PARE

Practical Assessment, Research, and Evaluation

A peer reviewed, open-access electronic journal: ISSN 1531-7714

Text Complexity Versus Task Complexity: Item Difficulty Modeling for Reading Items

M. Christina Schneider, *Cambium Assessment, Inc.*
Jing Chen, *Cambium Assessment, Inc.*
Jeremy Heneger, *ACT*

Abstract: This study investigates item features to aid in improving understanding of what makes items that measure reading comprehension easy or difficult. In this item difficulty modeling (IDM) study, item and passage features were included as predictors that represented text-task interactions and stimulus demands. The passage-level features included two common quantitative metrics of text complexity: the Lexile Framework® for Reading and Flesch-Kincaid. Passage word count, item type, Depth of Knowledge (DOK), and item to Range Achievement-Level Descriptor (RALD) match were held constant across conditions. Two IDM models were examined; one included passage-level text complexity features and not grade level, and the other included grade level and not passage level text complexity features. We found that quantitative metrics of text complexity added 3% to the IDM compared to when grade was substituted for those features. Text-task interactions as represented by RALDs and DOK levels were found to provide unique and significant information to the IDM model as did item type and particular standard topics. Implications for RALD construction and additional research related to RALDs for reading are discussed.

Keywords: Achievement level descriptors, Reading comprehension, Item difficulty modeling, Range achievement level descriptors, Test score interpretation and use

Introduction

Range Achievement Level Descriptors (RALDs; Egan et al., 2012) are intended to provide descriptions of increasing knowledge, skills, and abilities that are expected of students based on a state's content standards along a test score scale. RALDs are intended to provide content-based interpretations regarding what scale scores within a particular achievement level represent. Examples of RALDs for reading can be found online for many assessment programs, including Florida (FDOE, 2016), Georgia (GaDOE, 2025), Nebraska (NDOE, 2021), Texas (2024), and assessment collaboratives, such as Smarter Balanced (CTB/McGraw Hill, 2013). These RALDs, however, have different characteristics based on test developers' theories of reading development.

Reading is complex and requires a framework, such as RALDs, to guide test developers regarding what evidence to elicit to determine whether a student is a novice, proficient, or advanced reader. Having a framework teachers (and test developers) can use to match the appropriate tasks to the student's current stage of proficiency may be useful because vague guidance is found in the reading literature regarding how teachers should individualize reading comprehension tasks for different students during instruction and assessment (Hiebert & Pearson, 2014).

As a component of a principled assessment test design process (Huff et al., 2025), there is a growing set of converging theories regarding how RALDs and the items matched to them should function together. RALDs and the tasks that exemplify these levels need to be based on a theory of learning or cognition during the test development process. RALDs should describe the hypothesized trajectory regarding how students develop their abilities along the test scale to support collecting purposeful evidence of students' development of proficiency as described in the content standards (Perie & Huff, 2016). To date, however, there have not been studies showing that items matched to RALDs are useful in predicting item difficulty. Moreover, there are competing theories about the role text complexity plays in reading comprehension which has implications regarding how RALDs should be developed, especially for state accountability assessments. Therefore, we first describe the competing theories regarding the position of text complexity in the context of measuring reading comprehension. Next, we describe studies of item difficulty modeling (IDM) centered in findings from the reading literature. Finally, we situate our study within this theoretical and empirical context to examine how RALDs and other item level features used in the construction of reading comprehension items predict item difficulty when text complexity features are included and excluded from IDM models for a state accountability assessment.

Competing Theories of the Relationship between Text Complexity and Reading Comprehension

Reading is multidimensional in nature (Valencia et al., 2017), and there is disagreement in the field on the role that text complexity plays when measuring student learning. Rowe et al. (2006) captured the disagreement succinctly. Is reading comprehension measuring "how difficult a passage readers can comprehend," or "how difficult a question readers can answer about a passage?"

The relationship between the reader and the text in terms of how tasks (i.e., questions) elicit student reasoning about what was read (i.e., how difficult a question readers can answer about a passage) should be described in the RALDs. One of the first decision points test developers must make when creating RALDs is the degree to which text complexity will play a role in describing the tasks that show ability development along the test scale. Figure 1 shows a brief example of a RALD that exemplifies one way some states create them. As shown, text complexity was built into the sample RALD as the single contextual feature explaining how items (and student proficiency) are expected to differ along the test scale. The verbs (reflecting cognitive complexity) and the descriptions of the content (characters and the sequence of events) are written the same across levels; thus, we infer RALDs of this variety are designed to answer the question, "How difficult a passage can readers comprehend?"

Toyama et al. (2017) defined text complexity as the inherent property of text, which comprises a set of linguistic and discourse features. Text complexity is described in standards, such as the Common Core State Standards (CCSS) for English language arts (ELA) (National Governors Association Center for Best Practices & Council of Chief State School Officers [CCSSO], 2010). It is a component of assessment frameworks such as the *Reading Framework for the 2026 National Assessment of Educational Progress* (NAEP; National Assessment Governing Board, 2023). Nebraska's ELA standards assert that students will "listen to and read text of increasing length and/or complexity to increase reader stamina" and "construct meaning by using prior knowledge and text information while reading grade-level literary and informational text" (Nebraska Department of Education, 2014, p. 19). Toyama et al. wrote that text complexity indices, such as the Lexile Framework® for Reading (Lexiles; Stenner et al., 2006) and the Flesch-Kincaid Grade Level

(Kincaid et al., 1975), often serve as predictors of text difficulty, whereas text difficulty is measured by an index derived from student reading comprehension performance or expert judgments of the difficulty students are expected to have with particular texts.

Figure 1. Example RALD Centered in Text Complexity

Levels	Standard: Describe characters in a story (e.g., traits, motivations, feelings) and explain how their actions contribute to the sequence of events.
Level 1	Students in Level 1 describe characters in a story (e.g., traits, motivations, feelings) and explain how their actions contribute to the sequence of events in texts of low complexity.
Level 2	Students in Level 2 describe characters in a story (e.g., traits, motivations, feelings) and explain how their actions contribute to the sequence of events in texts of moderate complexity.
Level 3	Students in Level 3 describe characters in a story (e.g., traits, motivations, feelings) and explain how their actions contribute to the sequence of events in texts of moderate-to-high complexity.
Level 4	Students in Level 4 describe characters in a story (e.g., traits, motivations, feelings) and explain how their actions contribute to the sequence of events in texts of above-grade high complexity.

