Check for updates

INTERNATIONAL LITERACY ASSOCIATION READING RESEARCH QUARTERLY

TIMOTHY AND CYNTHIA SHANAHAN OUTSTANDING DISSERTATION AWARD FOR 2021

What Makes Reading Difficult? An Investigation of the Contributions of Passage, Task, and Reader Characteristics on Comprehension Performance

Yukie Toyama
University of California, Berkeley, USA

ABSTRACT

In this study, I investigated the simultaneous effects of the reader, the text, and the task factors, and their interactions, on reading comprehension, using explanatory item response models. Analyses of a large data set from a commercially available online assessment system with a wide range of readers (n = 10,547) and passages (n = 48) uncovered factors that contribute to reading challenge in complex ways. Among the passage features, sentence length, word frequency, syntactic simplicity, and temporality were found to significantly affect comprehension difficulty. More importantly, these textual features were moderated by student general vocabulary and task type. In general, high-vocabulary readers benefited more from traditional textual affordances (e.g., shorter sentences, familiar words, simpler grammatical constructions) than low-vocabulary readers, especially when asked to recall localized information without accessing the passage. However, a reverse effect was found with temporality: Passages with more time markers helped low-vocabulary readers, whereas low-temporality passages helped high-vocabulary readers. Ultimately, understanding these complex interactions, as highlighted in the RAND Reading Study Group's heuristic model, will be key in supporting students with their comprehension development.

Reading comprehension (RC) is a multifaceted, multilayered construct (Duke, 2005; Graesser & McNamara, 2011; Perfetti & Stafura, 2014; van den Broek & Espin, 2012) that is manifested within a complex interaction among three broad factors: the reader, the passage, and the task. These factors reside in a particular sociocultural context (RAND Reading Study Group, 2002). Investigating this complexity deepens understanding about the simultaneous contributions of the reader, the passage, and the task, as well as their complex interactions, in the face of a challenge to RC.

Although the RAND Reading Study Group's (2002) interactive view of RC is widely popular, surprisingly few attempts have been made to directly model the empirical phenomena related to how a reader interacts with a particular text[11] for a particular purpose. Rather, quantitative reading research has tended to investigate correlates of cognitive and affective predictors of RC within readers, through regression-based methods (e.g.,

INTERNATIONAL LITERACY ASSOCIATION

Reading Research Quarterly, 56(4)
pp. 633–642 | doi:10.1002/rrq.440
© 2021 International Literacy Association.

Adlof, Catts, & Lee, 2010; Cervetti, 2020; Kendeou, van den Broek, White, & Lynch, 2009; Language and Reading Research Consortium, 2015; Oakhill & Cain, 2012). Although this research has documented the important role that subword-level language skills can play for early and later reading outcomes, it has not addressed how the particular features of a passage or task contribute to students' RC in a certain context.

Another line of quantitative research has focused on the readability of text, in a quest, ultimately, to find the "just right" match between a reader and a text (e.g., Bormuth, 1969; Klare, 1984). Although this research has uncovered various linguistic and discourse characteristics of text that are highly predictive of text difficulty, it has paid little attention to task and reader characteristics. In fact, the cloze procedure (Taylor, 1953), which is prominent in readability research, is a product of an effort to eliminate nontextual factors such as how questions are crafted by test writers. Whereas proponents of the cloze method have argued that it is an efficient and nonsubjective measure of RC, the method has been criticized as being insensitive to RC beyond a single sentence (e.g., Shanahan, Kamil, & Tobin, 1982).

Recognizing these limitations in the literature, in the present study, I sought to disentangle the influences of passage, task, and reader characteristics, as well as their interactions, on students' comprehension of informational text. Specifically, I examined a large response data set from a commercially available online RC assessment, which included a wide range of readers and passages, covering grades 1–12. Prior studies, in contrast, have examined RC assessments that were narrowly designed for target groups (e.g., the Gates–MacGinitie Reading Tests–RC, Graduate Record Examinations–Verbal [GRE-V]). Such strong focusing inevitably limited the variability in passage and reader factors in these earlier studies.

