}_i = \hat{y}_\imath - t_{(0.20,\ 256)} \times s = -0.8430 \times 1.542 = \hat{y}_\imath - 1.299,$$
$$\text{Hi}_i = \hat{y}_\imath + t_{(0.20,\ 256)} \times s = \hat{y}_\imath + 0.8430 \times 1.542 = \hat{y}_\imath + 1.299. \quad (5)$$

where $t_{(0.20,\ 256)}$ represents the value on the $t$ distribution with 256 degrees of freedom that yields a cumulative probability of .20. Score ranges estimated via this alternative approach are expected to include the population parameter 60% of the time in repeated sampling.

**Table 4** Score Ranges Developed via Each of Three Different Approaches for the Five Grade Bands Defined in the Common Core State Standards

|  | Approach 1 |  | Approach 2 | Approach 3 |
|---|---|---|---|---|
| GB | Lexile | Flesch – Kincaid | TextEvaluator | TextEvaluator |
| 2 – 3 | 420 – 820 | 1.98 – 5.34 | 145 – 575 | 100 – 590 |
| 4 – 5 | 740 – 1010 | 4.51 – 7.73 | 425 – 705 | 405 – 720 |
| 6 – 8 | 925 – 1185 | 6.51 – 10.34 | 570 – 920 | 550 – 940 |
| 9 – 10 | 1050 – 1335 | 8.32 – 12.12 | 765 – 1095 | 750 – 1125 |
| 11 – CCR | 1185 – 1385 | 10.34 – 14.20 | 910 – 1350 | 890 – 1360 |

*Note*. GB = grade band, CCR = college and career ready. The score ranges listed for the Lexile tool and the Flesch – Kincaid tool are reprinted from Appendix C of Nelson et al. (2012).

## Evaluation

Score ranges generated via the three approaches documented above are listed in Table 4. Note that two sets of score ranges are provided for Approach 1: one defined in terms of Lexile scores and one defined in terms of Flesch – Kincaid scores. Both sets are reprinted from the table provided in Appendix C of Nelson et al. (2012). Table 4 also lists score ranges implemented via Approaches 2 and 3. Note that, while these additional score ranges are provided in a form that is similar to the form presented in Appendix C of Nelson et al. (2012), these ranges were actually generated via the fundamentally different estimation approach summarized above.

Three evaluations of the proposed score ranges were conducted. Each evaluation was implemented with respect to one of the following datasets: (a) the set of 168 exemplar texts provided in Appendix B of the CCSS, (b) a set of 10 texts selected to represent the reading ability needed to comprehend workplace and citizenship documents, and (c) a set of 59 texts selected from textbooks assigned in first-year credit-bearing college courses. Note that the first two datasets are composed of passages that were *not* included in the dataset used to estimate the proposed TextEvaluator/CC alignment function. Thus, these datasets provide an independent evaluation of the degree of confidence that can be placed in the proposed score ranges.

### Evaluation 1: Analysis of 168 CC Exemplar Texts

Appendix B of the CCSS presents a collection of 168 exemplar texts selected to illustrate the text complexity variation expected within each of the five GBs defined in the standards (i.e., Grades 2 – 3, 4 – 5, 6 – 8, 9 – 10, and 11 – CR). Nelson et al. (2012) described the selection and analysis of these texts as follows:

> A working group was assembled from among the constituencies guiding the writing of the Common Core Standards. . . . . These contributors were asked to recommend texts that they or their colleagues had used successfully with students in a given grade band and to justify and describe that use. Reviewing the recommendations and assembling the final collection was done using the following considerations:
>
> Complexity: Following the recommendations set forth in Appendix A of the CCSS, a three-part model for measuring complexity was used. The three parts were qualitative indices of inherent text complexity judged by human raters, quantitative measures using Lexiles and Coh-Metrix features of Easability, and professional (educator) judgment for matching texts to an appropriate band level. Final selection was made by the working group and vetted broadly during the Standards vetting process. (pp. 17 – 18)

The agreement between these independently assigned GB classifications and those obtained via application of the proposed score ranges are summarized in Table 5. Looking first at the agreement rates listed for Approach 1, note that both the Lexile tool and the Flesch – Kincaid tool yielded agreement rates below 50% at each of the top three GBs. By contrast, the corresponding agreement rates achieved via Approach 2 were noticeably higher at 63%, 62%, and 79%, and those achieved via Approach 3 were even higher at 68%, 64%, and 82%.

The trend toward higher agreement under Approaches 2 and 3 is also reflected in the overall agreement rates. These were 45% and 48% under Approach 1, and 64% and 69% under Approaches 2 and 3. Methodological issues that may account for the improved classification performance achieved using Approaches 2 and 3 are discussed below.

