Measuring the Power of Learning.®

Research Report
ETS RR–17-33

Helping Students Select Appropriately Challenging Text: Application to a Test of Second Language Reading Ability

Kathleen M. Sheehan

December 2017

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Senior Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Research Director

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Anastassia Loukina
Research Scientist

John Mazzeo
Distinguished Presidential Appointee

Donald Powers
Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Elizabeth Stone
Research Scientist

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Gontz
Senior Editor

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

Helping Students Select Appropriately Challenging Text: Application to a Test of Second Language Reading Ability

Kathleen M. Sheehan

Educational Testing Service, Princeton, NJ

A model-based approach for matching language learners to texts of appropriate difficulty is described. Results are communicated to test takers via a targeted reading range expressed on the reporting scale of an automated text complexity measurement tool (ATCMT). Test takers can use this feedback to select reading materials that are well matched to their abilities, that is, materials that are expected to be challenging, yet not so challenging as to cause frustration or reduce motivation. An application to the problem of helping students prepare to take the *TOEFL iBT*® test is presented.

Keywords Reader–text matching algorithms; *TOEFL iBT*® reading section; *TextEvaluator*®; Lexile

doi:10.1002/ets2.12160

Research over the past several years has highlighted the importance of encouraging readers to engage with texts that are well targeted to their abilities, that is, texts that are expected to be challenging, yet not so challenging as to cause frustration or reduce motivation (International Reading Association, 2004). This research, combined with recent advances in natural language processing techniques, and the increasing availability of large collections of electronic text, has sparked a renewed interest in automated approaches for matching language learners to texts of appropriate difficulty.

Existing approaches for matching readers to texts may be classified as belonging to one or another of two categories: indirect approaches or direct approaches. In the indirect approach, a linking study is used to establish a quantitative relationship between the reporting scales of two different assessments: a broad-based reading assessment designed to provide student-level evidence intended to support admissions or graduation decisions, and a second, more narrowly focused reading assessment administered for the sole purpose of distinguishing the types of texts expected to fall within a student's optimal reading range, that is, a reading range that is expected to be challenging, yet not too challenging. A *targeted reading range* can then be generated for any student who took the first assessment by inferring what his or her performance on the second, more narrowly focused assessment would have been if that second assessment had actually been administered. In the direct approach, by contrast, only one assessment is administered, and evidence collected via that one, more narrowly focused assessment is used to define a targeted reading range for each student.

Both the indirect and the direct approaches have been characterized as providing useful information about the types of texts that may help students improve their reading skills. Because each approach incorporates a different combination of advantages and limitations, however, a strategy of combining the most effective elements from each approach could lead to improved performance and, thus, better outcomes for students.

This report investigates a hybrid approach that combines the reader ability estimation technique implemented within the indirect approach with a variation of the text complexity estimation technique implemented within the direct approach. Although the resulting algorithm can be applied to any passage-based reading comprehension assessment, and can be implemented with respect to scores generated via any automated text complexity measurement tool (ATCMT), subsequent matches are likely to be most accurate when the reading proficiency construct targeted by the selected assessment is closely aligned with the reading proficiency construct adopted during the design and development of the selected ATCMT.

This report is organized as follows. First, existing approaches for matching readers to texts are reviewed, and the advantages and limitations of each approach are summarized. Next, a hybrid matching algorithm is introduced, and an illustrative application of the proposed approach is presented. The application is focused on a particular matching

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

problem: helping students prepare to take the *TOEFL iBT®* test. The TOEFL iBT is an English proficiency test taken by nonnative-English-speaking students as a requirement for admission to colleges and universities where English is the predominant language of instruction. An earlier study of reader–text matches generated for TOEFL iBT test takers is reported in Wendler, Cline, Sanford, and Aguirre (2010). In contrast to the current study, which employs a hybrid approach, the reader–text matching algorithm employed by Wendler and her colleagues was implemented via the indirect approach. Additional information about each application is summarized in the following section.

Existing Approaches for Matching Readers to Texts

Messick (1987) argued that large-scale assessments are likely to be most useful when scores exhibit three key features: relevance, comparability, and interpretability. Both the indirect approach and the direct approach are designed to enhance the interpretability of scores generated via a passage-based reading comprehension assessment by providing concrete information about the types of texts that students who score at specified points on the assessment's reported score scale are expected to be able to comprehend.

The Indirect Approach

Wendler et al. (2010) employed an indirect reader–text matching algorithm to link test takers' scores on the TOEFL iBT reading section to text complexity scores expressed on the Lexile scale (Stenner, Burdick, Sanford, & Burdick, 2006). A four-step approach was used to establish the needed link. First, students who registered for several spring 2009 TOEFL iBT administrations were invited to take part in a linking study, and a total of 3,420 students agreed to participate. Each participating student took an operational form of the TOEFL iBT reading assessment and received a reading ability score expressed on the TOEFL iBT reading scale.

Second, one of four Lexile linking tests was administered to each student. Each test included 45 items selected from the Lexile item bank, with 22 items common across all forms. Each item consisted of a single paragraph of text followed by a single fill-in-the-blank question presented with four options. A sample item is shown in Figure 1.

An important characteristic of the Lexile item bank is that only items that have passed a rigorous item review process are included. The review process is designed to identify and exclude any item that is not consistent with the text complexity construct targeted by the Lexile prediction model. This model posits that the difficulty level of a text (also called the *theoretical difficulty* of a text) can be accurately estimated from two machine-measurable text characteristics: log average sentence length (LASL), a proxy for sentence complexity, and average log word frequency (ALWF), a proxy for vocabulary difficulty. For shorter texts, the model is specified as follows:

$$L_j = \beta_o + \beta_1 \left(\mathrm{LASL}_j\right) + \beta_2 \left(\mathrm{ALWF}_j\right), \tag{1}$$

where L_j is the theoretical difficulty level of the jth text (expressed on the Lexile scale) and β_0, β_1, and β_2 are known constants. For longer texts, the text is first broken up into shorter segments, Equation 1 is independently applied within each segment, and the resulting segment-level scores are accumulated to form a single text-level score. Because each of the items in the Lexile item pool presents no more than a single paragraph of text, however, the strategy of breaking the text up into shorter segments was not needed and so was not implemented. Consequently, the theoretical item difficulty parameters generated for each item in the Lexile item pool, and therefore for each item on the Lexile linking tests, were entirely determined by applying Equation 1 to two particular features of each paragraph: the average length of its sentences and the average frequency of its words.

Although the Lexile linking test was administered to each of the 3,420 students identified at Step 1, several of the resulting response vectors indicated insufficient effort, so they were excluded from all subsequent analyses. A shifted, anchored Rasch analysis was then used to generate a reading proficiency score expressed on the Lexile scale for each retained test taker. Anchored Rasch analyses are frequently employed when item difficulty parameters from a previous Rasch analysis are available. In the application described in Wendler et al. (2010), however, the "anchor" difficulties submitted to the Rasch analysis were not developed from data collected in previous data collection efforts. Rather, theoretical item difficulty parameters obtained by applying Equation 1 to each of the paragraphs on the four Lexile linking tests were entered into the analyses as if they were known, true values.

It was Tim's first time flying in an airplane, and he didn't want to miss a single thing, so he leaned over in his seat and pressed his forehead against the cold glass of the window. Looking out the window, Tim saw the silver wing of the airplane with the jet engine hanging beneath it, but beyond he could only see white clouds. As Tim looked on, the clouds suddenly cleared away and he was able to see the ground far below. Tim could see rivers winding through forests and lakes reflecting the sunlight. As they passed above one city, Tim saw a baseball field, but from the air it looked so tiny that Tim couldn't believe players could fit on it.

Tim had an unusual _____.

A. livelihood
B. perspective
C. disposition
D. inspiration

Figure 1 A sample item from the Lexile item pool. Reprinted from "Linking TOEFL Scores to the Lexile Measure," by C. Wendler, F. Cline, E. Sanford, and A. Aguirre, paper presented at the Language Testing Research Colloquium, Cambridge, UK, 2010, p. 15.