College readiness assessments also note the role of text complexity in reading, albeit differently. The SAT and ACT have different theories of action and processes regarding how reading is measured. The SAT defines its construct in its test specifications as the ability to "read, analyze, and use reasoning" to comprehend texts (College Board, 2015). SAT test specifications separate reading comprehension from reasoning about difficult texts and explain how test developers use "feedback from secondary and postsecondary subject-matter experts and test data on student performance as well as quantitative and qualitative measures of text complexity" to determine the difficulty of the passage (p. 27). On the other hand, the ACT (2017) uses a qualitative text complexity rubric and shows student skills growing along the test scale based on the tasks the student can perform explicitly based on the level of text complexity.

Conflicting Findings on the Relationship between Text Complexity and Reading Comprehension in IDM studies

Researchers (e.g., Ferrara et al., 2022) engage in item difficulty modeling (IDM) to understand cognitive processes and as a tool for validating theories regarding what makes items easier or more difficult. Because items and students are typically on the same scale in the measurement models used for large-scale standardized assessments, the item content can be used to infer how students differ in their cognition across the test scale. Studies seeking to understand the difficulties of items associated with reading passages are centered on task analyses, meaning the researcher's purpose is to make explicit increasing levels of competencies related to higher levels of test performance to provide feedback to item writers, teachers, and learners. For example, Embretson and Wetzel (1987) found that item difficulty was associated with comparing and matching information in passages to response options, and the use of low-frequency words (a surrogate for vocabulary sophistication) in stems and options was a feature that predicted item difficulty.

IDM studies have produced conflicting results regarding how well text complexity predicts item difficulty. Ferrara et al. (2022) reviewed more than 100 IDM studies in reading. Thirteen studies explained at least 50% of the variance in item difficulty (Alderson et al., 2009; Drum et al., 1981; Freedle & Kostin, 1992; Gorin & Svetina, 2008; Kirsch & Mosenthal, 1990; Lumley et al., 2012; Mosenthal, 1996; Mosenthal, 1998; Sano, 2016; Scheuneman et al., 1991; Sheehan & Mislevy, 1990; Valencia et al., 2017; Toyama, 2019). In these studies, 7 passage-level features predicted item difficulty. However, only four of these studies investigated items administered to students in grades that ranged from Grade 3 through 8 (Drum et al., 1981; Sano, 2016; Valencia et al., 2017; Toyama, 2019). Two of these studies used NAEP data (Sano, 2016;

Valencia et al., 2017). Ferrara et al. also noted that only two of the 13 IDM studies cross-validated their results (Sano, 2016; Toyama, 2019). This means the variance of the item difficulties explained in 11 studies were taken from the training sets of the data, and the ability to explain 50% or more of the reason item difficulties varied in these studies is likely inflated.

Valencia et al. (2017) found that the average passage difficulty, an index derived by averaging the student reading comprehension item parameters from a passage, was not significantly correlated to two different readability formulas in a within-grade analyses of Grade 4 and Grade 8 items. They investigated the Lexile (Stenner et al., 2006) and TextEvaluator (Napolitano et al., 2015) formulas for text complexity. Ferrara and Steedle (2015) found that, while text complexity tended to increase with grade level, the correlation between text complexity and item difficulty, as measured based on *p*-values (the proportion of students who respond correctly to an item), was often negative and small in within-grade analyses of Grade 3–8 items. Text complexity explained only 1% of the variation in item difficulty. On the other hand, command of textual evidence (i.e., the amount of text that an examinee must process to respond correctly to a task) was found to be a significant, albeit weak, predictor of item difficulty. Similarly, Sheehan and Ginther (2001) found that item difficulty was related to the frequency and location of the target information in a passage. Thus, it appears that researchers may need to attend to the influence of text complexity in item prediction studies that appear across grades verses within grades.

Text Complexity Implications for RALD Development

We contend that providing teachers with feedback about the extent to which students compare response options to a passage does not generalize to the kind of evidence teachers observe in real time for formative use as students are reading. Instead, emphasizing how students match information in passages to response options may inadvertently promote test preparation rather than authentic reading development. Therefore, it is essential for test developers to plan assessment tasks that meaningfully generalize to the kinds of tasks students encounter outside of the testing context in the development of RALDs.

The RAND Reading Study Group (2002) and Valencia et al. (2014, 2017) argued that reading is centered on the task (the question), and text complexity is an insufficient determinant of reading comprehension. Valencia et al. critiqued many users, and even the developers of the CCSS, for overemphasizing the role of text complexity in reading comprehension. Instead, they posited that reading is found at the intersection of the task, text, and reader. They proposed that the primary focus of understanding reading comprehension must be on the task the students are asked to perform. This is consistent with the intended implementation of the RALD framework developed by Egan et al. (2012), which describes creating RALDs that delineate the conditions under which students "approach and process" content (i.e., the context), the difficulty of the content, and the cognitive complexity of the task as student performance increases across achievement levels.

We propose that, for reading comprehension, RALDs should describe both the task difficulty (e.g., a conclusion drawn from explicit or implicit information) and how the author sets up the reader to draw conclusions from the text within a particular grade. Valencia et al. (2017) posited that the explicitness, location, and amount of the target information in the text for a particular task influenced an item's difficulty. This is the text-task interaction. Easier and more complex texts often offer similar opportunities to ask questions of readers. The text-task interaction provides evidence that the reader is in a particular stage of reading development within a particular grade level (Valencia et al., 2014). Valencia et al.'s "Text-Task-Scenario" is a framework in which three features are hypothesized to interact to support reading comprehension: task features (i.e., what the student is asked to do), features of readers (e.g., engagement or reading ability), and text features (i.e., the attributes embedded in the text that support the student responding correctly to the task). They argued that the reading purpose, text type, and task features together form a learning progression.