**Table 5** Number and Percentage of Common Core Exemplar Texts Classified as Appropriate for Students in the Grade Band Recommended in the Common Cores State Standards, by Approach and Text Complexity Tool

| | Approach 1 | | | | Approach 2 | | Approach 3 | |
|---|---|---|---|---|---|---|---|---|
| | Lexile | | Flesch – Kincaid | | TextEvaluator | | TextEvaluator | |
| GB | *N* | % | *N* | % | *N* | % | *N* | % |
| 2 – 3 | 14 | 70 | 16 | 80 | 13 | 65 | 17 | 85 |
| 4 – 5 | 15 | 50 | 15 | 50 | 14 | 47 | 15 | 50 |
| 6 – 8 | 18 | 44 | 17 | 41 | 26 | 63 | 28 | 68 |
| 9 – 10 | 17 | 44 | 17 | 44 | 24 | 62 | 25 | 64 |
| 11 – CCR | 12 | 32 | 16 | 42 | 30 | 79 | 31 | 82 |
| Total | 76 | 45% | 81 | 48% | 107 | 64% | 116 | 69% |

*Note*. Lexile scores were obtained via the Lexile Analyzer (available at http://www.lexile.com). The analyzer did not provide scores for seven texts that contained more than 1,000 words. Flesch – Kincaid scores were generated via an in-house program that yields scores that are nearly identical to the Flesch – Kincaid scores generated via the Microsoft Word program. TextEvaluator scores were generated using TextEvaluator 5.0. Results for other tools are not included in this summary because scores generated via other tools were not available for this dataset. GB = grade band, CCR = college and career ready.

**Table 6** Number and Percentage of Career Texts Classified as Appropriate for Students in the 11 – CCR Grade Band, by Approach and Text Complexity Tool

| | Approach 1 | | Approach 2 | | Approach 3 | |
|---|---|---|---|---|---|---|
| | Flesch – Kincaid | | TextEvaluator | | TextEvaluator | |
| Category | *N* | % | *N* | % | *N* | % |
| Below 11 – CCR | 7 | 70 | 3 | 30 | 3 | 30 |
| Within 11 – CCR | 3 | 30 | 7 | 70 | 7 | 70 |
| Above 11 – CCR | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 10 | 100% | 10 | 100% | 10 | 100% |

*Note*. Flesch – Kincaid scores were generated via an in-house program that yields scores that are nearly identical to the Flesch – Kincaid scores generated via the Microsoft Word program. TextEvaluator scores were generated via TextEvaluator 5.0. Results for other tools are not included in this summary because scores generated via other tools were not available for this dataset. GB = grade band, CCR = college and career readiness.

## Evaluation 2: Classification Results for Career Texts

A collection of 10 texts was available to represent the reading ability needed to comprehend workplace and citizenship documents.[3] Workplace documents included a selection from an army field manual, a classroom management guide for teachers in community arts programs, an article about new licensing requirements for cosmetologists in the State of Florida, a guide for emergency medical technicians, and a guide for addressing insect infestations in apple trees in Minnesota. Citizenship documents included a study guide for the U.S. Naturalization Exam, a handbook for trial jurors published by the U.S. District Court of Kansas, instructions for obtaining a marriage license in North Carolina, instructions for obtaining a driver’s license in Illinois, and an article about hybrid cars.

The CCSS note that all U.S. students should be able to comprehend career texts by the end of Grade 12. Consequently, the collection of 10 career texts summarized above was evaluated relative to the score ranges estimated for the 11 – CCR GB.

Results are summarized in Table 6. Because Lexile scores were not available for these texts, Approach 1 is evaluated using Flesch – Kincaid scores only. Note that all three approaches placed all of the career texts at or below the 11 – CCR GB. This is consistent with findings previously reported in Williamson (2008) and Stenner et al. (2012), which demonstrated that workplace and citizenship texts tend to be less complex than college texts.

## Evaluation 3: Classification Results for Texts Assigned in First-Year Credit-Bearing College Courses

This evaluation differs from the two evaluations reported above in that an independent set of college texts was not available so the evaluation was conducted with respect to the same set of 59 college texts used to estimate the TextEvaluator/CC

**Table 7** Number and Percentage of College Texts Classified as Appropriate for Students in the 11 – CCR Grade Band, by Approach and Text Complexity Tool

| | Approach 1 | | Approach 2 | | Approach 3 | |
|---|---|---|---|---|---|---|
| | Flesch – Kincaid | | TextEvaluator | | TextEvaluator | |
| Category | *N* | % | *N* | % | *N* | % |
| Below 11 – CCR | 10 | 17 | 3 | 5 | 3 | 5 |
| Within 11 – CCR | 47 | 80 | 52 | 88 | 53 | 90 |
| Above 11 – CCR | 2 | 3 | 4 | 7 | 3 | 5 |
| Total | 59 | 100% | 59 | 100% | 59 | 100% |

*Note*. Flesch – Kincaid scores were generated via an in-house program that yields scores that are nearly identical to the Flesch – Kincaid scores generated via the Microsoft Word program. TextEvaluator scores were generated via TextEvaluator 5.0. Results for other tools are not included in this summary because scores generated via other tools were not available for this dataset. CCR = college and career readiness.

alignment function. Results are summarized in Table 7. Note that the proposed TextEvaluator score ranges were successful at classifying 88% and 90% of these texts, while the success rate achieved by the Flesh – Kincaid tool was slightly lower at 80%. Because Lexile scores and Flesch – Kincaid scores are highly correlated and because the Flesch – Kincaid score ranges were developed via the equipercentile equating described above, it is likely that the classification rate for the Lexile tool is also near 80%.