Let x_{ij} represent the response provided by the ith test taker when responding to the jth item on the Lexile linking test. The shifted anchored Rash model specifies the probability that $x_{ij} = 1$ (a correct response) rather than 0 (an incorrect response) as follows:

$$P\left(x_{ij} = 1|\theta_i, L_j\right) = \left[\frac{\exp\left[a\left(\theta_i - L_j\right) + 1.1\right]}{1 + \exp\left[a\left(\theta_i - L_j\right) + 1.1\right]}\right], \quad (2)$$

where θ_i is the unknown ability level of the ith test taker, L_j is the theoretical difficulty parameter generated for the jth item by applying Equation 1, $a = 0.0056$ is a scaling factor used to translate θ_i and L_j onto a logit scale (instead of the more familiar Lexile scale), and 1.1 is a shift parameter that reflects the decision to define a successful comprehension episode as one in which the reader has a 75% chance of responding correctly. In other words, when $\theta_i - L_j = 0$, Equation 2 reduces to $[\exp(1.1)]/[1 + \exp(1.1)] = 0.75$, so a *close match* between reader ability and text complexity is defined as one that yields a correct response probability of 0.75, rather than 0.50, as would be the case in a traditional Rasch model (Stenner et al., 2006; Stenner, Fisher, Stone, & Burdick, 2013; Swartz et al., 2014).

Next, a subset of test takers judged to have responded in a manner that was consistent with the proposed Rasch model was selected, and Equation 2 was used to generate a reading ability estimate expressed on the Lexile scale ($\widehat{\theta}_i$) for each of those test takers. The resulting estimates are plotted versus test takers' TOEFL iBT reading scores in Figure 2. A total of 2,867 points are plotted, one for each of the 2,867 test takers retained in the final sample. The plot suggests that, after excluding 553 test takers with missing or misfitting responses, there is a moderate positive relationship ($r = .66$) between students' TOEFL iBT reading scores and their scores on the Lexile linking test.

In the final step of the approach, an equipercentile method was used to establish a link between the TOEFL iBT reading scores obtained for each student at Step 1 and the corresponding set of Lexile reading proficiency scores obtained for each retained student at Step 2. Wendler et al. (2010) argued that the resulting correspondence table can help us distinguish the types of reading materials that TOEFL iBT test takers at successive points on the TOEFL iBT reading scale are likely to be able to comprehend.

Score correspondences generated via the preceding approach have been incorporated into an interactive tool that is currently available online.[1] This tool suggests, for example, that a TOEFL iBT reading score of 10 corresponds to a Lexile reading score of 1040, so that a test taker with a TOEFL iBT reading score of 10 is likely to be well matched to any text

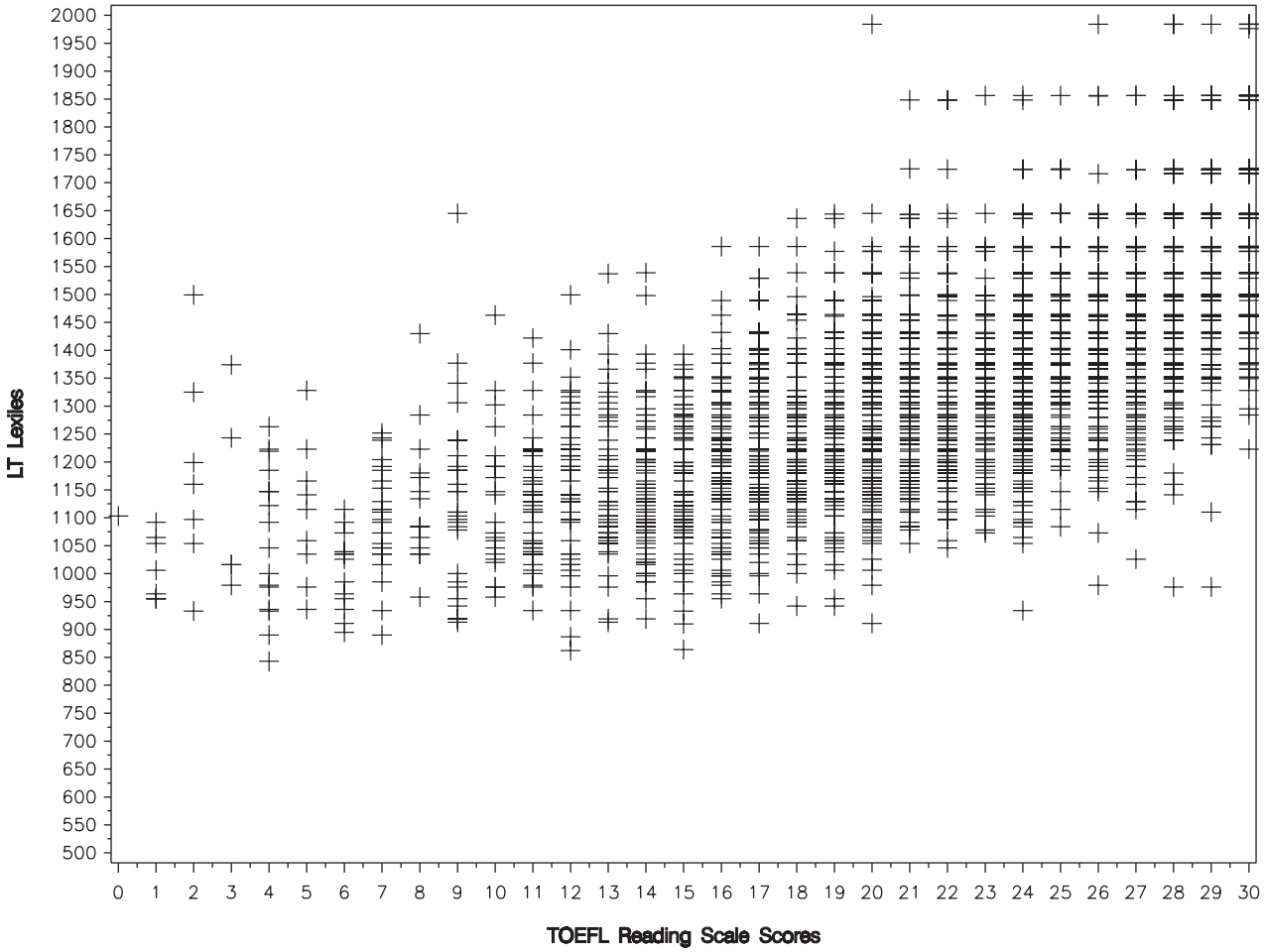

Figure 2 Relationship between TOEFL iBT reading scores and Lexile scores ($r = .66$). Reprinted from "Linking TOEFL Scores to the Lexile Measure," by C. Wendler, F. Cline, E. Sanford, and A. Aguirre, paper presented at the Language Testing Research Colloquium, Cambridge, UK, 2010, p. 16.

that has a Lexile text complexity score that falls between 940 and 1090, that is, anywhere from 100 points below the test taker's concorded Lexile reading ability score to 50 points above the test taker's concorded Lexile reading ability score.

Several limitations of this matching procedure should be noted. First, Dorans (1999) argued that the method of matching score distributions via an equipercentile approach is likely to be most appropriate when the observed correlation between two sets of scores is at least .866. As is noted earlier, however, the two sets of scores considered in the Wendler et al. (2010) analysis yielded an observed correlation of just .66, and this was *after* a number of outlying scores had already been deleted.

A variety of factors may account for the low correlation between test takers' scores on the TOEFL iBT reading section and on the Lexile linking test. Three key factors include (a) differences in the types of passages included on each assessment, (b) differences in the types of items included on each assessment, and (c) the additional estimation errors that may have been introduced during the process of generating a reading ability score expressed on the Lexile scale for each test taker. Additional information about these factors is summarized in the following paragraphs.