Research Questions

Two research questions guided the present study:

1. Which set of text and task features best explain variability in the difficulty of RC items after controlling for student general vocabulary knowledge?
2. Are any of the text feature effects moderated by student and/or task characteristics?

Conceptual Frameworks

In addition to the RAND Reading Study Group's (2002) heuristic model, the item difficulty modeling paradigm, especially Embretson and Wetzel's (1987) processing model for multiple-choice RC assessment, framed this study. Based partially on Kintsch's (1988, 1998) construction–integration model of comprehension, Embretson and Wetzel's framework specifies two phases of the response process: (1) At the text representation phase, examinees build mental model(s) of the source text as they read it, followed by (2) the response decision phase, when they build the mental model of the question and answer choices, map them against the one(s) built for the text, and evaluate the plausibility and falsifiability of response options as they respond to a question.

Embretson and Wetzel (1987) and subsequent studies (Gorin, 2005; Gorin & Embretson, 2006; Ozuru, Rowe, O'Reilly, & McNamara, 2008) investigated the effect of text features (e.g., propositional density) and task features (e.g., the extent to which distractors can be falsifiable) on item difficulty, using standardized RC assessments such as the GRE-V and the Gates–MacGinitie Reading Tests–RC. The findings from these studies are inconclusive as to whether the passage or the item features explain greater variance in RC in assessment contexts. Additionally, researchers have only begun to incorporate the reader characteristics and examine the simultaneous effects of the reader, the task, and the text, as well as their interactions (e.g., Kulesz, Francis, Barnes, & Fletcher, 2016). Yet, to my knowledge, there has been no study that has investigated the full three-way interactions among the text, the reader, and the task.

Method

Data and Analytic Samples

During the winter of the 2015–2016 academic year, 41,555 students in the United States and Canada took Reading-Plus's RC assessment, InSight. It is a commercially available, computer-adaptive assessment designed to determine students' reading levels for placement in an online reading intervention program. InSight is formatted as a series of testlets, each consisting of a short informational text followed by five multiple-choice questions.[2] Each passage was designed for one of 12 levels, roughly corresponding to grades 1–12, based on their vocabulary demand.[3] The analytic sample entailed four testlets per level, totaling 48 passages and 240 items.

Student grade levels ranged from first grade through postsecondary. Typically, students responded to five testlets during one test administration. Of the five testlets, four were adaptively selected, and the fifth was randomly selected from the test bank. I took advantage of item responses from the fifth testlet (i.e., the randomly selected ones) because they were administered to students with a wide range of comprehension ability, independent of the assessment's adaptive logic.

To construct a vertical scale that covers a wide range of testlets and students, a response data set needs an anchor, through common items, common persons, or both, to link an otherwise sparse matrix unless all students take all items (Kolen & Brennan, 2004). In this study, if the data matrix

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

included only the responses from the fifth randomly chosen testlet, it would represent 48 groups of students taking 48 different tests, with no common items or persons linking the sparse data matrix. To overcome this issue, seven testlets were selected as anchors, and I included only those students who took any of the anchor tests as second, third, fourth, or fifth testlets in the analytic sample. I excluded responses from the first testlet from the analysis, as I considered this first testlet to be a practice test for students to get familiar with the assessment system. To provide chaining linkages across the data matrix, I chose these anchor testlets to be ones that spread across the assessment developer's 12 passage levels. Additionally, I selected the anchor testlets to minimize the effect of the assessment's adaptive logic on the item difficulty estimates.[4]

The resulting analytic sample comprised 10,547 students, who were randomly divided into two samples: sample 1 for model building and sample 2 for cross-validation. Table 1 shows descriptive statistics for general vocabulary knowledge, grade levels, and the number of items responded to in the two samples. As expected, no statistically significant difference was found between them.

Variables

In this study, I sought to explain student performance on each of the comprehension items on the InSight assessment with three types of predictors: the text, the task, and the reader. I drew the text predictors from those used in three prominent studies in the literature: Gorin and Embretson's (2006) model, the Lexile framework for reading (Stenner, Burdick, Sanford, & Burdick, 2006), and Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011). Table 2 provides descriptive statistics for each of the text predictors used in the study.