We posit that task features can and should include the item types used to collect information. For example, item type (e.g., polytomous or technology-enhanced) has been found to predict item difficulty (Wools et al., 2019). If students have more reading ability, they notice and track more details in separate sections of evidence from the text. Their more advanced ability, for example, allows them to select multiple pieces of evidence across paragraphs within a passage to support an interpretation. An item type such as technology-enhanced that allows the student to select multiple pieces of evidence supports a test developer's ability to measure this more advanced skill. Thus, the technology enhanced item type may predict a more difficult task. The hypothesis for this study is that, when texts are appropriately complex for students within a grade level, the task-text interaction (i.e., RALDs that embed task-text interactions into the descriptors) is a significant predictor of item difficulty along the ability continuum, with literal comprehension representing more novice stages of understanding and analysis and reasoning across a passage representing more sophisticated stages. That is, we have a hypothesis that readability formulas that are surrogates of text complexity do not explain variance in item difficulty in an accountability assessment when the grade that represents the typical readability ranges are included in the IDM model. We also expect Depth of Knowledge (Webb, 2005) to be a significant predictor of item difficulty.

Purpose of Study

The purposes of this study follow:

1. Evaluate if RALD-to-task match is a significant predictor of task difficulty.
2. Evaluate if text complexity features are a significant predictor of task difficulty compared to the grade-level alignment.

To address our study purposes, we compare two IDMs. The first model includes the same predictors but excludes text complexity predictors. Instead, it includes the grade level to which each item is aligned. The second model includes the same predictors but excludes grade level. Instead, it includes text complexity predictors using proxy reading formulas and word count.

Test Design Process

A principled assessment design (PAD) approach was used to create the items for this study based on RALDs (Huff et al., 2016; Egan et al., 2012). The state from which these reading items are drawn implemented an amalgamated test design process (Forte, 2017), which involved the creation of RALDs after the adoption of new achievement-level standards. After the cut scores were established using threshold ALDs (descriptions of students at the borderline of an achievement level), teachers and item writers collaborated to construct RALDs based on the recommended cut scores using 80 items per grade so that state could publish Range as Reporting ALDs (Michaels et al., 2018) and align future passage development and item writing to this framework. Additional items were field tested, aligned to the RALDs, and added to the pool for a computer-adaptive assessment from 2018–2020, thereby ensuring the blueprint for the new assessment was met.

Passages were reviewed by item writers for potential tasks that aligned to the RALDs at different achievement levels without the use of task templates because item writers felt this supported their work efficiently (M. Veazey, personal communication, August 2020). Item writers typically developed 10 items per passage. The state used Level 1 to denote students in earlier stages of college and career readiness in reading comprehension and Level 3 to denote students in more advanced stages of college and career readiness in reading comprehension, as shown in the Sample RALD in Figure 2. Within the standard, the RALD focuses on the cognitive complexity of each task and the location of evidence.

Figure 2. Sample RALD from a Midwestern State

Text Complexity	**With a range of texts with text complexity commonly found in Grade 3, a student performing in Level 1 can likely**	**With a range of texts with text complexity commonly found in Grade 3, a student performing in Level 2 can likely**	**With a range of texts with text complexity commonly found at the intersection of Grade 3 and Grade 4, a student performing in Level 3 can likely**
Standard. Identify and describe elements of literary text (e.g., characters, setting, plot, point of view).	Identify a basic element of literary text (e.g., characters, setting).	Identify and describe elements of literary text (e.g., characters, setting, plot, point of view).	Analyze elements of literary text (e.g., characters, setting, plot, point of view) from across the passage to support a thorough understanding of the text.

Method

Outcome and Predictor Variables

The data for this study came from a summative large-scale assessment from a midwestern state. As is common in K–12 state assessment, reading items were calibrated and assumed to be conditionally independent even when related to a common stimulus. Psychometricians used the Rasch model (Rasch, 1960) for dichotomous items and the partial credit model (PCM; Masters, 1982) for polytomous items. The vertical scale was created by concurrently calibrating items across grades with the item mean centered at 0 with a standard deviation of 1. Item parameter estimations were implemented using WINSTEPS 3.91.0.0 (Linacre, 2015) using joint maximum likelihood estimation (JMLE), as described by Wright and Masters (1982), with item score data from a representative sample of students. The step parameters for the polytomous items were averaged to provide a single difficulty parameter to be comparable to dichotomous items. A benefit of using item response theory (IRT) models to calibrate items is that student ability and item difficulty are estimated on the same continuum.

Task features were the predictors in the IDMs. Most task features were available in the item banking software; however, passage length (i.e., word count), Flesch-Kincaid, and Lexile were derived for this study. The nine item features used in the study and the IDM in which the features were used are given in Table 1.

The state designed its standards with 12 topics in common across grades. Examples of topics include common expectations for reading such as applying context clues to infer the meaning of unknown words, identifying author's purpose, and describing elements of literary text. The item pool comprised 1,493 items, of which 89% were multiple choice items, 6% were gap match items, and 5% were choice multiple items. Five percent of items were polytomous. Choice multiple items are sometimes referred to as multi-select item types. They allow students to select more than one response for a task. For a gap-match item, a student selected or used a drag-and-drop feature to use text to support an analysis or make an inference. The RALD and DOK levels of each item were coded by subject-matter experts as a component of the test development process. Table 2 presents the number of items included in the study by RALD level. For DOK, 18% of items were DOK 1, 68% were DOK 2, and 14.1% of the items were DOK 3 across grades. For RALD, only 8.8% of the items were at the highest level across grades.

Reading vocabulary comprised 24% of items, and reading comprehension comprised 76% of items. Text complexity was estimated using two measures: the Flesch-Kincaid Grade Level (Kincaid et al., 1975) and the Lexile Framework® for Reading (MetaMetrics, 2007). The Flesch-Kincaid Grade Level indicates the grade level a student would need to be in to comprehend the passage's material. For example, if the Flesch-Kincaid passage grade level is around 2.0, an average student in Grade 2 can read the text. The Flesch-Kincaid grade level uses the average number of words per sentence and the average number of syllables per word to estimate the grade level. The score typically ranges from 0 to 18. Lexile measures the complexity of the text using features such as sentence length and word frequency. Stenner et al. (2006) noted that word frequency is generated using a dictionary based on a 550-million-word corpus. Generally, longer sentences and words of lower frequency lead to higher Lexile measures, whereas shorter sentences and higher frequency words lead to lower Lexile measures.