It is useful to consider similarities and differences between the agreement rates reported for the 11 – CCR GB in Table 5, and those reported for the college texts in Table 7. For both Approaches 2 and 3, agreement rates are slightly lower for texts in the 11 – CCR GB than for college texts. For example, under Approach 2, the agreement rate for college texts is 88%, while that for texts in the 11 – CCR GB is slightly lower at 79%. Similarly, under Approach 3, the agreement rate for college texts is 90%, while that for texts in the 11 – CCR GB is slightly lower at 82%. While the agreement rates estimated for Approach 1 also show a decline, the decline is much larger (e.g., the Flesch – Kincaid score yielded an agreement rate of 80% for college texts, but just 42% for texts in the 11 – CCR GB). Although Lexile scores were not available for the college texts, the Lexile/CC agreement rate for texts in the 11-CCR GB was even lower at just 32%. What may account for these large decreases? One possibility is that the higher agreement rates achieved by all three approaches when scoring the college texts is due to the fact that all of the college texts belong to the informational genre, while a large proportion of the CC exemplars are either literary or mixed.[4] As has been reported in a large number of previous studies, both the Lexile tool and the Flesch – Kincaid tool are much more precise when scoring informational texts and much less precise when scoring literary and mixed texts (CCSSI, 2010; Hiebert & Mesmer, 2013a, 2013b; Nelson et al., 2012; Sheehan, 2014a, 2014b; Sheehan, Kostin, Futagi, & Flor 2010, Sheehan et al., 2014). Because the TextEvaluator tool includes distinct prediction models for informational, literary, and mixed texts, however, classification accuracy remains relatively high, even when scoring literary and mixed texts. This difference may explain why the Lexile and Flesch – Kincaid tools were unable to achieve a correct classification rate that exceeded 50% when scoring the CC exemplar texts at each of the top four CC GBs.

The correlation coefficients reported in several previous studies support this explanation. For example, after analyzing the set of 168 exemplar texts in Appendix B of the CCSS, Nelson et al. (2012) reported that the correlation between Lexile scores and scores assigned by human experts was just .50. By contrast, the same correlation calculated with TextEvaluator scores instead of Lexile scores is .72 (Sheehan, 2014a). These results suggest that a strategy of attending to genre effects when scoring informational, literary, and mixed texts can lead to score predictions that are more appropriate for use when placing texts into GBs.

## Summary, Recommendations, and Discussion

This paper documented three approaches for establishing an alignment between the scores generated by an automated text complexity evaluation tool and a proposed text complexity scale. Score ranges developed via each approach were evaluated. In each case, analyses were designed to characterize the agreement between text complexity scores generated via a proposed tool and text complexity classifications intended to be reflective of the accelerated text complexity exposure trajectory specified in the CCSS. Implications with respect to the goal of determining an optimal GB classification for a text with a known text complexity score are summarized below.

### Selecting a Dependent Variable for Use When Constructing an Alignment Function

John Tukey famously argued that “[it is] far better [to provide] an approximate answer to the right question, which is often vague, than an exact answer to the wrong question which can always be made more precise” (Tukey, 1962, p. 13). When aligning scores generated via an automated text complexity tool to any proposed text complexity scale, the right question is “Which GL is most likely given the complexity score estimated for this text?” The polynomial regression model introduced in this paper was specifically constructed to answer this question. By contrast, the polynomial regression model presented in Williamson et al. (2013), and the corresponding score ranges provided in Appendix C of Nelson et al. (2012) and in CCSSO and NGA (2014), were designed to answer a fundamentally different question, “What is the range of Lexile scores that we should expect to observe among texts at each of 12 known GLs?” The greater success of the alternative approaches introduced in this paper suggests that a strategy of focusing on the right question can lead to estimated score ranges that are more appropriate for use when determining an optimal GL or GB placement for a text with a known text complexity score.

### Accounting for Genre Effects

Researchers have frequently argued that many important indicators of comprehension difficulty tend to function differently within informational, literary, and mixed texts. For example, the authors of the CCSS argued as follows “The Lexile Framework, like traditional formulas, may underestimate the difficulty of texts that use simple, familiar language to convey sophisticated ideas, as is true of much high-quality fiction written for adults and appropriate for older students” (CCSSI, 2010, Appendix A, p. 7). Hiebert and Mesmer (2013a) provided a more detailed explanation: “Content area texts often receive inflated readability scores because key concept words that are rare (e.g., *photosynthesis*, *inflation*) are often repeated which increases vocabulary load, even though repetition of content words can support student learning (Cohen & Steinberg, 1983). … [By contrast] Readabilities of narrative texts are especially prone to deflation due to the presence of dialogue that typically consists of short sentences” (p. 46).

Analyses summarized above suggested that the poor performances of both the Lexile tool and the Flesch–Kincaid tool when attempting to place the CC exemplar texts into the GBs assigned by human experts may be due to the fact that neither tool is designed to address the unique processing challenges of literary and mixed texts. In particular, in each case, a single prediction model is assumed to hold equally well for informational, literary, and mixed texts. Analyses summarized above suggest that this assumption can be evaluated by comparing the correct classification rates achieved for the college texts, which are all informational, to those achieved for the CC exemplar texts in the 11–CCR GB, which include a large number of literary and mixed texts. This comparison strongly supports the genre differences reported in previous research. In particular, the Lexile and Flesch–Kincaid tools achieved correct classification rates at or near 80% when scoring the college texts, which are all informational, yet just 32% and 42% when scoring the exemplar texts in the 11–CCR GB, which include a large number of literary and mixed texts, even though both sets of texts were evaluated with respect to the exact same score ranges (i.e., 1185–1385 for the Lexile tool, and 10.34–14.20 for the Flesch–Kincaid tool).

Note that a similarly dramatic drop is not present when the correct classification rates achieved by the TextEvaluator tool are compared. In particular, the TextEvaluator tool yielded correct classification rates of 88% and 90% when scoring the college texts, and 79% and 82% when scoring the CC exemplar texts in the 11–CCR GB. These results support the claim that the TextEvaluator strategy of estimating distinct prediction models for informational, literary, and mixed texts has succeeded in addressing at least some of the unique processing demands of literary and mixed texts, and thus, may be more appropriate for use when classifying texts that are similar to the exemplar texts presented in Appendix B of the CCSS (CCSSI, 2010).