Passages on the TOEFL iBT reading section are selected from introductory-level college textbooks and university press books judged to be suitable for use in introductory-level college courses. Each passage typically includes approximately 700 words. By contrast, all of the passages on the Lexile linking test present a single paragraph of text, yielding an average passage length of 120 words.

The items included on each assessment are also very different. For example, items on the TOEFL iBT reading section are designed around a view of reading that highlights the reader purpose perspective, that is, the notion that reading takes

place "in the service of a goal or purpose" (Enright & Schedl, 2000, p. 4). This perspective acknowledges that the specific goal set forth in a reading task may require the reader to engage in effortful processes, such as anticipating information, distinguishing between primary and secondary ideas, organizing and mentally summarizing information, monitoring comprehension, repairing comprehension breakdowns, and aligning comprehension outputs with reading goals (Grabe, 2009).

By contrast, each Lexile reading item consists of a short fill-in-the-blank item stem followed by four single-word options (see Figure 1). This item format, combined with the use of single-paragraph passages, suggests that test takers may have spent a large amount of time puzzling over individual words rather than implementing the broader array of reading competencies targeted by TOEFL iBT reading items. Thus the reading skills measured by the Lexile linking test may not be closely aligned with those measured by the TOEFL iBT reading assessment.

Errors introduced during the process of using an anchored Rasch analysis to generate a reading ability score expressed on the Lexile scale ($\widehat{\theta}_i$) for each test taker may have also contributed to the relatively low correlation reported by Wendler et al. (2010). To see why this might be the case, consider a test in which each test taker reads exactly one passage. If the theoretical difficulty score generated for that passage were 100 points too high, the reading ability scores generated for each of the students who read that passage would also be approximately 100 points too high. In other words, the anchored Rasch analysis is structured such that any errors in the theoretical difficulty parameters generated via Equation 1 are automatically translated into corresponding errors in the reading ability estimates generated by applying Equation 2 to the item responses provided by each student. This characteristic of the anchored Rasch approach, combined with the differences in passages and items noted earlier, suggests that alternative approaches for matching readers to texts could lead to improved feedback and, thus, better outcomes for students.

The Direct Approach

The Metametrics Oasis platform (now called Edsphere) is an example of a reader–text matching platform implemented via the direct approach (Stenner et al., 2013; Swartz et al., 2014). Stenner et al. (2013) described this innovative platform as follows:

> The Edsphere platform enables students to select articles of their choosing from a vast range of content. Selected articles are targeted to ±100 L of each student's developing reading ability. Thus, as students' reading ability grows, the machine adjusts the text complexity of the articles from which the student chooses the next reading. The target success rate is 75%. The machine generates a reading comprehension item on the fly about every 70 words such that two students sitting side by side at computers and reading the same article will respond to different items. (p. 540)

The reader–text matching algorithm implemented within the Edsphere platform is similar to the algorithm described in Wendler et al. (2010) in some respects, yet different in others. One key similarity is that theoretical text complexity parameters generated via Equation 1 are employed in the analyses as if they were known, true values. But a number of distinctive elements are also present. For example, (a) texts vary in length and are administered with multiple items instead of just one; (b) items are automatically generated on the fly and are not reviewed prior to administration; (c) students' reading abilities are assumed to be constantly growing so that a new reading ability estimate ($\widehat{\theta}_i$) is generated for each student at the conclusion of each reader–text encounter; and (d) students are only allowed to read texts with theoretical complexity scores (L_j) that are no more than 100 Lexile points different from their estimated reading ability scores ($\widehat{\theta}_i$) so that each reading experience involves some challenge, but not too much challenge.

The unusual nature of the item response data collected via the Edsphere platform is such that calibration via traditional item response theory (IRT) models is not feasible. Lattanzio, Burdick, and Stenner (2012) addressed this problem by introducing a new type of item response model, called an ensemble Rasch model (ERM). The ERM differs from traditional IRT models in that each individual item difficulty parameter is modeled as a random instance drawn from a difficulty distribution constructed to characterize the difficulty of the text that an item refers to rather than the difficulty of the individual item administered. Because resulting parameter estimates refer to texts instead of individual items, they are called *text difficulty estimates* or *text complexity estimates* rather than item difficulty estimates.

Let x_{ijk} represent the response provided by the ith student when responding to the kth item administered with the jth Edsphere text. The ERM specifies the probability that $x_{ijk} = 1$ (a correct response) rather than 0 (an incorrect response), as follows:

$$P\left(x_{ijk} = 1|\theta_i, L_j, \sigma\right) = E\left[\frac{\exp\left[a\left(\theta_i - L_j\right) + 1.1 + \varepsilon_{ijk}\right]}{1 + \exp\left[a\left(\theta_i - L_j\right) + 1.1 + \varepsilon_{ijk}\right]}\right], \quad (3)$$

where $E[\ ]$ is the expected value operator, θ_i is the unknown ability level of the ith student, L_j is the difficulty parameter generated for the passage that Item ijk refers to by applying Equation 1, $a = 0.0056$ is a scaling factor used to translate θ_i and L_j onto a logit scale (instead of the more familiar Lexile scale), 1.1 is a shift parameter used to establish 0.75 as the expected proportion correct score when $\theta_i - L_j = 0$, and ε_{ijk} is an error term that is assumed to be normally distributed with mean 0 and standard deviation σ, where σ is a known constant, not a parameter to be estimated.

Because Equation 3 has only one unknown parameter (θ_i), Swartz et al. (2014) noted that the ERM estimation process can be implemented via "a look-up table that accounts for text complexity and percent correct" (p. 366). Table 1 presents six rows from the referenced look-up table. The selected rows show the reader ability estimates ($\widehat{\theta}_i$) generated via Equation 3 for any reader who achieved a proportion correct score of 0.62, 0.75, or 0.82 on a text with a theoretical text complexity score of $L_j = 500$ or a text with a theoretical text complexity score of $L_j = 1000$. As is illustrated in the table, an observed score of 0.75 signifies that reader and text are aligned, that is, $\theta_i - L_j = 0$, so the reader is assigned an ability estimate ($\widehat{\theta}_i$) equal to the theoretical Lexile score (L_j) of whatever text was read. By contrast, a reader who only achieved an observed score of 0.62 on the single text read would be rated as being 100 points less able than the ability needed to read the text with 75% comprehension, so the reader would be assigned an ability estimate that is 100 points below the theoretical Lexile score of whatever text was read. Similarly, a reader with an observed score of 0.82 on the single text read would be rated as being 100 points more able than the ability needed to read the text with 75% comprehension, so the reader would be assigned an ability estimate that is 100 points greater than the Lexile score of whatever text was read.

Under the usual assumption of conditional independence, Equation 3 can be used to "forecast the level of comprehension a reader will have with a specific text" (Swartz et al., 2014, p. 360). Furthermore, because the Edsphere algorithm is structured such that students are matched to texts that are predicted to fall within ±100 Lexile points of their estimated reading ability scores, Equation 3 predicts that many of the observed proportion correct scores collected via the Edsphere platform will fall within the predicted interval from 0.62 to 0.82, inclusive.

Figure 3 summarizes a collection of 76,538 proportion correct scores collected as students read one or more of 372 texts within the Edsphere platform.[2] Each score represents the observed performance of a single student when reading a single text and responding to the computer-generated multiple-choice cloze items administered with that text. When describing these data, Stenner et al. (2013) noted that "well-estimated reader measures were available prior to an encounter between an article and a reader" (p. 15). Swartz et al. (2014) presented a similar description that "most of the articles read were well-targeted to student ability (±100 L)" (p. 359). Consequently, Equation 3 predicts that many of the observed proportion correct scores will fall within the predicted interval from 0.62 to 0.82, inclusive.