Table 1. Item features used in the IDMs

Predictor	Explanation	Model in which Predictor was Used
Grade level	The grade to which an item was aligned that ranged from Grade 3 to Grade 8	Model 1
Standard	Identifier for the topic that the state held in common across grades	Model 1; Model 2
Item type	multiple choice, gap match, choice multiple	Model 1; Model 2
DOK level	Depth of Knowledge (Webb, 2005) based on a framework of cognitive complexity commonly used in state assessment programs that in practice ranges from 1–3	Model 1; Model 2
ALD level	The achievement descriptor level to which an item was aligned using the state's RALDs (Egan et al., 2012).	Model 1; Model 2
CCSSType	A dummy variable that indicated if the passage was literary or informational	Model 1; Model 2
Passage length	Number of words in a passage	Model 2
Flesch-Kincaid	Score representing the grade level a student would need to be in to comprehend the passage's material (Kincaid et al., 1975)	Model 2
Passage Lexile score	complexity of the text using features such as sentence length and word frequency (MetaMetrics, 2007)	Model 2

Table 2. Number of Items by Grade and RALD Level

	Number of Items by RALD Level							
	Level 1		Level 2		Level 3		Total	
Grade	N	%	N	%	N	%	N	%
3	93	22.3	289	69.3	35	8.4	417	100.0
4	29	14.2	161	78.9	14	6.9	204	100.0
5	74	33.5	121	54.8	26	118	221	100.0
6	64	31.2	121	59.0	20	9.8	205	100.0
7	64	33.7	111	58.4	15	7.9	190	100.0
8	83	32.4	152	59.4	21	8.2	256	100.0
Total	407	27.3	955	64.0	131	8.8	1493	100.0

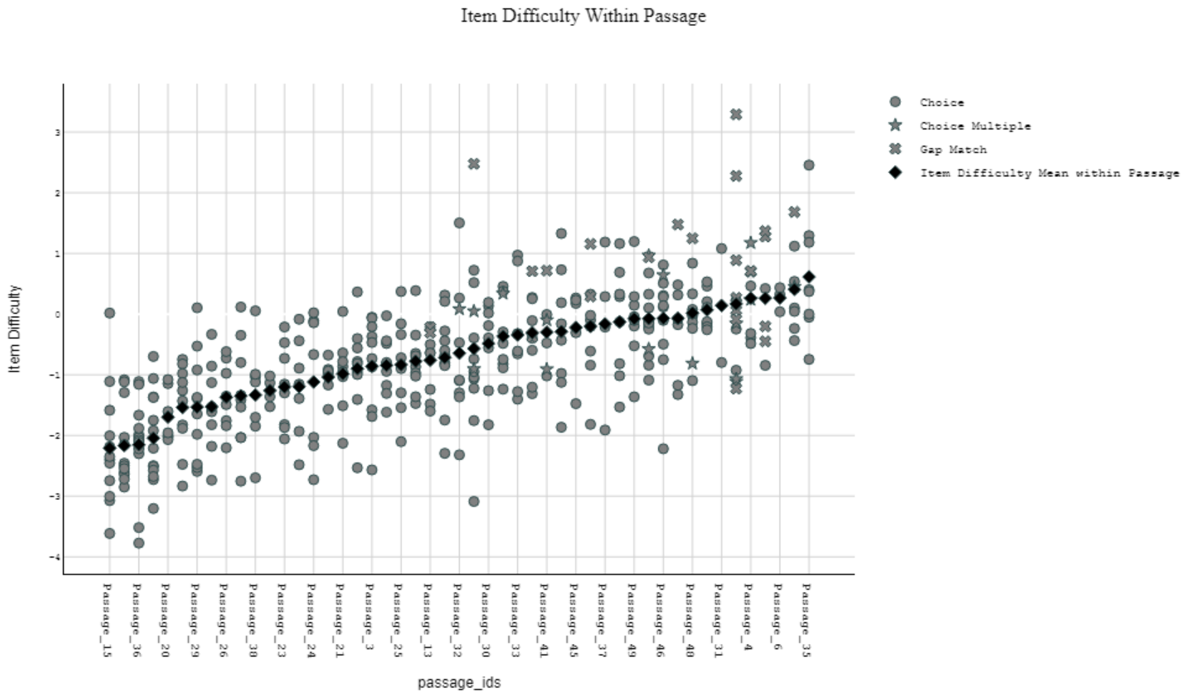
The reading items included in this study were from 180 passages. The Flesch-Kincaid Grade Level was calculated for each passage, and Lexile measures were determined by the Lexile Analyzer provided by MetaMetrics, which approximates official Lexile scores. A Lexile score indicates the reading difficulty of a text and the reading ability of a student because both are measured on the same Rasch scale (Stenner et al., 2006). Stenner et al. wrote that when a student's Lexile matches a text's Lexile, the student is expected to comprehend approximately 75% of the text. Table 3 presents the descriptive statistics for text complexity metrics and passage length by grade, as well as the CCSS-recommended Lexile range proposed by MetaMetrics (n.d.). Lexile scores fall within the CCSS-recommended Lexile range at each grade level. As grade level increases, the averages of Flesch-Kincaid scores, Lexile scores, and passage lengths also increase. However, the maximum Lexile is the same for Grades 7 and 8, and the maximum word count does not always increase as grade level increases.