### Selecting an Optimal Alignment Approach

All of the approaches discussed in this study performed better at some GBs and worse at others. For example, the score ranges estimated via Approach 1 yielded fairly high agreement at the 2–3 GB, yet exhibited much lower agreement (at or below 50%) at each of the other GBs. By contrast, the score ranges estimated via Approaches 2 and 3 were consistently higher, with only one GB yielding an agreement rate that failed to surpass 50%. The greater success of the score ranges estimated via Approaches 2 and 3 is also evident when the results from all GBs are combined: Approach 1 yielded overall agreement rates of 45% and 48%, while those achieved under Approaches 2 and 3 were noticeably higher at 64% and 69%, respectively. Because the score ranges estimated via Approach 3 performed much better than those estimated via Approach 1 and slightly better than those estimated via Approach 2, the ranges estimated via Approach 3 are most appropriate for use by teachers, curriculum specialists, textbook publishers, and text developers when placing texts into GBs. As is recommended in Appendix A of the CCSS (CCSSI, 2010), however, both quantitative and qualitative analyses should be considered when making final placement decisions.

### Can TextEvaluator Scores Help One Distinguish College-level Texts?

A key purpose of the accelerated text complexity exposure trajectory specified in the CCSS is to ensure that all HS graduates have been exposed to the types of complex texts they will likely encounter in college and careers. This study demonstrated that score ranges defined in terms of TextEvaluator scores can be used with confidence when selecting texts for use at the 11 – CCR GB. For example, analyses of the exemplar texts from Appendix B of the CCSS (CCSSI, 2010) yielded agreement rates of 79% and 82% at the 11 – CCR GB, and analyses of a large collection of college texts yielded agreement rates of 88% and 90%. These high agreement rates suggest that text complexity evidence collected via TextEvaluator can help users distinguish the types of complex texts that students can expect to encounter in both college and careers.

### Can TextEvaluator Scores be Used With Confidence When Selecting Texts for Use in Instruction and Assessment?

Analyses reported in Nelson et al. (2012) have been cited as supporting the claim that text complexity scores generated via each of six different text complexity evaluation tools are appropriately structured to "guide curriculum decisions, assist assessment development, and support the efforts of educational publishers to meet complexity guidelines" (Nelson et al., 2012, p. 3). Analyses summarized above suggest that TextEvaluator scores may be even more effective at providing feedback for use in these types of activities. Therefore, teachers and other educators are encouraged to use the TextEvaluator score ranges estimated via Approach 3 when selecting texts for use in instruction and assessment.

### Selecting a High School Graduation Target

All of the score ranges evaluated in this paper are designed to culminate at the HS graduation target referenced in the CCSS. As was demonstrated in Stenner et al. (2012), however, this target is not closely focused on the reading ability needed to comprehend college and workplace texts, as has frequently been reported. Rather, it appears to be more closely focused on the reading ability needed to comprehend Wikipedia articles and articles from international newspapers. Because evidence that confirms the unusually high Lexile scores obtained for these alternative types of texts has not yet been provided, the decision to expand the definition of a postsecondary text to include these alternative genres should be reexamined. Methodological issues that should be considered when conducting these additional investigations are outlined in the appendix.

### Reexamining the Decision to Allocate 25% of Total Estimated Growth to the Interval Between Grades 1 and 2

Nelson et al. (2012) reported that the Lexile tool was successful at distinguishing gradations of text complexity at the lowest GLs, but tended to "flatten out" at the highest GLs (p. 46). This finding suggests that the conclusion reported in Sanford-Moore and Williamson (2012) that reading growth in the interval between Grades 1 and 2 is five times greater than reading growth in the interval between Grades 10 and 11 may not be accurate, and that the CC strategy of requiring students at the lowest GB to grow much faster than students at the highest GB may not be appropriate. Thus, additional research focused on the goal of understanding grade-to-grade trends in reading growth is needed. Growth curves estimated via alternative text complexity evaluation tools should be included as part of this additional research.

### Alternative Approaches for Communicating Alignment Information to Score Users

The alignment table provided in Appendix C of Nelson et al. (2012) and in the supplement to Appendix A provided by the CCSSO and NGA (2014) specified the expected range of text complexity scores: ($x_{Lo}$, $x_{Hi}$), at each of five CC GBs. Because other approaches for communicating alignment information to score users may also facilitate the goal of helping users place texts into GBs, alternative approaches for communicating alignment information to score users should also be investigated.

## Acknowledgments

I am grateful to Neil Dorans and Chaitanya Ramineni for many helpful discussions regarding the alignment methodology reported in this paper.

## Notes

[1] The reported $R^2$ of .99 is not particularly informative because five predictors are used to model variation in just 12 data points. According to one rule of thumb, the $R^2$ statistic only provides useful information about the predictive power of an estimated regression equation when the number of observed data points is at least $(50 + 8 \times p)$ where $p$ is the number of parameters that must be estimated (Tabachnick & Fidell, 2001). Because the regression equation presented in Williamson et al. (2013) includes five estimated parameters and because the number of available data points is far less than $50 + 8 \times 5 = 90$, the reported $R^2$ statistic of 0.99 is not a valid indicator of predictive power and should not have been reported.