Table 1 Look-up Table Constructed to Provide Reader Ability Estimates Expressed on the Lexile Scale When Theoretical Text Complexity and Proportion Correct Are Known

Student–text encounter	Theoretical difficulty of the specific text read[a] (L_j)	Observed proportion of correct responses to cloze items (p_{ij})	Estimated reader ability parameter expressed on the Lexile scale[b] ($\widehat{\theta}_i$)
1	500	0.62	400
2	500	0.75	500
3	500	0.82	600
4	1,000	0.62	900
5	1,000	0.75	1,000
6	1,000	0.82	1,100

[a]Estimated using Equation 1 and then assumed to be known without error. [b]Estimated using Equation 2 and then rounded to the nearest 100.

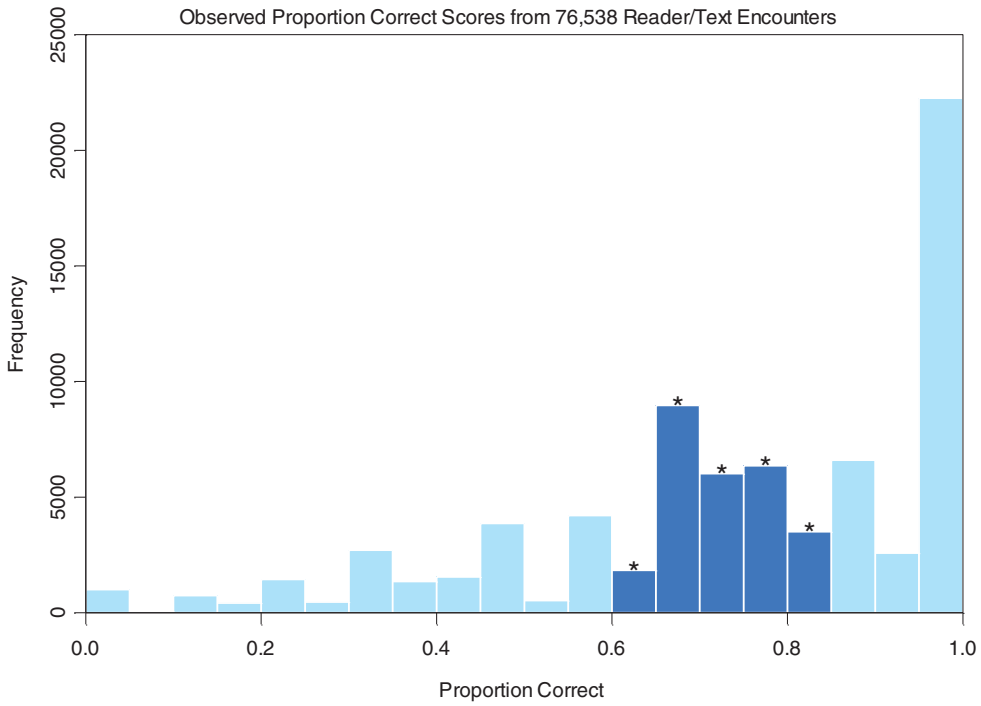

Figure 3 Histogram of 76,538 proportion correct scores collected as students read one or more of 372 texts and responded to multiple-choice fill-in-the-blank cloze items administered via the Edsphere platform. Proportion correct scores that fell within the range of variation forecasted by the Lexile theory are plotted within the five bars rendered with darker shading. These bars are also marked with a star.

Raw proportion correct scores that fell within the predicted interval from 0.62 to 0.82 are plotted within the five bars rendered with darker shading. These five bars account for a total of 26,412 observed proportion correct scores, or 34.5% of the total set of scores analyzed. Thus only a small proportion of the observed encounters (less than 35%) resulted in an observed proportion correct score that fell within the interval predicted by the Lexile theory. These results suggest that alternative approaches for matching readers to texts may be more effective at helping test takers select texts that are well matched to their abilities.

A New Approach for Matching Readers to Texts

The reader–text matching algorithm introduced in this article can be implemented with respect to any reading assessment that includes a sufficient number of passage-based reading items and any text complexity measurement tool that is expected to be closely aligned with the reading construct targeted by that assessment. The approach incorporates elements selected from both the indirect approach and the direct approach, while also introducing a number of completely new elements.

An element selected from the indirect approach is the strategy of characterizing students' current reading proficiency levels via their scores on a reading assessment used in making high-stakes decisions, such as the TOEFL iBT. Elements selected from the direct approach include the strategy of estimating both student ability and text complexity from the same set of observed item responses.

This new approach also adopts a key element of both previous approaches: Passage comprehension is operationally defined as the reading ability needed to respond correctly to 75% of the items administered with a passage. Unlike the previous research summarized earlier, however, difficulty estimates generated via a theoretical model are not used to estimate this ability level. Rather, a modification of Kirsch's (2001) approach is used to generate a threshold score for each passage, and the relationship between those scores, and corresponding passage difficulty estimates generated via an ATCMT, is determined. This new approach is illustrated in Figure 4. A more detailed description is presented in the following pages, after which an application focused on the goal of helping students prepare to take the TOEFL iBT reading assessment is presented.

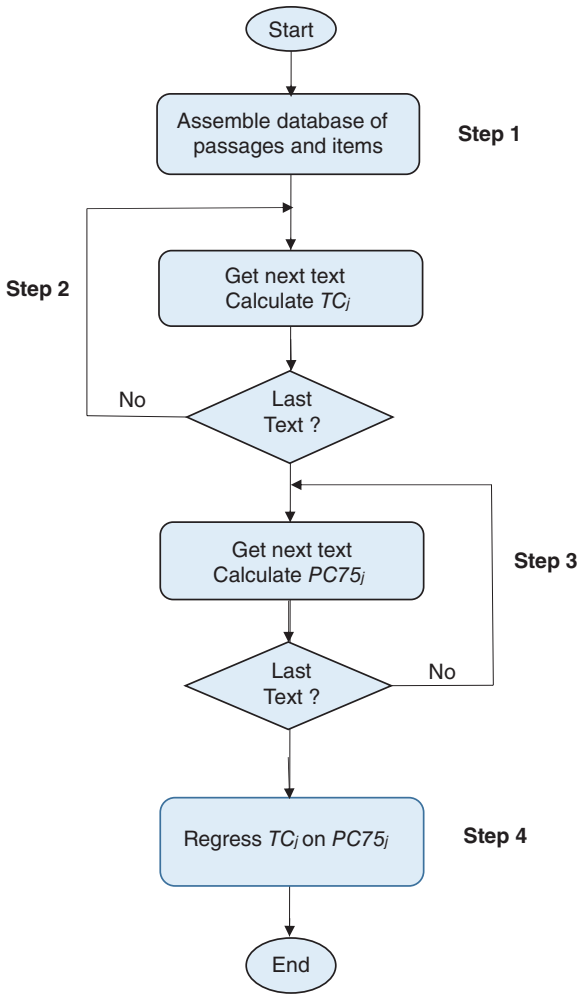

Figure 4 The four-step process used to quantify the relationship between text complexity scores generated via an automated text complexity measurement tool (TC_j) and passage difficulty scores estimated from students' responses to test questions on a passage-based reading assessment ($PC75_j$).

Step 1: Assemble a Database of Reading Comprehension Passages and Items

Many reading assessments are structured such that each item is designed to simulate the types of reading tasks that students would be expected to perform in real-life reading episodes. For example, items on the TOEFL iBT reading assessment are designed to simulate the types of reading tasks that students are likely to engage in at colleges and universities in North America. The reading skills needed to respond correctly to these types of passage-based reading items have been described as encompassing two types of processes: (a) processes focused on the goal of developing a coherent mental representation of the information, argument, situation, or experience presented in the passage and (b) processes focused on the goal of applying additional verbal reasoning skills, as needed, to address additional processing demands, such as clarifying the type of information requested in an item stem or ruling out a close distractor (Embretson & Wetzel, 1987; Gorin, 2005; Gorin & Embretson, 2006). Gorin (2005) referred to these two types of processes as *text representation* and *response decision*.