Additionally, each of the 240 items was coded by researchers on three task features: item type/comprehension process, abstractness of information requested by the question, and falsifiability. These were found to be salient in the previous studies and were assumed to affect the response decision phase of Embretson and Wentzel's (2006) framework. Additionally, the vocabulary demands of the question, and the answer choices in terms of the Flesch–Kincaid grade level (Kincaid, Fishburne, Rogers, & Chissom, 1975) were computed using the Quantitative Discourse Analysis Package in R (Rinker, 2013). Descriptive statistics for each of these task predictors are shown in Tables 3 and 4.

Finally, I used students' general vocabulary knowledge as the reader covariate, which was measured within the InSight assessment system prior to the RC section. The vocabulary scores ranged from 0 to 13 (in grade levels) with a mean of 5.4 (see Table 1 for summary statistics for the two samples).

Explanatory Item Response Models

I used doubly explanatory item response models (De Boeck & Wilson, 2004), which include both item and person predictors that explain item responses. Figure 1 illustrates the model specification: The outcome of interest is a student's response to a RC item, or more precisely, the log odds of a person p correctly answering item i, which is modeled as a function of person p's reading ability (θ_p) and the difficulty of item i (δ_i), according to the Rasch model. This model is then extended to include explanatory components: On the item side, the model postulates that the item difficulty (δ_i) can be explained by text features, such as mean sentence length (MSL) and mean log word frequency (MLWF), and/or task features such as falsifiability—the number of falsifiable distractors with explicit information in the source text. Note that the text and task features are both treated as item features in the model. On the person side, person p's reading ability (θ_p) is modeled with reader characteristics, which in this study was general vocabulary knowledge. The magnitude and directionality of coefficients (β_1, β_2, β_3, and ϑ_1 in Figure 1) show the effects of reader, text, and task predictors on students' success at answering a RC item.

I examined four broad categories of models: (1) the text models with different sets of text predictors from the three models in the literature, (2) the task model, (3) the text and task combined model, and (4) the interaction models. I ran all models using the meglm command in Stata/SE 15.0 (StataCorp, 2017), first with the model-building sample and then with the cross-validation sample.

Results

Table 5 shows the results from the four models: the text model, the task model, the text and task combined model,

TABLE 1
Comparison of the Two Student Samples

	Sample 1 (n = 5,274)		Sample 2 (n = 5,273)			
Variable	*M*	*SD*	*M*	*SD*	*t*	*p*
General vocabulary knowledge	5.42	1.58	5.45	1.58	1.20	.23
Grade level	6.78	2.38	6.82	2.42	0.82	.41
Number of items responded to	10.89	2.06	10.90	2.05	0.29	.77

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2
Summary Statistics for the Text Feature Variables (*N* = 48 passages)

Description	Scale/unit	*M*	*SD*	Min	Max
Gorin and Embretson's (2006) model					
1. Modifier propositional density	Count of adjectives per 1,000 words	91.58	2.49	43.10	136.36
2. Predicate propositional density	Count of verbs per 1,000 words	128.17	22.29	88.24	195.98
3. Text content vocabulary level	Word frequency for content words	2.14	0.21	1.75	2.64
Lexile framework for reading (Stenner, Burdick, Sanford, & Burdick, 2006)					
4. Mean sentence length	Count of sentences per passage	14.33	2.76	7.95	17.50
5. Mean log word frequency[a]	Logarithm of the word frequency	3.46	0.21	3.13	3.73
Coh-Metrix (Graesser, McNamara, & Kulikowich, 2011)					
6. Narrativity[a]	*z*-score	−0.47	0.39	−1.12	0.13
7. Syntactic simplicity[a]	*z*-score	0.85	0.58	−0.25	1.76
8. Word concreteness[a]	*z*-score	1.61	0.57	0.53	2.44
9. Referential cohesion[a]	*z*-score	0.42	0.56	−0.78	1.26
10. Causal cohesion[a]	*z*-score	1.16	1.52	−1.37	4.05
11. Verb cohesion[a]	*z*-score	1.13	1.58	−1.17	3.58
12. Logical cohesion/connectivity[a]	*z*-score	−2.32	1.20	−4.36	−0.08
13. Temporal cohesion[a]	*z*-score	−0.08	0.86	−1.49	1.46

Note. Statistics in the table are all in their original unit. All variables were standardized for analyses.
[a]The variables that make text easier to process and comprehend (i.e., higher values indicate easier texts).