[2] According to the rule of thumb provided in Tabachnick and Fidell (2001), the reported $R^2$ of 0.92 is a useful measure of the predictive power of the estimated regression equation because the number of observations included in the analysis, $n = 161$ is greater than $50 + 8 \times 4 = 82$, where 4 is the number of predictors.

[3] These texts were provided by staff at Student Achievement Partners. Additional analyses of these texts are reported in Nelson et al. (2012).

[4] A mixed text is a text that contains a mixture of informational and literary elements. Many of the exemplar texts in Appendix B of the CCSS belong to the mixed genre because they were originally written both to provide information and to address important literary goals.

[5] This appendix is adapted from Sheehan (2014b), a paper which was awarded the Lorne H. Woollatt Distinguished Paper Award by the Northeastern Educational Research Association (NERA).

[6] Two medians are estimated from the chart presented in Stenner et al. (2012). Estimates are used because actual values were not reported.

## References

Afflerbach, P., Blachowicz, C. L. Z., Boyd, C. D., Izquierdo, E., Juel, C., Kame'enui, E., … Wixson, K. K. (2013). *Reading street (Common Core)*. Glenview, IL: Scott Foresman.

Biber, D., Conrad, S. M., Reppen, R., Byrd, P., Helt, M., Clark, V., … Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph Series No. 25). Princeton, NJ: Educational Testing Service.

Chall, J. S., Bissex, G. L., Conrad, S. S., & Harris-Sharples, S. (1996). *Qualitative assessment of text difficulty: A practical guide for teachers and writers*. Cambridge, MA: Brookline Books.

Common Core State Standards Initiative. (2010, June). *Common Core State Standards for English language arts & literacy in history/social studies, science and technical subjects*. Washington, DC: Council of Chief State School Officers & National Governors Association.

Council of Chief State School Officers, & National Governors Association. (2014). *Supplemental information for Appendix A of the Common Core State Standards for English language arts and literacy: New research on text complexity*. Retrieved from http://www.corestandards.org/wp-content/uploads/Appendix-A-New-Research-on-Text-Complexity.pdf

Dagget, W. R. (2003). *Achieving reading proficiency for all*. Rexford, NY: International Center for Leadership in Education.

Hiebert, E. H. (2012). The Common Core State Standards and text complexity. *Teacher Librarian*, *39*(5), 13–19.

Hiebert, E. H., & Mesmer, H. A. (2013a). Meeting standard 10: Reading complex text. *Principal Leadership*, *13*(5), 30–33.

Hiebert, E. H., & Mesmer, H. A. (2013b). Upping the ante of text complexity in the Common Core State Standards: Examining its potential impact on young readers. *Educational Researcher*, *42*(1), 44–51.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York, NY: Student Achievement Partners.

Sanford-Moore, E. E. (2013). *The Lexile framework and myON reader* (MetaMetrics Whitepaper). Durham, NC: MetaMetrics, Inc.

Sanford-Moore, E. E., & Williamson, G. L. (2012). *Bending the text complexity curve to close the gap* (MetaMetrics Research Brief). Durham, NC: MetaMetrics, Inc.

Sheehan, K. M. (2014a). *An analysis of key claims in the TextEvaluator validity argument* (Unpublished manuscript), Educational Testing Service, Princeton, NJ.

Sheehan, K. M. (2014b, October). *What proportion of the high school/college text complexity gap is due to genre DIF?* Paper presented at the Northeastern Educational Research Association (NERA), Trumbull, CT.

Sheehan, K. M., Flor, M., & Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In *Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA), Annual Conference of the Association for Computational Linguistics* (pp. 49–58). Stroudsburg, PA: Association for Computational Linguistics.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2013). *U.S. Patent No. 8,517,738*. Washington, DC: U.S. Patent and Trademark Office.

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards* (Research Report No. RR-10-28). Princeton, NJ: Educational Testing Service.

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. *Elementary School Journal*, *115*(2), 184–209.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). *The Lexile framework for reading technical report*. Durham, NC: MetaMetrics, Inc.

Stenner, A. J., Koons, H. H., & Swartz, C. W. (2009). *Re-conceptualizing the text complexity demand curve and using technology to promote growth towards college and career readiness*. Durham, NC: MetaMetrics, Inc.

Stenner, A. J., Sanford-Moore, E. E., & Williamson, G. L. (2012). *The Lexile framework for reading quantifies the ability needed for "college & career readiness"* (Metametrics Research Brief). Durham, NC: Metametrics, Inc.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Tukey, J. W. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*, 1–67.

Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics*, *19*, 602–632.

Williamson, G. L., Fitzgerald, J., & Stenner, A. J. (2013). The Common Core State Standards' quantitative text complexity trajectory: Figuring out how much complexity is enough. *Educational Researcher*, *42*(2), 59–69.

Williamson, G. L., Koons, H., Sandvik, T., & Sanford-Moore, E. (2012). *The text complexity continuum in Grades 1–12* (Metametrics Research Brief). Durham, NC: MetaMetrics, Inc.

## Appendix

## Accounting for Genre-Based Differential Feature Functioning When Estimating the High School/College Text Complexity Gap[5]

A key assumption underlying the methodology used to estimate the HS/college text complexity gap referenced in the CCSS is that text complexity scores generated via the Lexile tool are unbiased with respect to genre. This appendix introduces a revised gap estimation methodology designed to be more robust against violations of this assumption. The revised methodology is similar to the methodology employed in existing research except for two relatively minor refinements: (a) peripheral genres such as Wikipedia articles and articles from international newspapers are not included in the sample of texts used to represent the reading demands of college and workplace texts and (b) a correction for genre bias is included when quantifying differences between HS and college texts. Additional information about each refinement is summarized below. A subsequent section examines the impact of each proposed revision on the estimated magnitude of the gap.