When selecting texts that are well matched to a test taker's reading abilities, we are primarily interested in the text representation aspect of reading ability rather than in the response decision aspect. Thus the reader–text matching algorithm introduced in this study is implemented with respect to a subset of items judged to be closely focused on the text representation aspect of comprehension. This subset is selected by starting with a large pool of passages and items and then retaining just those items that are judged to be most appropriate for use in the proposed application. As is illustrated in what follows, item classifications developed as part of the item development process and the degree of alignment between

empirical and theoretical estimates of text complexity are considered when selecting an optimal subset of items for each passage.

Step 2: Generate a Text Complexity Score (TC_j) for Each Passage

In this step, an ATCMT is used to generate an estimated text complexity score (TC_j) for each passage. Although a variety of different types of automated tools could be employed, resulting feedback is likely to be most accurate when the reading ability construct employed in the development of the selected tool is closely aligned with the reading ability construct targeted by the selected reading assessment.

Step 3: Generate a Reading Difficulty Score ($PC75_j$) for Each Passage

Previous research reported in Kirsch (2001) has also played a significant role in the development of the new reader–text matching algorithm described in this report. This research introduced the RP80 score as an approach for helping test users understand the types of reading skills needed to score at lower and higher levels on the reading scale of the International Adult Literacy Survey (IALS).

An RP80 score was generated for each IALS item as follows. First, a three-parameter logistic IRT model (Lord, 1980) was used to model students' observed responses to each IALS item. Next, an item characteristic curve (ICC) was generated for each item. Each ICC provides the probability that a test taker will respond correctly to an item expressed conditional on the test taker's IALS reading proficiency score. Finally, an RP80 score was defined for each item by determining the reading proficiency score needed to achieve a correct response probability of at least 80%. By definition, then, a test taker with a reading proficiency score that falls below an item's RP80 score is expected to have less than an 80% chance of responding correctly to the item, whereas a test taker with a reading proficiency score that falls above an item's RP80 score is expected to have an 80–100% chance of responding correctly to the item. Thus the RP80 measure provides a method for distinguishing items that are likely to be more or less challenging for test takers located at any specified point on a reading proficiency scale.

The RP80 measure introduced by Kirsch and his colleagues was conceptualized as an item characteristic. In many reading assessments, however, items are clustered within passages. For example, each form of the TOEFL iBT includes a reading section comprising three, four, or five passages, with each passage followed by 12 to 14 items. To properly address this alternative format, this report introduces an extension of the RP80 concept that is conceptualized as a passage characteristic rather than as an item characteristic. In this new approach, a passage comprehension curve (PCC) is generated for each passage by summing the IRT-based correct response probabilities estimated for each retained item in each passage set and then dividing by the total number of retained items. Each resulting PCC provides the probability that a test taker will respond correctly to any of the text representation items presented with a passage, expressed conditional on the test taker's reading proficiency score. A threshold difficulty score can then be generated for each passage by solving for the reading proficiency score at which a test taker achieves some minimum proportion of correct responses. Consistent with earlier research presented in Stenner et al. (2006), a threshold score of 75% correct was selected for use in this research. A PC75 score was then defined for each passage as the passage comprehension score at which a student is expected to respond correctly to at least 75% of the text representation items administered with a passage. A passage that is well matched to a test taker's reading ability can then be defined as any passage that has a PC75 score that falls within a relatively narrow interval centered about the test taker's reading proficiency score.

Step 4: Estimate the Regression of Text Complexity Scores (TC_j) on Passage Difficulty Scores ($PC75_j$)

A key limitation of the matching process outlined earlier is that only those passages that were administered on the specified assessment, and thus have operational IRT item parameters expressed on the same scale as a student's reading ability score, can be considered as a potential match. Because readers may also want to read texts that were not included on the targeted reading assessment, however, a procedure for extending the definition of a well-matched text to a broader class of texts is needed.

In the innovative reader–text matching approach introduced in this report, a more broadly applicable definition of a well-matched text is developed by establishing a link between passage difficulty estimates specified in terms of PC75

scores, which are only available for texts that were included on the targeted assessment, and passage difficulty estimates specified via an ATCMT, which can be generated for any text. If a valid link can be established, then an *expected* PC75 score can be generated for each potential text, and a well-matched text can then be defined as any text that has an expected PC75 score that falls near the test taker's reading proficiency score, where each expected PC75 score is expressed on the measurement scale of the selected ATCMT.

Consistent with the guidelines specified in Dorans (1999), a regression technique is used to establish a link between the text complexity scores estimated for each passage in Step 2 and the PC75 scores estimated for each passage in Step 3. In particular, a locally weighted scatterplot smoother (Cleveland & Devlin, 1988) is used to characterize TC_j conditional on $PC75_j$. The resulting smoothed curve provides the range of text complexity scores that corresponds most closely to each possible PC75 score. A concordance table based on this relationship can then be developed. The resulting table will support inferences from a test taker's score on the selected reading assessment to a corresponding range of text complexity scores expressed on the reporting scale of the selected ATCMT.

For example, consider a test taker who receives a score of 20 on the TOEFL iBT reading assessment. By definition, this test taker is expected to be well matched to any text that has a PC75 score near 20. A concordance table generated via the preceding approach would enable us to translate any PC75 score into a corresponding range of text complexity scores expressed on the reporting scale of an ATCMT, so the set of texts that are likely to be well matched to a test taker's current reading ability can also be specified more generally. This greatly expands the universe of texts that test takers can consider, because text complexity scores generated via a variety of ATCMTs are readily available (Nelson, Perfetti, Liben, & Liben, 2012).

Application to the *TOEFL iBT*® Reading Assessment

This section presents an application of the proposed reader–text matching algorithm to the problem of helping TOEFL iBT test takers select texts that are well matched to their abilities. The application is implemented with respect to the *TextEvaluator*® text analysis tool, an ATCMT designed to provide text complexity scores that are closely aligned with the reading proficiency constructs targeted by many reading assessments that aid in making high-stakes decisions, including assessments targeted at L1 readers and assessments targeted at L2 readers (Chen & Sheehan, 2015; Sheehan, 2015, 2016; Sheehan, Kostin, Napolitano, & Flor, 2014).

Database of Passages and Items

A database of 582 TOEFL iBT passage sets was assembled for consideration in the analysis. Each set had been included on an operational TOEFL iBT form administered between 2010 and 2015 and included between 12 and 14 items. Because item type classifications were considered at subsequent stages of the analysis, passage sets that did not include a valid type classification for each item were excluded.

Developers of the *TOEFL*® test classify each TOEFL iBT reading item as belonging to one of three main categories: (a) basic comprehension items, (b) inferencing items, or (c) reading to learn items. Basic comprehension items are further divided into five subtypes: vocabulary, fact, negative fact, sentence, and reference. Each of these five subtypes is designed to assess lexical, syntactic, and semantic abilities, along with the ability to understand information presented in single sentences and to connect information across sentences.

Inferencing items differ from basic comprehension items in that students are also required to infer information that is not directly stated in the text but is inferable from the information presented in the text. Items in this category belong to three subtypes: rhetorical items, inference items, and insert sentence items.

Reading to learn tasks are designed to assess additional abilities, such as recognizing the organization and purpose of a text, distinguishing major from minor ideas, and understanding rhetorical functions, such as the text features used by authors to establish cause-and-effect relationships.

Items included on the TOEFL iBT reading section also differ in terms of the number and type of options included with each item. In particular, whereas all of the basic comprehension and inferencing items are presented with a single correct option and three incorrect options, reading to learn tasks are presented with more than four choices and more than one correct answer, allowing for partial-credit scores.