TABLE 3
Summary Statistics for the Continuous Task Feature Variables (*N* = 240 items)

Variable	*M*	*SD*	Min	Max
Vocabulary demand of the question[a]	4.99	3.09	−2.62	18.22
Vocabulary demand of the correct answer[a]	5.66	5.21	−3.40	26.49
Vocabulary demand of the distractors[a] (mean)	5.29	3.13	−1.07	13.56
Number of falsifiable distractors	0.63	0.95	0.00	3.00

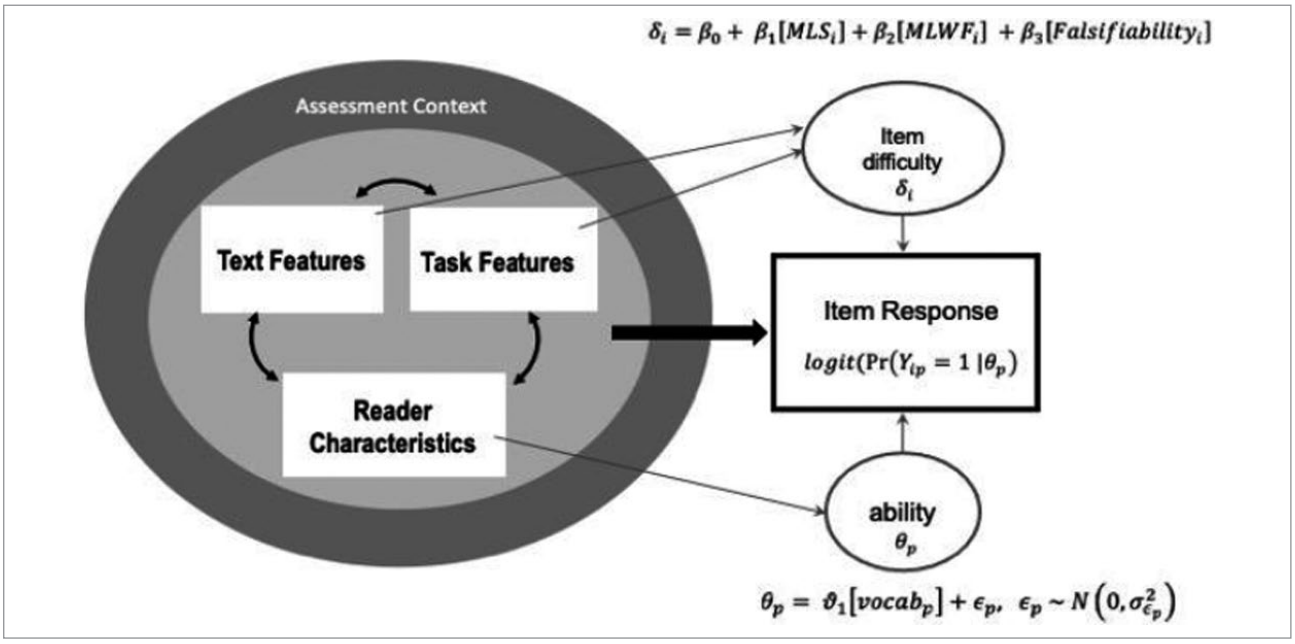
[a]Automatically computed in the Flesch–Kincaid grade levels (Kincaid, Fishburne, Rogers, & Chissom, 1975) with the Quantitative Discourse Analysis Package in R (Rinker, 2013).