### Refinement 1: Exclude Peripheral Genres When Estimating the Reading Demands of College and Workplace Texts

Williamson (2008) argued that all HS graduates should be able to comprehend four types of documents: citizenship documents, workplace documents, military documents, and college textbooks. In a subsequent paper, Stenner et al. (2012) argued that this set should be expanded to also include two additional text genres: Wikipedia articles and articles extracted from international newspapers. Table A1 presents median Lexile scores for the four types of postsecondary texts proposed in Williamson (2008), and for the additional types of postsecondary texts proposed in Stenner et al. (2012). Note that,

**Table A1** Median Lexile Scores for the Different Types of Postsecondary Texts Analyzed in Williamson (2008) and in Stenner et al. (2012)

| Text collection | No. of texts | Median Lexile score | Distance from combined median of 1300 |
|---|---|---|---|
| Military documents | 22 | 1180[b] | −120 |
| Citizenship documents | 54 | 1230[b] | −70 |
| University textbooks—GA | *nr* | 1220[a] | −80 |
| University textbooks—TX | *nr* | 1230[a] | −70 |
| University textbooks—TN | *nr* | 1260[a] | −40 |
| Workplace texts | 1,401 | 1260[b] | −40 |
| University textbooks | 294 | <1300[*est*] | |
| Combined sample | 2,990 | 1300[c] | 0 |
| *Chicago Tribune* | *nr* | 1310[d] | +10 |
| *Wall Street Journal* | *nr* | 1320[d] | +20 |
| *Los Angeles Times* | *nr* | 1330[d] | +30 |
| *Washington Post* | *nr* | 1350[d] | +50 |
| *New York Times* | *nr* | 1380[d] | +80 |
| Wikipedia articles | 945 | >1300[*est*] | |

*Note*. *nr* = not reported. *est* = estimated from the chart provided in Stenner et al. (2012).
[a] Reported in Sanford-Moore (2013).
[b] Reported in Williamson (2008).
[c] Reported in Stenner et al. (2012).
[d] Reported in Dagget (2003).

while each of the four types of texts discussed in Williamson (2008) yielded median Lexile scores that were less than 1300, the additional types of postsecondary texts introduced in Stenner et al. (2012) each yielded median Lexile scores that were greater than 1300.[6] For example, the median Lexile score of articles extracted from the *Wall Street Journal* is more than 100 points greater than the median Lexile score of textbooks read by students attending universities in Georgia. Similarly, the median Lexile score of articles extracted from the *New York Times* is more than 100 points greater than the median Lexile score of textbooks read by students attending universities in Tennessee. These differences suggest that, in many cases, students who have no problems comprehending their college textbooks are, nevertheless, unable to comprehend successfully articles extracted from international newspapers and from Wikipedia. What may account for these unexpected findings? One possibility is that the Lexile prediction equation may overestimate the complexity levels of Wikipedia articles and articles extracted from international newspapers because these alternative types of texts may include names of persons, places, and things that were not well-represented within the sample of texts considered when the Lexile word frequency (WF) index was estimated. For example, familiar terms such as *software* and *digital* occur frequently in Wikipedia, yet are rarely found in the Touchstone Applied Science Associates (TASA) corpus, a corpus that, like the Lexile corpus, includes a large number of texts selected to represent materials read by students in elementary and secondary classrooms. Conversely, everyday words like *Mom* and *Dad* occur with high frequency in the TASA corpus, yet are more than 10 times less frequent in Wikipedia. As these few examples illustrate, WF indices estimated from the Lexile corpus may not provide valid information about the familiarity of words in Wikipedia articles or in articles extracted from international newspapers.

In addition to presenting median Lexile scores for the various types of college and workplace texts discussed in Williamson (2008) and Stenner et al. (2012), Table A1 also shows how each median compares to the proposed Common Core Grade 12 target of 1300 Lexile points. Note that all four of the text types discussed in Williamson (2008) yielded median Lexile scores that fell *below* this target, while only the additional genres introduced in Stenner et al. (2012) yielded median Lexile scores that fell *above* this target. This suggests that the CCR target defined in Stenner et al. (2012) is higher than it would have been if the postsecondary sample had not been expanded to include additional genres such as Wikipedia articles and articles from international newspapers. Because the additional research needed to validate the unusually high Lexile scores obtained for these newer types of texts has not yet been reported, it seems prudent to limit the analyses to the four types of postsecondary documents proposed in Williamson (2008), that is, military documents, citizenship documents, university textbooks, and workplace texts. Note that, even after implementing this limitation, the postsecondary sample described in Stenner et al. (2012) would still be quite large (i.e., the sample would then include a total of 2,013 college and career texts).

### Refinement 2. Include an Adjustment for Genre Bias

In many educational contexts, inferences about students' knowledge, skills, and accomplishments are based on evidence extracted from observed item responses. Although each new item is designed to provide unbiased evidence about examinee standing relative to the targeted proficiency construct, item pretest statistics are routinely evaluated in order to ensure that the items accepted for use on operational assessments do not incorporate unintended biases. In some cases, however, analyses of item pretest statistics confirm that one or more items exhibit differential item functioning or DIF. DIF occurs when test takers with similar levels of a measured attribute or trait tend to score lower or higher on an item depending on the particular subpopulation they belong to (Holland & Wainer, 1993). When items with significant levels of DIF are included on an assessment, subsequent mean scores may indicate an achievement gap among test takers in some subpopulations (e.g., male examinees vs. female examinees) even when no gap is actually present.