Table 2 Numbers of Items, by Type of Item, for the Original and Text Representation Samples

	Original sample		Text representation sample		
Type of item[a]	Modal number per passage	Total	Modal number per passage	Total	Retained (%)
Basic comprehension					
Vocabulary	4	2,247	3	1,476	66
Fact	4	1,996	3	1,471	74
Negative fact	1	763	1	559	73
Sentence	1	423	1	325	77
Reference	1	31	0	19	61
All basic comp.	10	5,460	7	3,850	71
Inferencing					
Rhetorical	1	730	1	539	74
Inference	1	686	1	447	65
Insert	1	566	1	402	71
All inferencing	3	1,982	2	1,388	70
Reading to learn					
Prose summary	1	582	0	0	0
All items	14	8,024	9	5,238	65

[a]All item type classifications were assigned by TOEFL assessment developers for consideration during test assembly.

Table 2 shows the numbers of items of each type included in the original item pool. In most cases, passages were administered with 10 basic comprehension items, three inferencing items, and one reading to learn item, yielding a total of 14 items for the passage and a total of 8,024 items across all 582 passages.

Analyses focused on selecting an optimal subset of items were conducted in two steps. First, all reading to learn items were excluded, as these items require students to evaluate each of six different options, a requirement which may emphasize response selection processes over text representation processes. Next, the item difficulty parameters obtained for each of the remaining items were compared to passage difficulty estimates generated via TextEvaluator, and nine closely aligned items were selected from each passage. As is indicated in Table 2, most passages were then represented by seven basic comprehension items and two inferencing items, yielding a total of nine items per passage and $9 \times 582 = 5{,}238$ items across all passages.

Generating a *TOEFL®*-to-*TextEvaluator®* Concordance Table

Two difficulty scores were then generated for each passage: a TextEvaluator score (TC_j) and a passage difficulty score ($PC75_j$). TextEvaluator scores were obtained by first extracting more than 100 features known to be indicative of comprehension ease or difficulty and then using that evidence to infer the location of each text on the TextEvaluator scale, a developmental scale that ranges from 100 (indicating that the text is appropriate for beginning readers) to 2000 (indicating that the text is appropriate for advanced, college-level readers).

Passage difficulty scores were obtained by first generating a PCC for each passage and then solving for the reading ability scores needed to achieve an expected proportion correct score of .75. The process of generating a PC75 score for a text is illustrated in Figure 5.

As is shown in Figure 5, each PC75 score is originally expressed on the theta scale, that is, the standardized scale employed during item calibration and form assembly. When presenting assessment results to test takers and other score users, however, it is standard practice to employ a reporting scale that does not include negative numbers. Consequently, a previously estimated theta to scaled score translation function was used to reexpress each PC75 score on the TOEFL iBT reporting scale, which ranges from 0 to 30.

Next, a locally weighted scatterplot smoother (Cleveland & Devlin, 1988) is used to estimate the regression of TextEvaluator scores on PC75 scores. The 582 score pairs employed in the analysis are plotted in Figure 6, along with the resulting smoothed curve. The analysis yielded an estimated standard error of 91 and a correlation of .73, suggesting that reading ability scores expressed on the TOEFL iBT reading scale can be reliably translated into corresponding reading ability scores expressed on the TextEvaluator scale.

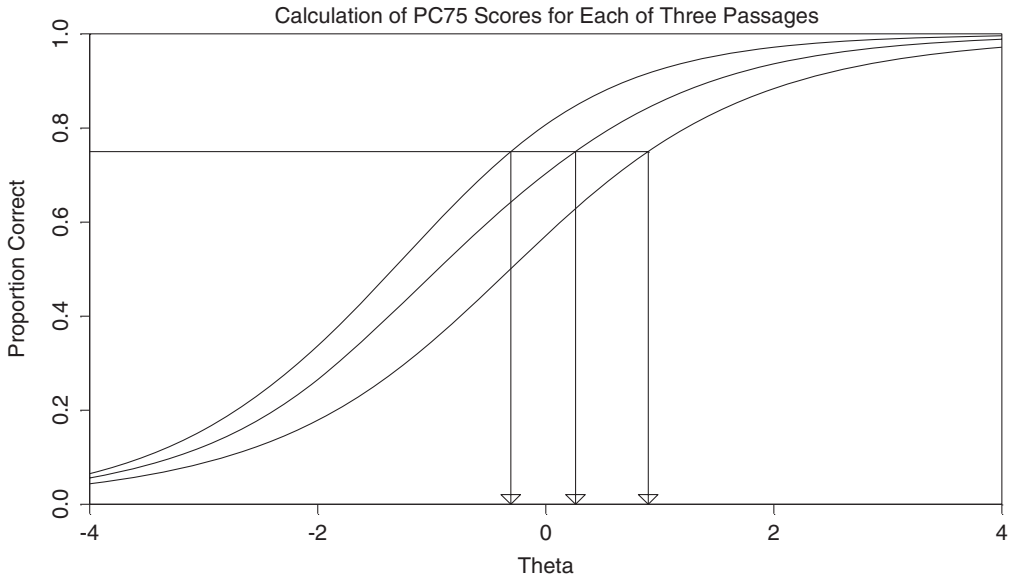

Figure 5 Calculation of PC75 scores for each of three passages. Each curve is estimated from the item response theory item parameters of nine items. Resulting PC75 scores (listed from left to right) are as follows: −.31, .26, and .90.

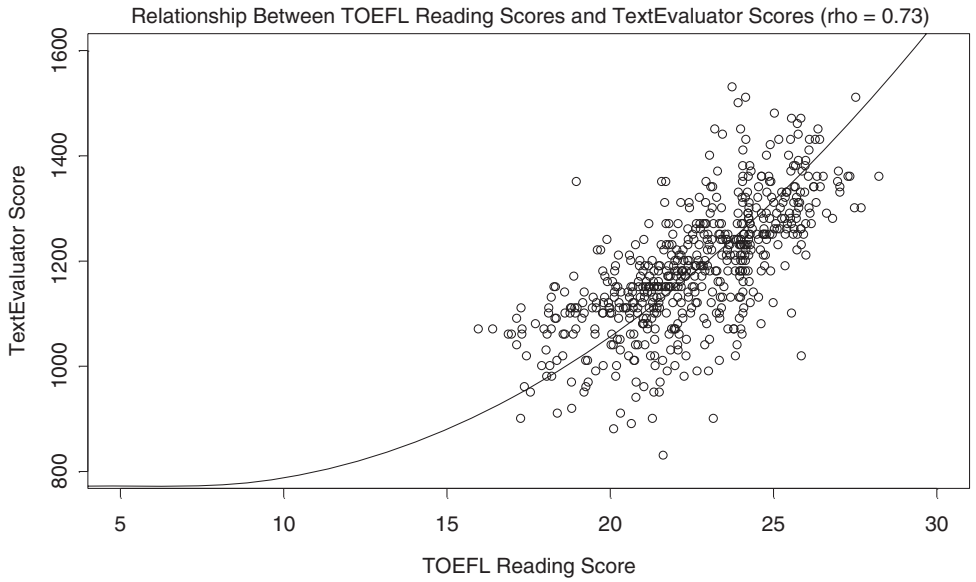

Figure 6 Relationship between TOEFL iBT reading scores and TextEvaluator scores.

Table 3 is a score concordance table estimated from the smoothed regression curve in Figure 6. The table provides an approximate 2 *SD* targeted reading range expressed on the TextEvaluator scale for students with specified TOEFL iBT reading scores. For example, consider a test taker who received a score of 20 on the TOEFL iBT reading section. The estimates in Table 3 suggest that this test taker is likely to be well matched to any text with a TextEvaluator score in the range from 910 to 1110. Practically speaking, this means that the test taker is expected to know the meaning of many, but not all, of the words presented in the texts at this range and is also likely to be familiar with many, but not all, of the sentence- and discourse-level structures found in such texts.