TABLE 4
Summary Statistics for the Categorical Task Feature Variables (*N* = 240 items)

Classification scheme	Type of question	*N*	%
Item type/comprehension process	Text-based/literal recall	23	9.58
	Reconstruct	84	35.00
	Integrate	65	27.08
	Knowledge-based	68	28.33
Abstractness	Highly concrete	45	18.75
	Somewhat concrete	60	25.00
	Somewhat abstract	84	35.00
	Highly abstract	51	21.25

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

FIGURE 1
A Diagram of an Explanatory Item Response Model



Note. $\in_p$ = a random error term for person ability; $MLWF_i$ = mean log word frequency; MSL_i = the source text's mean sentence length; $vocab_p$ = vocabulary knowledge. The model depicts how the Rasch model—$\text{logit}[Pr(Y_{ip} = 1 \mid \theta_p)] = \theta_p - \delta_i$—can be extended to include two additional equations: one modeling item difficulty (δ_i) with text and/or task features (e.g., MSL_i, $MLWF_i$, $Falisifiability_i$) and another modeling reading comprehension ability (θ_p) with reader characteristics (e.g., $vocab_p$). The probability of person *p*'s response to an item (*i*) is transformed into log odds (a continuous metric) so linear regression techniques can be used.

and one of the three-way interaction models. According to a pseudo-R^2 index,[5] the text model with ten text predictors explained 54.5% of variance in item difficulty after controlling for student's general vocabulary knowledge. In contrast, the Task model, also with ten predictors, explained 25.4%.

These results indicate that the text features, which were thought to influence the text representation phase in Embretson and Wentzel's (1987) framework, had a larger impact on student performance than did the task features, which were thought to affect the response decision phase. When the text and task features were combined, the explanatory power increased to 59%. Further, a three-way interaction model examining the effect modification of temporality accounted for 63% of the variance in item difficulty.

Notably, MSL, MLWF, syntactic simplicity, and temporality were the four textual predictors that were consistently found to have statistically significant unique effects on item difficulty in the text model and in the text and task combined model, while controlling for all other variables in the models. As expected, texts with longer sentences made the items more difficult (notice the positive sign for MSL's coefficient in Table 5). In contrast, passages with more frequent words, simpler grammatical construction, and more cohesive ties with temporal markers made the items easier (notice the negative sign for these predictors' coefficients). When all of these models were calibrated with the cross-validation sample, similar patterns were found. The effect size of these text features in terms of the range of the sample students' RC ability varied from 0.12 for temporality to 0.38 for MSL in the text and task combined model.

As for the task predictors, one of the most interesting and consistent findings was that, contrary to expectations, literal recall/text-based questions were not the easiest type of task among the four types examined (notice the negative sign for the reword and integrate item types). I conjectured that this might be partially due to the InSight assessment's unique feature: Students were not able to access the source passage while answering RC questions, so reading time could be captured. This setup required students to rely more on the text representation that they constructed than would be the case if they had been allowed to reread the text after reading the questions. Consequently, this without-text condition makes it more difficult to recall specific information in a localized section of the source text as compared with rewording or integrating/bridging inference tasks that typically are considered more cognitively demanding (Anderson, 1972; Hua & Keenan, 2014; Ozuru et al., 2008; Pearson, Hansen, & Gordon, 1979).

I further explored whether each of the four textual main effects was moderated by a reader's general vocabulary knowledge, the task type, and both the reader and the task types. The most noteworthy results were found with the models that examined the effect modification of temporality—the extent to which passages have temporal markers (e.g., *then*, *after*, the

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5
Parameter Estimates From Selected Explanatory Item Response Models