Table 3. Descriptive Statistics of Passage Readability Measures by Grade

Variable	Grade	N	Mean	SD	Min.	Max.	CCSS Lexile Lower Bound	Range Upper Bound
Flesch-Kincaid	3	50	3.7	1.1	0.3	6.1	NA	NA
	4	24	4.9	1.3	1.2	7.0	NA	NA
	5	26	5.7	0.9	3.5	8.3	NA	NA
	6	25	7.1	1.1	5.0	10.4	NA	NA
	7	24	7.5	1.1	6.1	9.5	NA	NA
	8	31	8.0	0.9	6.1	9.7	NA	NA
Lexile	3	50	633.6	73.3	520	820	520	820
	4	24	843.3	61.0	740	940	740	940
	5	26	901.5	51.8	830	1000	830	1010
	6	25	993.2	44.0	940	1070	925	1070
	7	24	1045.4	50.8	970	1120	970	1120
	8	31	1053.2	36.0	1010	1120	1010	1185
Passage Length (#Words)	3	50	562.6	164.4	219	1103	NA	NA
	4	24	624.6	273.6	173	1442	NA	NA
	5	26	651.9	236.1	186	1424	NA	NA
	6	25	719.6	157.8	285	1006	NA	NA
	7	24	803.0	208.8	267	1195	NA	NA
	8	31	835.6	249.3	299	1688	NA	NA

Item Difficulty Within Passage. Figure 3 presents the difficulties of items within each passage at one grade level (Grade 3) as an example of the range in item difficulties observed within passages. The item difficulties are based on the Rasch model (Rasch, 1960; Masters, 1982). Dots on the same vertical line represent difficulties of items from the same passage. The dark diamond represents the average difficulty of the items within each passage (or overall passage difficulty). The grey circles, stars, and Xs represent the item difficulties for multiple choice, choice multiple, and gap match items, respectively. The passages are ordered by passage difficulty from lowest to highest on the x-axis.

Figure 3. Item Difficulty Within Passage



From Figure 3, we can see that item difficulties are related to the passage to some extent (i.e., passages on the left side tend to have easier items and passages on the right side tend to have harder items). Items below -.32 are Level 1. Items at or above .79 are Level 3. There is large variation in item difficulty within most passages. Another pattern that can be seen from the figure is that gap match items and choice multiple items seem to be more difficult than multiple choice items when the step values are averaged for analysis purposes.

RALD, DOK, and grade-level variables were treated as ordinal variables with ordered categories from low to high. The targeted standard, item type, and passage type features were treated as dummy variables. For example, an item that assessed a particular standard was coded as 1 for that standard and 0 for all other standards. An item that belonged to a particular passage type was coded as 1 for literary and 0 for informational. Text complexity scores were continuous variables. After all the dependent and independent variables were derived, correlations and multicollinearity among variables were inspected, model assumption checks were implemented, and IDMs were fit to the data.

Item Difficulty Modeling

An Ordinary Least Squares multiple linear regression model was applied to build each IDM. This model was chosen for its relative ease in interpretation because our goal was testing how RALDs and text complexity proxies are associated with item difficulties based on a theory of action. Moreover, we have found this model is easier to explain to non-statisticians such as item writers. Because predictors were not fully crossed, we did not include interaction terms. We modeled the regression across grades because items were on a vertical scale, and because within grades (e.g., Grade 7), the ratio of the number of items to number of predictors would be too small given the recommendation of a 15:1 ratio (Stevens, 2009).

The items were split randomly into training and validation datasets, with 80% of the items in the training dataset and 20% of the items in the validation dataset. The random splitting process was repeated 10 times for each IDM. Each time, the model was built and validated using the randomly split training and validation datasets. This was done because it was expected that R-squared values would decrease from the training to the cross-validated test set (Ferrara et al., 2022). Models 1 and 2 contained all predictors such as DOK and item type. The differences in the models were centered in comparing text complexity features versus grade. Model 1 included grade as a predictor but not text complexity variables. Text complexity variables were included, and grade was excluded in Model 2. This approach allowed us to test the theory that readability formulas were serving as a surrogate for grade level, and the other features were predicting item difficulty when all else was held constant.

Results

Relationship Between Item Difficulty and Predictors

The correlations between the dependent variable, item difficulty, and the most salient independent variables are shown in Table 4. Item difficulty had a weak to moderate correlation with grade, DOK, Flesh-Kincaid, and RALD, and a moderate correlation with passage length and Lexile. Flesch-Kincaid and Lexile measures had strong correlations with grade level and each other. The correlations between item RALD and text complexity measures were close to 0, suggesting that there was little relationship between increases in reasoning as defined by RALDs and increases in text complexity. Similarly, there was little relationship between item DOK and text complexity measures. The variance inflation factor (VIF, Stevens, 2009) for all variables was calculated. VIF statistics identify predictors with multicollinear relationships. The VIF statistic indicates if there is a strong linear association between each predictor and the other predictors used in the model (Stephens, 2009). A VIF less than 5 was found except for grade and Lexile which were near 7.0. This supported the need to compare the features in two different models. When the VIF was computed separately for Model 1 and Model 2, no variable reached a VIF of 5.0.

Table 4. Correlations Among Difficulty and the Ordinal and Continuous Independent Variables

Variable	Difficulty	Grade	RALD	DOK	Flesch-Kincaid	Lexile	#Words
Difficulty	1.00	–	–	–	–	–	–
Grade	0.25	1.00	–	–	–	–	–
RALD	0.43	-0.09	1.00	–	–	–	–
DOK	0.23	0.10	0.22	1.00	–	–	–
Flesch-Kincaid	0.26	0.84	-0.04	0.09	1.00	–	–
Lexile	0.30	0.90	-0.03	0.12	0.84	1.00	–
#Words	0.33	0.46	0.13	0.13	0.49	0.43	1.00

Model Predictions

Table 5 presents model-data fit indices for the two models. Three evaluation metrics were applied: mean absolute error (MAE), mean square error (MSE), and R-squared. MAE measures the average absolute difference between the actual value and the predicted value of the dependent variable, and MSE measures the average squared difference between the actual value and the predicted value of the dependent variable. Compared to MAE, MSE punishes large errors in prediction. R-squared is the proportion of the variance in the dependent variable explained by the predictors. The results are the averages from the 10 cross-validation datasets. The evaluation statistics show that Model 2, in which text complexity features were included in the model rather than grade, explained 3% more of the variance in item difficulty. Both training and validation indexes are included to allow for comparisons with the research literature. For Model 2, 37% of reading item difficulty could be explained by the variables.