Wendler et al. (2010) evaluated the validity of the Lexile reader–text matching algorithm by subdividing the TOEFL iBT reading scale into six performance levels and then considering the range of Lexile scores associated with each of those levels. Table 4 presents a similar analysis for the reader–text matching algorithm developed in this study. The table shows TextEvaluator score ranges and corresponding grade-level classifications for each of six prespecified TOEFL performance

Table 3 A Concordance Table for Use When Translating Reading Ability Scores Expressed on the TOEFL iBT Reading Scale Into Corresponding Text Complexity Scores Expressed on the TextEvaluator Scale

TOEFL iBT reading score	Expected TextEvaluator score	Recommended range of TextEvaluator scores
0 – 10	Varied	500 – 840[a]
11	800	650 – 850[a]
12	820	670 – 870[a]
13	830	680 – 880[a]
14	860	710 – 910[a]
15	880	730 – 930[a]
16	910	760 – 960
17	940	790 – 990
18	980	830 – 1030
19	1010	860 – 1060
20	1060	910 – 1110
21	1100	950 – 1150
22	1150	1000 – 1200
23	1200	1050 – 1250
24	1260	1110 – 1310
25	1310	1160 – 1360
26	1380	1230 – 1430
27	1440	1290 – 1490
28	1500	1350 – 1550
29	1580	1430 – 1630
30	1600	1450 – 1650

[a]Results obtained via extrapolation.

Table 4 Correspondence Between TOEFL Performance Levels and TextEvaluator Grade Levels

TOEFL Reading performance level	TOEFL iBT reading score	TextEvaluator difficulty score	TextEvaluator grade-level score[a]
High	27 – 30	1440 – 1600	Graduate school
High	22 – 26	1150 – 1380	College
Intermediate	18 – 21	980 – 1100	Grades 10 – 12
Intermediate	15 – 17	880 – 940	Grades 8 – 10
Low	9 – 14	500 – 860	Grades 6 – 8
Low	0 – 8	200 – 500	Grades 4 – 6

[a]Grade-level scores are reported on the accelerated scale specified in the Common Core State Standards. This new scale is structured such that students achieve college and career readiness in reading by the end of Grade 12.

levels. Results can be summarized as follows: Students with TOEFL iBT reading scores at the low performance level are matched to texts with TextEvaluator scores in the range from Grade 4 to Grade 8, students with TOEFL iBT reading scores at the intermediate performance level are matched to texts with TextEvaluator scores in the range from Grade 8 to Grade 12, and students with TOEFL iBT reading scores at the high performance level are matched to texts with TextEvaluator scores in the range from college to graduate school. These results suggest that the proposed matching algorithm may help TOEFL iBT text takers select texts that are well matched to their abilities. When interpreting these results, however, it is important to remember that TextEvaluator grade levels have been adjusted to align with the accelerated text complexity guidelines specified in the Common Core State Standards (Sheehan, 2015). These new, more challenging guidelines are designed to ensure that all students achieve college and career readiness in reading by the end of Grade 12.

Incorporation Within an Automated Text Selection Application

The targeted text complexity score ranges generated via the proposed approach will only be useful for test takers if texts that have been scored via the selected ATCMT are readily available. Researchers at ETS are in the process of developing a Web site that addresses this need. The Web site will provide TextEvaluator scores for books and articles likely to be of interest to readers in elementary, secondary, and college courses. Sample books, with corresponding TextEvaluator scores, are shown in Table 5 (fiction) and Table 6 (nonfiction). A much larger set of titles, including a large number of books selected

Table 5 TextEvaluator Scores for a Sample of Fiction Titles

TextEvaluator score	Book
195	*The Stories Julian Tells* by Ann Cameron
530	*The Lighthouse Family: The Storm* by Cynthia Rylant
590	*M. C. Higgins the Great* by Virginia Hamilton
600	*The Little Prince* by Antoine de Saint-Exupery
675	*P.S. I Still Love You* by Jenny Han
680	*Bud, Not Buddy* by Christopher Paul Curtis
680	*Roll of Thunder, Hear My Cry* by Mildred D. Taylor
710	*A Wrinkle in Time* by Madeleine L'Engle
710	*Dragonwings* by Lawrence Yep
755	*Me and Earl and the Dying Girl* by Jesse Andrews
760	*The Secret Garden* by Frances Hodgson Burnett
760	*Tuck Everlasting* by Natalie Babbitt
770	*The Girl on the Train* by Paula Hawkins
845	*Alice's Adventures in Wonderland* by Lewis Carroll
850	*The Dark Is Rising* by Susan Cooper
990	*The Adventures of Tom Sawyer* by Mark Twain
1030	*Little Women* by Louisa May Alcott

Table 6 TextEvaluator Scores for a Sample of Nonfiction Books

TextEvaluator score	Book
575	*My Librarian Is a Camel* by Margriet Ruurs
585	*A Long Walk to Water* by Linda Sue Park
600	*We Are the Ship: The Story of Negro League Baseball* by Kadir Nelson
625	*A History of US* by Joy Hakim
640	*Quest for the Tree Kangaroo* by Sy Montgomery
670	*Math Trek: Adventures in the Math Zone* by Ivars Peterson
725	*Toys! Amazing Stories Behind Some Great Inventions* by Don Wulffson
819	*I Am Malala* by Malala Yousafzai, with Patrick McCormick
840	*Harriet Tubman: Conductor on the Underground Railroad* by Ann Petry
980	*Freedom Walkers: Story of the Montgomery Bus Boycott* by Russell Freedman
990	*A Night to Remember* by Walter Lord
1120	*Vincent Van Gogh: Portrait of an Artist* by Jan Greenberg and Sandra Jordan
1185	*The Wright Brothers* by David McCullough
1200	*The Life-Changing Magic of Tidying Up* by Marie Kondo
1205	*And the Good News Is …* by Dana Perino
1220	*Hard Choices* by Hillary Clinton
1300	*Living History* by Hillary Clinton

from the Project Gutenberg corpus ($n \cong 700$) and thousands of articles selected from Wikipedia ($n \cong 60{,}000$) and Simple Wikipedia ($n \cong 40{,}000$), is also being prepared for inclusion on the proposed site. Figure 7 provides a high-level overview of this new application. Four key components are highlighted: the score entry module, the score translation module, the text analysis module, and the text selector module. Test takers who choose to access the new application will be able to select texts that are well matched to their abilities, a strategy that could lead to increased confidence and higher levels of reading proficiency.

Discussion

Assessment publishers have long been challenged to provide additional information about the meaning of test scores. This report introduced a powerful new approach for addressing that challenge. The approach combines evidence extracted from test takers' observed item responses with information developed from a large collection of previously administered reading passages and information obtained via an automated analysis of the observable features of those passages to provide a quantitative description the types of texts that test takers at different score levels are expected to be able to comprehend.