	Text model		Task model		Text + Task combined model		Text × Reader × Task model	
Variable	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***	**Est.**	***SE***
Fixed effects								
Text features								
Mean sentence length	**0.21*** **	**0.04**			**0.22*** **	**0.04**	**0.24*** **	**0.04**
Mean log word frequency	**−0.20*** **	**0.03**			**−0.18*** **	**0.03**	**−0.16*** **	**0.03**
Syntactic simplicity	**−0.12***	**0.05**			**−0.11***	**0.05**	**−0.12***	**0.05**
Temporality	**−0.09*** **	**0.02**			**−0.08*** **	**0.02**	0.03	0.04
Narrativity	−0.06	0.04			0.07	0.05	−0.09	0.05
Word concreteness	>0.01	0.02			0.03	0.02	0.03	0.02
Referential cohesion	0.02	0.02			0.01	0.02	0.03	0.02
Deep cohesion	>−0.01	0.01			−0.02	0.01	−0.02*	0.01
Verb cohesion	>−0.03	0.02			−0.01	0.02	−0.02	0.02
Logical cohesion	0.03	0.01			0.03	0.01	0.01	0.02
Task features								
Vocabulary demand of the question			<0.01	0.01	−0.03*	0.01	**−0.05*** **	**0.01**
Vocabulary demand of the correct answer			**0.20*** **	**0.01**	**0.03***	**0.01**	<0.01	0.01
Vocabulary demand of the distractors			**0.02****	**0.01**	**−0.04****	**0.01**	**−0.04****	**0.01**
Item type (reference = literal recall)								
• Reword			**−0.19*** **	**0.04**	−0.15***	0.04	−0.10*	0.04
• Integrate			**−0.18*** **	**0.04**	−0.12**	0.04	−0.09	0.05
• Knowledge-based			**0.09***	**0.04**	>0.01	0.04	−0.03	0.05
Abstractness (reference = highly concrete)								
• Somewhat concrete			**0.30*** **	**0.03**	**0.22*** **	**0.04**	**0.26*** **	**0.04**
• Somewhat abstract			**0.17*** **	**0.03**	**0.20*** **	**0.03**	**0.24*** **	**0.03**
• Highly abstract			**0.15****	**0.04**	**0.14****	**0.04**	**0.16*** **	**0.04**
Falsifiability			−0.02*	0.01				
Reader characteristics								
Vocabulary knowledge	**0.35*** **	**0.02**	**0.32*** **	**0.02**	**0.35*** **	**0.02**	**0.25*** **	**0.04**
Text × Reader								
Temporality × Vocabulary Knowledge							**0.09***	**0.04**
Text × Task								
Temporality × Reword							0.06	0.04
Temporality × Integrate							0.09	0.05
Temporality × Knowledge-Based							0.18***	0.04

(continued)

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

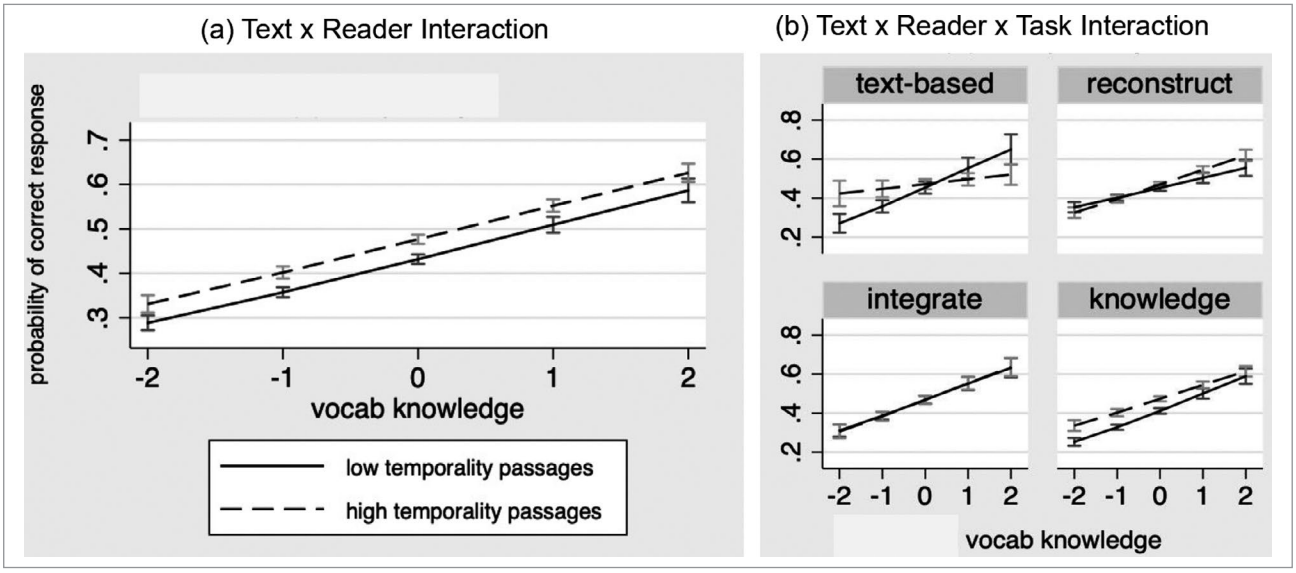
TABLE 5
Parameter Estimates From Selected Explanatory Item Response Models (*continued*)