Table 5. Model Comparison: Average Across 10 Training and Validation Sets

Predictors	Model	Training			Validation		
		MSE	MAE	R^2	MSE	MAE	R^2
All variables except text complexity	1	0.85	0.68	0.37	0.86	0.69	0.34
All variables except grade	2	0.83	0.66	0.40	0.85	0.68	0.37

Appropriateness of Applying the Linear Regression Model. A linear relationship existed between the outcome variable and the predicted outcome variable based on each combination of predictors, which met the assumption of the linear regression model for both Model 1 and Model 2. The residuals from the linear regression model were checked for each case to see if the linear regression model assumptions, such as independence, homoscedasticity, and normality of residuals, were met. The residuals were normally distributed using a histogram and probability (Q-Q) plots for inspection. The quantiles of residuals formed a straight line when plotted against quantiles of a normal distribution, suggesting that the residuals are normally distributed. This was the case for each model.

Significance of Predictors from Multiple Linear Regression.

How the predictors predicted item difficulties in each model were examined. What follows are the significant predictors common to both models: DOK, RALD, choice multiple item type, Gap Match item type, Standard 1.6c (identify and explain why authors use literary devices), Standard 1.6f (use text features to locate information and explain how the information contributes to understanding print and digital text), and Standard 1.6g (compare and contrast the characteristics that distinguish a variety of literary and informational texts). In Model 1, grade was a significant predictor. In Model 2 passage length (i.e., number of words), Lexile score, and Standard 1.5b (Apply context clues) were additional significant predictors. Table 6 presents the IDM results for both models. When the grade level alignment of items was removed, Lexile and passage length became significant predictors, and all coefficients of the other variables remained highly similar to those in Model 1.

While holding all other predictors constant, each increase in RALD level was associated with an increase in the average difficulty of items of .74 in Model 1 and .68 in Model 2 on the IRT scale. Each DOK level was associated with an increase in the average difficulty of items of .12 in Model 1 and .11 in Model 2.

Discussion

Our first research question was to determine if RALD-to-task match was a significant predictor of item difficulty. RALD-to-task match was associated with a positive increase in item difficulty, which provides validity evidence for their intended purpose and use of providing content-based interpretations regarding what scale scores within a particular achievement level represent. Their relationship with item difficulty supports Perie and Huff's (2016) claim that RALDs are serving as a model of cognition. Predictors should

Table 6. Model Comparison Coefficients of Independent Variables

	Model 1						Model 2					
				95% CI						95% CI		
Variable	β	SE β	t	*LL*	*UL*	ϱ	β	SE β	t	*LL*	*UL*	ϱ
const	-2.96	0.13	-22.08	-3.23	-2.70	0.00	-3.72	0.17	-22.24	-4.04	-3.39	0.00
DOK	0.12	0.05	2.64	0.03	0.21	**0.01**	0.11	0.05	2.50	0.02	0.20	**0.01**
ald	0.74	0.04	17.69	0.66	0.82	**0.00**	0.68	0.04	16.49	0.60	0.76	**0.00**
grade	0.16	0.01	12.63	0.13	0.18	**0.00**	-	-	-	-	-	-
ItemType_ChoiceMultiple	0.62	0.10	6.34	0.43	0.81	**0.00**	0.58	0.10	6.08	0.40	0.77	**0.00**
ItemType_GapMatch	0.95	0.11	9.05	0.74	1.15	**0.00**	0.90	0.10	8.80	0.70	1.11	**0.00**
CCSSType_Literary	-0.02	0.06	-0.28	-0.13	0.10	0.78	0.07	0.06	1.21	-0.04	0.18	0.23
Standard_1.5.b	-0.23	0.12	-1.97	-0.46	0.00	0.05	-0.27	0.12	-2.31	-0.49	-0.04	**0.02**
Standard_1.5.d	0.09	0.12	0.74	-0.15	0.32	0.46	0.07	0.12	0.56	-0.16	0.30	0.57
Standard_1.6.a	0.09	0.13	0.66	-0.17	0.34	0.51	0.06	0.13	0.46	-0.19	0.31	0.64
Standard_1.6.b	0.12	0.13	0.96	-0.13	0.37	0.34	0.06	0.12	0.45	-0.19	0.30	0.65
Standard_1.6.c	0.45	0.13	3.56	0.20	0.70	**0.00**	0.43	0.13	3.41	0.18	0.67	**0.00**
Standard_1.6.d	-0.12	0.17	-0.70	-0.45	0.21	0.48	-0.16	0.16	-0.96	-0.48	0.16	0.34
Standard_1.6.e	0.18	0.13	1.46	-0.06	0.43	0.15	0.14	0.12	1.15	-0.10	0.38	0.25
Standard_1.6.f	0.38	0.12	3.13	0.14	0.61	**0.00**	0.33	0.12	2.82	0.10	0.56	**0.01**
Standard_1.6.g	0.30	0.14	2.25	0.04	0.57	**0.03**	0.29	0.13	2.18	0.03	0.55	**0.03**
Standard_1.6.h	0.37	0.22	1.68	-0.06	0.80	0.09	0.28	0.21	1.32	-0.14	0.70	0.19
Standard_1.6.i	0.08	0.11	0.79	-0.13	0.29	0.43	0.08	0.10	0.73	-0.13	0.28	0.46
Standard_1.6.j	0.43	0.12	3.49	0.19	0.68	**0.00**	0.40	0.12	3.32	0.17	0.64	**0.00**
n_word	-	-	-	-	-	-	0.00	0.00	6.39	0.00	0.00	**0.00**
fk_score	-	-	-	-	-	-	0.00	0.02	0.07	-0.04	0.04	0.95
lexile	-	-	-	-	-	-	0.00	0.00	5.82	0.00	0.00	**0.00**

be related to item difficulty when evaluating cognitive complexity (S. Ferrara, personal communication, March 2024). The model of creating RALDs in this study followed the Egan et al. (2012) framework and provides preliminary positive evidence to support the construction framework model that is generally depicted as intertwining content features, cognitive complexity features, and contextual features. They also need to be carefully crafted based on learning science literature. The RALD framework appears to work when conceptualized as text-task interactions (Valencia et al., 2014) that move from literal comprehension ability to analysis and reasoning using evidence across a passage. Creating RALDs and then coding items for RALD match—along with the degree of explicitness of evidence, location of evidence, and amount of evidence (Sheehan & Ginther, 2001; Valencia et al., 2017)—may codify a set of contextual features to embed into RALDs for reading consistently. To our knowledge this is the first study that investigates RALDs as a predictor of reading comprehension item difficulty.