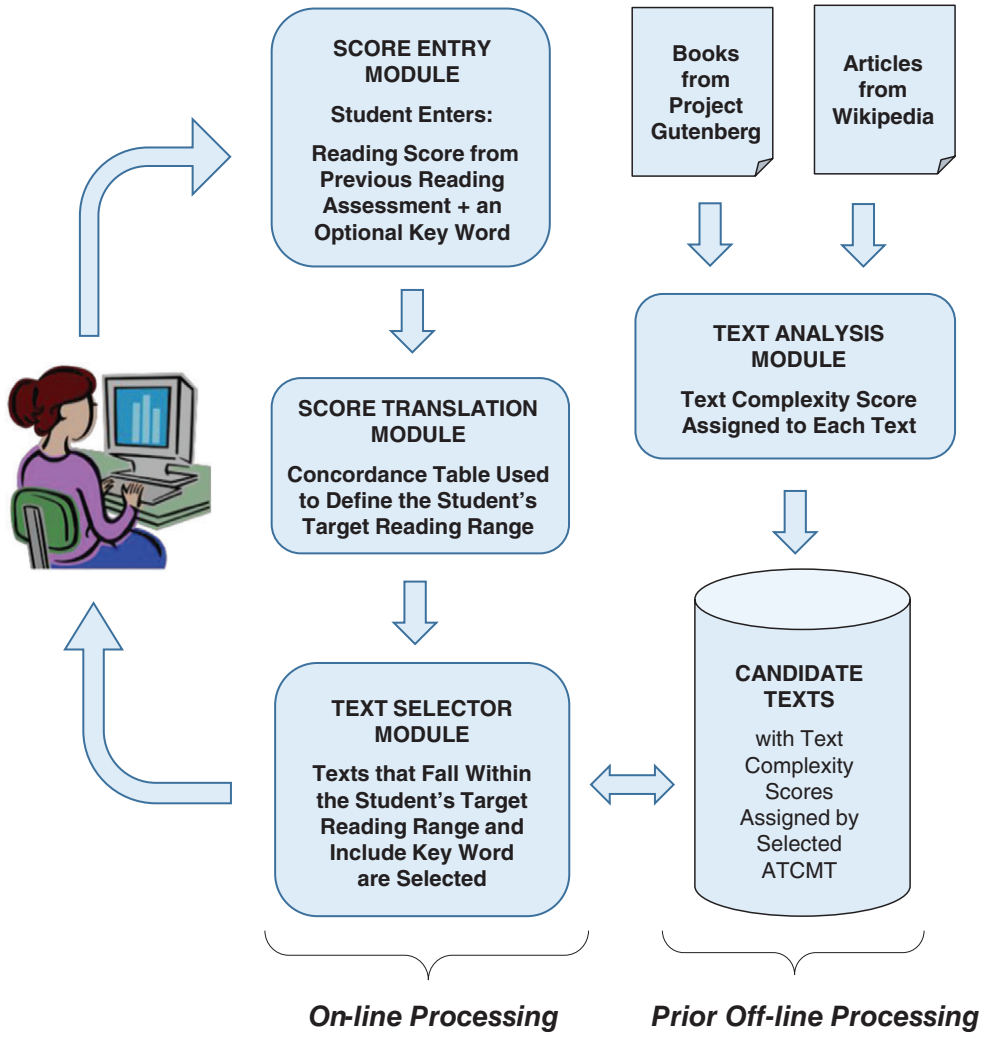

Figure 7 Architecture of the reader–text matching system.

Resulting information is intended to help both test takers and test score users. Test takers can use the information to select texts that are expected to be well matched to their abilities, that is, texts that are expected to be challenging, yet not too challenging. This information may also help test score users understand the types of reading skills that students at different score levels are likely to have mastered.

Although additional research focused on the measurement properties of this new approach is clearly needed, the analyses summarized in this report suggest that the approach offers a number of advantages over existing technologies. One key advantage is that a separate linking test is not required, so the time and expense associated with developing, administering, and scoring a separate linking test are avoided. The strategy of not administering a separate linking test also means that the resulting concordance table will not be subject to the types of biases that may arise when linking tests are administered to self-selected samples of test takers, as is the case under the indirect approach (Pommerich, Hanson, Harris, & Sconing, 2000).

An additional advantage is that the strategy of incorporating text complexity evidence obtained from a large number of passages and items from numerous previous administrations of the selected assessment means that a larger, more diverse set of passages can be included in the analysis, and evidence about the difficulties experienced by test takers when reading those passages can be based on the responses provided by a larger, more diverse set of test takers.

Third, because passage complexity is evaluated using TextEvaluator, resulting estimates may be more closely aligned with cognitive theories of how readers process text and thus may be more effective at distinguishing passages that are likely to be more or less challenging for test takers (Chen & Sheehan, 2015; Sheehan, 2015, 2016; Sheehan et al., 2014).

One limitation of the proposed approach should also be mentioned. In particular, the results in Figure 6 confirm that there can be many test takers whose reading ability scores fall below the 75% cutoff score adopted throughout the procedure. Consequently, the data needed to establish a concordance between students' scores on the TOEFL iBT reading assessment and TextEvaluator scores are not available. This issue could be addressed by providing recommendations based on expert judgment or by administering an additional, less demanding set of passages to the lowest scoring students. Additional research focused on these and other options is needed.

Acknowledgments

I am grateful to Mary Schedl for many useful discussions about the reading skills tapped by different types of TOEFL iBT reading items, to Carmen Parker for assisting with data collection tasks, to Diane Napolitano for assisting with TextEvaluator analyses, and to the Metametrics Corporation for providing a collection of raw proportion correct scores collected as students read one or more of 372 passages within the Oasis platform (now called Edsphere).

Notes

1 See http://www.lexile.com/toefl/

2 The proportion correct scores summarized in Figure 3 were distributed by the Metametrics Corporation as part of a research study "undertaken in support of the Common Core State Standards' emphasis on students reading text of appropriate complexity" (Nelson et al., 2012, p. 5). Seven groups who had developed text analysis tools participated in the study. Each group "committed to offering transparency in revealing both the text features it analyzed and the general means of analysis" (p. 5). Consistent with that commitment, a collection of 372 texts administered via the Edsphere platform were distributed. The selected texts included all informational passages that had been read by at least 50 different students and had response data for at least 1,000 computer-generated items. The data summarized in Figure 3 are the raw proportion correct scores collected from students as they responded to those items.

References

Chen, J., & Sheehan, K. M. (2015). *Analyzing and comparing reading stimulus materials across the TOEFL family of assessments* (TOEFL iBT Research Report No. 26). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12055

Cleveland, W. S., & Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association, 83*, 596–610.

Dorans, N. J. (1999). *Correspondences between ACT and SAT I scores* (Report No. 99-1). New York, NY: College Entrance Examination Board.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension. *Applied Psychological Measurement, 11*, 175–193.

Enright, M. K., & Schedl, M. (2000). *Reading for a reason: Using reader purpose to guide test design* (TOEFL internal report). Princeton, NJ: Educational Testing Service.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*, 351–373.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*, 394–411.

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.

International Reading Association. (2004). *Standards for reading professionals*. Newark, DE: Author.

Kirsch, I. (2001). *The International Adult Literacy Survey (IALS): Understanding what was measured* (Research Report No. RR-01-25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2001.tb01867.x

Lattanzio, S. M., Burdick, D. S., & Stenner, A. J. (2012). *Ensemble Rasch models*. Durham, NC: Metametrics.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Messick, S. (1987). Large-scale educational assessment as policy research: Aspirations and limitations. *European Journal of Psychology and Education, 2*, 157–165.

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.

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2000). *Issues in creating and reporting concordance results based on equipercentile methods* (Research Report No. 2000-1). Iowa City, IA: ACT.

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. https://doi.org/10.1002/ets2.12068

Sheehan, K. M. (2016). *A review of evidence presented in support of three key claims in the TextEvaluator validity argument* (Research Report No. RR-16-12). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12100

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*, 184–209.

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

Stenner, A. J., Fisher, W. P., Stone, M. H., & Burdick, D. S. (2013). Causal Rasch models. *Frontiers in Psychology, 4,* 536–564.

Swartz, C. W., Burdick, D. S., Hanlon, S. T., Stenner, A. J., Burdick, H., & Kyngdon, A. (2014). Toward a theory relating text complexity, reader ability, and reading comprehension. *Journal of Applied Measurement, 15*, 359–371.

Wendler, C., Cline, F., Sanford, E., & Aguirre, A. (2010). *Linking TOEFL scores to the Lexile measure.* Paper presented at the Language Testing Research Colloquium, Cambridge, UK.

Suggested citation:

Sheehan, K. M. (2017). *Helping students select appropriately challenging text: Application to a test of second language reading ability* (Research Report No. RR-17-33). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12160

Action Editor: Donald Powers

Reviewers: Ikkyu Choi and Michael Flor

ETS, the ETS logo, MEASURING THE POWER OF LEARNING., TEXTEVALUATOR, TOEFL, and TOEFL iBT are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/