Just as the evidence provided by a proposed test item may be biased in favor of examinees in some subgroups (e.g., male examinees), the evidence provided by a proposed text complexity feature may be biased in favor of texts in some subgroups (e.g., informational texts such as those included in college textbooks). We refer to this phenomenon as *differential feature functioning* or DFF. That such biases are possible has been noted in a number of recent publications. For example, the authors of the CCSS argued as follows: "The Lexile Framework, like traditional formulas, may underestimate the difficulty of texts that use simple, familiar language to convey sophisticated ideas, as is true of much high-quality fiction written for adults and appropriate for older students" (CCSSI, 2010, Appendix A, p. 7).

The finding that traditional readability formulas tend to underestimate the complexity levels of literary texts is also reported in Hiebert (2012); Hiebert and Mesmer (2013a, 2013b); Nelson et al. (2012); Sheehan, Flor, and Napolitano (2013); Sheehan et al. (2010); and Sheehan et al. (2014).

Hiebert (2012) explained this phenomenon as follows:

> First, when rare words are repeated—as they often are in informational texts where precise vocabulary is used (e.g., photosynthesis, refraction)—the level of a text is frequently overestimated. Second, when texts contain large amounts of dialogue as is often the case with narrative texts, text levels are frequently underestimated since people typically speak in short sentences. (p. 13)

A refinement designed to address the genre-specific estimation errors documented in the above studies has been developed. The proposed refinement is implemented in four steps as follows. First, two samples of passages are assembled: one composed entirely of informational passages and one composed entirely of literary passages. Second, both a Lexile score and a GL classification provided by a human expert are obtained for each passage and two text complexity exposure trajectories are defined: one estimated entirely from informational passages and one estimated entirely from literary passages. Third, the estimated trajectories are smoothed, and the average difference ($d$) between the smoothed informational text complexity trajectory and the smoothed literary text complexity trajectory is determined. Fourth, an adjusted Grade 12 score ($y$*) is estimated by assuming that the Grade 12 median reported in Williamson (2008), 1130, can be modeled as a mixture of $p$ texts from the literary trajectory, and (1 – $p$) texts from the informational trajectory, where $p$ is the proportion of literary texts in the HS sample analyzed in Williamson (2008). This approach can be summarized as follows:

$$1130 = p\left(y^{*} - d\right) + \left(1{-}p\right)y^{*}, \tag{A1}$$

which reduces to

$$y^{*} = 1130 + p\left(d\right). \tag{A2}$$

Finally, an updated estimate of the HS/college text complexity gap is obtained by subtracting $y$* from the college median estimated above. Note that this new estimate is less subject to distortions due to genre DFF because each of the resulting medians is constructed to represent text complexity variation among informational text.

**Table A2** Genre-Specific Current-Day Reading Demand Curves, Before Smoothing

| | Numbers of passages | | | Median Lexile scores | | |
|---|---|---|---|---|---|---|
| Grade | Inf. | Literary | Total | Inf. | Literary | Difference |
| 3 | 34 | 44 | 78 | 785 | 610 | 175 |
| 4 | 31 | 54 | 85 | 880 | 725 | 155 |
| 5 | 31 | 30 | 61 | 910 | 755 | 155 |
| 6 | 23 | 22 | 45 | 960 | 755 | 205 |
| 7 | 26 | 43 | 69 | 1030 | 860 | 170 |
| 8 | 39 | 34 | 73 | 1090 | 945 | 145 |
| 9 | 22 | 16 | 38 | 1180 | 1030 | 150 |
| 10 | 22 | 40 | 62 | 1200 | 950 | 250 |
| 11–12 | 15 | 22 | 37 | 1190 | 1070 | 120 |
| Total | 243 | 305 | 548 | 1030 | 860 | 170[a] |

*Note*. Lexile scores were obtained by sending each passage through the Lexile Analyzer available at www.lexile.com. Inf. = informational.
[a] After smoothing, the median difference is about 150 Lexile points.

Table A2 shows the results obtained when genre-specific trajectories are estimated from the collection of 548 passages described in Sheehan et al. (2010). All of the passages in this collection were selected from high-stakes state assessments constructed to provide evidence of students' proficiencies relative to the reading skills specified in published state reading standards. Note that the informational trajectory is higher than the literary trajectory at each GL. After smoothing the trajectories the median difference is calculated as $d = 150$ Lexile points. This suggests that a correction factor of $d = 150$ Lexile points can be used to estimate the median Lexile score that would have been observed if the Grade 12 sample had been entirely composed of informational text.

## Results

Table A3 shows how estimates of the HS/college text complexity gap vary as one or both of the two refinements introduced above are implemented. The gap estimate used to generate the accelerated text complexity exposure trajectory referenced in the CCSS is listed in the first row. The second row shows how this estimate would change if Refinement 1 was implemented (i.e., if peripheral genres such as Wikipedia articles and articles from international newspapers were excluded from the sample used to represent the complexity levels of college and career texts). Note that this one change reduces the estimated magnitude of the gap from 170 Lexile points to 120 Lexile points, a reduction of 29%.