We also found other predictors in our model were significant such as DOK and item type. We found that technology-enhanced item types tended to be more difficult, replicating findings from Wools et al. (2019). Our findings that standards may predict item difficulty should be replicated with other testing programs using a PAD approach based on RALDs. While this finding makes intuitive sense, only 8.8% of items had a RALD-to-item match to Level 3, as an example. It could be that which standards had more items matched to other predictors of item difficulty contributed to this finding. In an OLS regression, if standards were underrepresented in the dataset in combination with other predictors, then the model has less information to accurately estimate the effect of those standards. This most certainly occurred. While adding interaction effects into the model would be possible, predictors need to be fully crossed in the dataset to do so.

Our second purpose for this study was to better understand the role text complexity plays in measuring reading comprehension in large-scale state accountability assessments in Grades 3 – Grade 8. Therefore, we compared a model that included all predictors in our study and used grade as a predictor (excluding text complexity predictors) to a model that included all predictors in our study and used text complexity predictors (excluding grade as a predictor). We found the OLS multiple linear regression models achieved similar R-squared values (0.34 and .37) when excluding or including text complexity predictors in the model. The results of this study closely replicate the findings of Ferrara and Steedle (2015), who found text complexity explained 1% of the variation in item difficulty in Grade 3–8 assessments when investigating within-grade item difficulty and text complexity relationships. In our study, text complexity features contributed an additional 3% of explained variance in item difficulty beyond the grade level of the item and other predictors.

It appears that, in large part, quantitative metrics of text complexity are serving as a proxy for grade level given the strong relationship between the two. When quantitative metrics of text complexity were removed from the model, grade became a significant predictor of item difficulty. This suggests that the quantitative text complexity metrics may be functioning as intended. Item developers typically use such metrics to help establish what grade-appropriate text complexity means. Toyama et al. (2017) argued that text complexity metrics within a grade should be similar and that a stair-step approach of increases in text complexity difficulty should be observed across grades. The suggestions of Toyama et al., the findings of Ferrara and Steedle (2015), and our findings taken together may indicate that, within a grade for a state accountability assessment, text complexity statements should not be the central feature that varies in RALDs.

Received: 2/27/2025. **Accepted:** 1/13/2026. **Published:** 2/3/2026.

Citation: Schneider, M.C., Chen, J., & Heneger, J. (2026). Text complexity versus task complexity: Item difficulty modeling for reading items. *Practical Assessment, Research, & Evaluation*, 31(1)(5). Available online: https://doi.org/10.7275/pare.2928

Corresponding Author: Christina Schneider, Cambium Assessment, Inc.
Email: cschne5934@aol.com

References

ACT. (2017). *ACT college & career readiness standards: Reading.* https://www.act.org/content/dam/act/unsecured/documents/CCRS-ReadingStandards.pdf

Alderson, J. C., de Jong, J., Kirsch, I., Lafontaine, D., Lumley, T., Mendelovits, J., & Searle, D. (2009). *How can we predict the difficulty of PISA reading items? The process of describing item difficulty*. Paper presented at the Language Testing Forum, University of Bedfordshire, England.

College Board (2015). *Test Specifications for the Redesigned SAT*®. https://collegereadiness.collegeboard.org/pdf/test-specifications-redesigned-sat-1.pdf

CTB/McGraw Hill. (2013). *Initial achievement level descriptors: Technical report*. Smarter Balanced. https://portal.smarterbalanced.org/hubfs/44715956/technical-report-initial-achievement-level-descriptors.pdf

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, *16*(4), 486-514.

Egan, K. L., Schneider, M. C., & Ferrara, S. (2012). Performance-level descriptors. In G. J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 79–106). Routledge.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*(2), 175–193.

Ferrara, S., & Steedle, J. (2015, April). *Predicting item parameters using regression trees: Analyzing existing data to understand and improve item writing*. Presentation at the annual meeting of the National Council on Measurement in Education, Chicago.

Ferrara, S., Steedle, J., & Franz, R. (2022). Response demands of reading comprehension test items: A review of item difficulty modeling studies. *Applied Measurement in Education,35*(3), 237–253.

Florida Department of Education (2016). Florida standards assessments achievement level descriptors. https://www.fldoe.org/core/fileparse.php/5663/urlt/2015fsarangesummary.pdf

Forte, E. (2017). *Evaluating alignment in large scale standards based assessment systems*. Washington, DC: Council of Chief State School Officers. https://files.eric.ed.gov/fulltext/ED586799.pdf

Freedle, R., & Kostin, I. (1992). *The prediction of SAT reading comprehension item difficulty for expository prose passages*. ETS Research Report RR-91-29. https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2333-8504.1991.tb01396.x

Georgia Department of Education (2025). Georgia Milestones Assessment System: Draft achievement level descriptors grade 3 English language arts. https://lor2.gadoe.org/gadoe/file/0398b5b8-bcad-46a5-af4c-643df515341c/1/GM_ALDs_Gr3_ELA.pdf

Gorin, J. S., & Svetina, D. (2008). *SAT Critical Reading Q-matrix study: LLTM analysis of Q-matrix attributes*. Report submitted to the College Board. Available from the authors.

Hiebert, E. H., & Pearson, P. D. (2014). Understanding text complexity: Introduction to the special issue. *The Elementary School Journal*, *115*(2), 153–160.

Huff, K., Warner, Z., & Schweid, J. (2016). Large-scale standards based assessments of educational achievement. In A. A. Rupp & J.P. Leighton, (Eds). *The handbook of cognition assessment: Frameworks, methodologies, and applications* (pp. 399–426).

Huff, K., Nichols, P., & Schneider, M. C. (2025). Designing and developing educational assessments. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 441–511). National Council on Measurement in Education. DOI: 10.1093/oso/9780197654965.003.0007

Kincaid, J. P., Fishburne, L. R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Research Branch Report* (pp. 8–75). Memphis, TN: Chief of Naval Technical Training: Naval Air Station.

Kirsch, I. S., & Mosenthal, P. B. (1990). Exploring Document Literacy: Variables Underlying the Performance of Young Adults. *Reading Research Quarterly*, *25*(1), 5-30. https://doi.org/10.2307/747985

Linacre, J. M. (2015). Winsteps® Rasch measurement computer program (V3.91.0.0). Beaverton, OR: winsteps.com.

Lumley, T., Routitsky, A., Mendelovits, J., & Ramalingam, D. (2012, April). *A framework for predicting item difficulty in reading tests*. Paper presented at the annual meeting of the American Educational Research Association, Vancouver, BC. https://research.acer.edu.au/cgi/viewcontent.cgi?article=1004&context=pisa

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174.

Metametrics. (n.d.). *Lexile & Quantile measures: Supporting student college and career readiness*. https://hub.lexile.com/lexile-quantile-measures-supporting-student-college-and-career-readiness/

Metametrics. (2007). The Lexile framework for reading technical report. https://metametricsinc.com/wp-content/uploads/2017/07/Stenner_Burdick_Sanford__Burdick-_The_LFR_Technical_Report.pdf

Mosenthal, P. B. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, 88*(2), 314–332.

Mosenthal, P. B. (1998). Defining prose task characteristics for use in computer-adaptive testing and instruction. *American Educational Research Journal*, *35*(2), 269–307. https://doi:10.2307/1163425

National Assessment Governing Board. (2023). *Reading framework for the 2026 National Assessment of Educational Progress.* U.S. Department of Education. https://www.nagb.gov/content/dam/nagb/en/documents/publications/frameworks/reading/2026-reading-framework/naep-2026-reading-framework.pdf

Napolitano, D., Sheehan, K. M., & Mundkowsky, R. (2015). *Online readability and text complexity analysis with TextEvaluator*. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations (pp. 96–100). Association for Computational Linguistics. https://aclanthology.org/N15-3020/

National Governors Association Center for Best Practices & Council of Chief State School Officers (CCSSO). (2010). *Common core state standards for English language arts & literacy in history/social studies, science,* *and technical subjects*. https://learning.ccsso.org/wp-content/uploads/2022/11/ADA-Compliant-ELA-Standards.pdf

National Institute of Child Health and Human Development (2000). *Report of the National Reading Panel: Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*. https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf

Nebraska Department of Education. (2014). *Nebraska English Language Arts Standards*. https://cdn.education.ne.gov/wp-content/uploads/2018/06/2014-ELA-Standards.pdf

Nebraska Department of Education. (2021). 2021 Range achievement level descriptors (ALDs) Grade 1-12. https://www.education.ne.gov/wp-content/uploads/2022/08/2021-Range-ALDs_June-2022-2.xlsx

Perie, M., & Huff., K. (2016). Determining the content and cognitive demand for achievement tests. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 119–143). Routledge.

RAND Reading Study Group. (2002). *Reading for understanding: Toward an R & D program in reading comprehension.* RAND.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Paedagogiske Institut.

Rowe, M., Ozuru, Y., & McNamara, D. (2006). *An Analysis of Standardized Reading Ability Tests: What Do Questions Actually Measure?* In ICLS 2006 - International Conference of the Learning Sciences, Proceedings (vol. 2, pp. 627-633). https://repository.isls.org/handle/1/3566

Sano, M. (2016 April). *Improvements in automated capturing of psycho-linguistic features in reading assessment text*. Paper presented at the annual meeting of the National Council on Measurement in Education, Washington, DC.

Scheuneman, J., Gerritz, K., & Embretson, S. (1991). *Effects of prose complexity on achievement test item difficulty*. ETS Research Report ETS-RR-91-43. Princeton, NJ: Educational Testing Service. https://files.eric.ed.gov/fulltext/ED389717.pdf

Sheehan, K., & Ginther, A. (2001). *What do multiple choice verbal reasoning items really measure? An analysis of the cognitive skills underlying performance on a standardized test of reading comprehension skill*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.

Sheehan, K., & Mislevy, R. J. (1990). Integrating cognitive and psychometric models to measure document literacy. *Journal of Educational Measurement*, *27*(3), 255–272.

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

Stevens, J.P. (2009). *Applied multivariate statistics for the social sciences.* 5th Edition, Routledge, New York.

Texas Education Agency. (2024). Range performance level descriptors: Grade 8 math. https://tea.texas.gov/student-assessment/assessment-initiatives/2024-ttap-8-math-range-performance-level-descriptors.pdf

Toyama, Y. (2019). *What makes reading difficult? An investigation of the contribution of passage, task, and reader characteristics on item difficulty, using explanatory item response models*. Doctoral dissertation, University of California, Berkeley.

Toyama, Y., Hiebert, E.H., & Pearson, P.D. (2017). An analysis of the text complexity of leveled passages in four popular classroom reading assessments. *Educational Assessment, 22*(3), 193–170

Valencia, S. W., Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *The Elementary School Journal*, *115*(2), 270–289.

Valencia, S. W., Wixson, K. K., Ackerman, T., & Sanders, E. (2017). *Identifying Text-Task-Reader Interactions Related to Item and Block Difficulty in the National Assessment for Educational Progress Reading Assessment*. https://www.air.org/sites/default/files/downloads/report/Identifying-Text-Task-Reader-Interactions-Related-to-Item-and-Block-Difficulty-NAEP-Oct-2017.pdf

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Webb, N. L. (2005). *Web alignment tool (WAT): Training manual*. Draft Version 1.1. Wisconsin Center for Education Research, Council of Chief State School Officers. http://watv2.wceruw.org/

Wools, S., Molenaar, M., & Hopster-den Otter, D. (2019). The validity of technology enhanced assessments—threats and opportunities. In B. P. Veldkamp & C. Sluijter (Eds.), *Theoretical and practical advances in computer-based educational measurement* (pp. 3–19). Springer.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago: MESA Press.