The next three rows of Table A3 show how the magnitude of the gap changes when Equation A2 is implemented with $d$ set at 150 and with the percentage of literary texts in the HS sample ($p$) set at any of three different levels: 10%, 20%, or 30%. More than one value of $p$ is evaluated because the actual proportion of literary texts in the HS sample has not been reported. The results show that each increase in the percentage of literary texts in the HS sample leads to an additional decrease in the estimated magnitude of the gap because the genre correction is then implemented for a larger number of texts. Table A3 also shows that a strategy of implementing both refinements simultaneously leads to a gap estimate that is 40% to 50% lower than the gap estimate referenced in the current CCSS. These findings suggest that the text complexity exposure trajectory referenced in the current CCSS may accelerate text complexity expectations beyond the level that is actually needed to ensure that students are exposed to the types of complex texts that they will be expected to read in college and careers.

## Discussion

There are important reasons for getting the size of the HS/college text complexity gap right: the size of the gap determines the amount by which text complexity expectations must be accelerated in order to ensure that all HS graduates are adequately prepared for the advanced reading demands of college and careers. A smaller gap means less acceleration is needed; a larger gap means more acceleration is needed. Because both underacceleration and overacceleration could have negative consequences for both students and teachers, a precise estimate is needed.

**Table A3** The High School/Postsecondary Text Complexity Gap Before and After Implementing Proposed Refinements

| Refinements included[a] | | CCR target | % of literary texts in the HS sample | Grade 12 median complexity | Estimated gap | % Reduction |
|---|---|---|---|---|---|---|
| 1 | 2 | | | | | |
| No | No | 1300 | *nr* | 1130 | 170 | — |
| Yes | No | 1250 | *nr* | 1130 | 120 | 29 |
| Yes | Yes | 1250 | 10% | 1145 | 105 | 38 |
| Yes | Yes | 1250 | 20% | 1160 | 90 | 47 |
| Yes | Yes | 1250 | 30% | 1175 | 75 | 56 |

*Note*. CCR = college and career ready, HS = high school, *nr* = not reported.

[a] Refinement 1 = exclude Wikipedia articles and articles from international newspapers, Refinement 2 = adjust the HS median to account for the lower Lexile scores typically assigned to literary texts.

This appendix evaluated two relatively minor refinements to the methodology used to estimate the HS/college test complexity gap adopted for use in defining the accelerated text complexity exposure trajectory referenced in the current CCSS: (a) excluding peripheral genres when representing the reading demands of college and workplace texts and (b) addressing genre effects when assigning a complexity score to each text. Key results are discussed below.

## Refinement 1: Exclude Peripheral Genres

As in any inferential problem, methods for reducing uncertainty can lead to estimates that are more stable, less subject to problematic biases, and more likely to hold up over the long term. The analyses summarized above suggested that the decision to include Wikipedia articles and articles from international newspapers in the postsecondary sample increased the median complexity of the resulting sample from a Lexile score of about 1250 to a Lexile score of 1300. This large increase yielded a correspondingly large increase in the estimated magnitude of the HS/college text complexity gap. Because the unusually high Lexile scores obtained for Wikipedia articles and articles from international newspapers have not yet been validated, however, a strategy of excluding these alternative genres from the postsecondary sample seems warranted. Note that this one slight change reduces the estimated magnitude of the gap by about 29%.

## Refinement 2: Include an Adjustment for Genre Differential Feature Functioning

The authors of the CCSS have argued that understanding and measuring text complexity are fundamental to determining if students are adequately prepared for the academic and professional reading demands they will face after HS. The current analyses have suggested, however, that certain aspects of the text complexity measurement process are still not adequately understood. For example, while the finding that the Lexile Framework and traditional readability metrics tend to underestimate the complexity levels of literary texts is reported in Appendix A of the CCSS, this important source of estimation bias was not accounted for when estimating the accelerated text complexity exposure trajectory that teachers and students throughout the United States are now being asked to embrace.

If the samples used to estimate the HS/college text complexity gap had included similar proportions of informational and literary texts, the failure to account for genre DFF might have had a much smaller impact on subsequent inferences. Because the HS sample appears to have included a higher proportion of literary texts, however, the strategy of simply assuming that genre-based DFF is either not present, or if present, can always be safely ignored, means that Lexile scores for HS texts may be slightly *underestimated*, while those for college texts may be slightly *overestimated*, thereby yielding an unintended increase in the estimated magnitude of the gap. The updated gap estimates in Table A3 suggest that these genre-based distortions may have added as much as 50 Lexile points to the estimated magnitude of the gap. By contrast, a strategy of implementing both Refinement 1 and Refinement 2 could decrease the estimated magnitude of the gap by as much as 60 to 95 Lexile points.

## Recommendations for Additional Research

This appendix introduced two relatively minor refinements to the gap estimation methodology referenced in the CCSS. Although the exact impact of these refinements cannot be stated precisely (because median Lexile scores for some types of texts have not been reported and because the proportion of literary texts in the HS sample has also not been reported), current estimates suggest that a strategy of implementing both refinements simultaneously could lead to an updated gap estimate that is noticeably lower than the current estimate. Because this difference could have important consequences for both students and teachers, the two refinements described above should be considered as additional research focused on the CC text complexity exposure trajectory is planned, implemented, and reviewed.

**Suggested citation:**

Sheehan, K. M. (2015). *Aligning TextEvaluator® scores with the accelerated text complexity guidelines specified in the Common Core State Standards* (Research Report No. RR-15-21). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12068

**Action Editor:** James Carlson

**Reviewers:** Chaitanya Ramineni and Neil Dorans