	Text model		Task model		Text + Task combined model		Text × Reader × Task model	
Variable	Est.	*SE*	Est.	*SE*	Est.	*SE*	Est.	*SE*
Task × Reader								
Reword × Vocabulary Knowledge							−0.09	0.05
Integrate × Vocabulary Knowledge							**−0.15****	**0.05**
Knowledge-Based × Vocabulary Knowledge							**−0.10***	**0.05**
Text × Reader × Task								
Temporality × Vocabulary Knowledge × Reword							**0.12***	**0.05**
Temporality × Vocabulary Knowledge × Integrate							**0.12***	**0.05**
Temporality × Vocabulary Knowledge × Knowledge-Based							0.06	0.05
Random effects								
Reader variance							**0.41*****	**0.02**
Pseudo-R^2	.55		.25		.59		.63	

Note. Est. = estimate; *SE* = standard error. Bolded values are the statistically significant effects replicated in the cross-validation analysis.
**p* < .05. ***p* < .01. ****p* < .001.

consistency in verb tense). As can be seen in Figure 2a, the Text × Reader model found no moderation of temporality by student vocabulary knowledge, indicated by the parallel lines representing the two contrasting levels of temporality: one standard deviation above and below the mean temporality scores for the 48 passages examined.

FIGURE 2
(a) A Line Plot Depicting Interactions Between Temporality (the Text Feature) and General Vocabulary Knowledge (in *z*-Scores), and (b) Line Plots Depicting Three-Way Interactions Among Temporality, General Vocabulary Knowledge, and Four Separate Item Types



Note. Figure 2a shows that the text's main effect is not moderated by general vocabulary knowledge, as evident with parallel lines representing two levels of the text feature (±1 *SD* from the mean). Figure 2b shows a statistically significant three-way interaction involving text-based item type with the crossed lines. Hinge-like shapes along the lines indicate the 95% confidence intervals for the success rate estimates.

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In contrast, the Text × Reader × Task model found what experimental researchers have called a reverse effect (McNamara, Kintsch, Songer, & Kintsch, 1996; O'Reilly & McNamara, 2007): As shown in Figure 2b for text-based item type, readers with below-average vocabulary knowledge (i.e., <0 *z*-scores for general vocabulary knowledge on the *x*-axis) are expected to do better on high-temporality passages than low-temporality passages, as evident with the dashed line being above the solid line. However, these lines are crossed, indicating that the reverse is true for readers with above-average vocabulary knowledge: The passages with fewer temporal markers give them a greater boost in their expected success rate. Interestingly, this reverse effect is limited to the literal recall/text-based item type. In light of findings from prior research (e.g., McNamara et al., 1996; McNamara & Kintsch, 1996), I conjectured that less cohesive passages (in terms of temporal markers) encouraged readers with high vocabulary knowledge and, by extension, with high background knowledge to engage more actively in text processing and generating inferences, which likely helped them build refined text representations, enabling them to perform better on the literal recall task even without access to the source passage.

In contrast, Figure 3 shows results for another textual feature, MSL. The widening gap between the two lines in Figure 3a indicates a differential effect of MSL by a reader's general vocabulary knowledge: Passages with short sentences gave a greater boost in the probability of success for readers with a higher level of general vocabulary knowledge than for their peers with lower vocabulary knowledge. In contrast, Figure 3b shows no three-way interaction, indicating that this text–reader interaction was not further differentiated by item type. Results for the remaining two text features examined, MLWF and syntactic simplicity, showed similar patterns as MSL (for details, see Toyama, 2019).

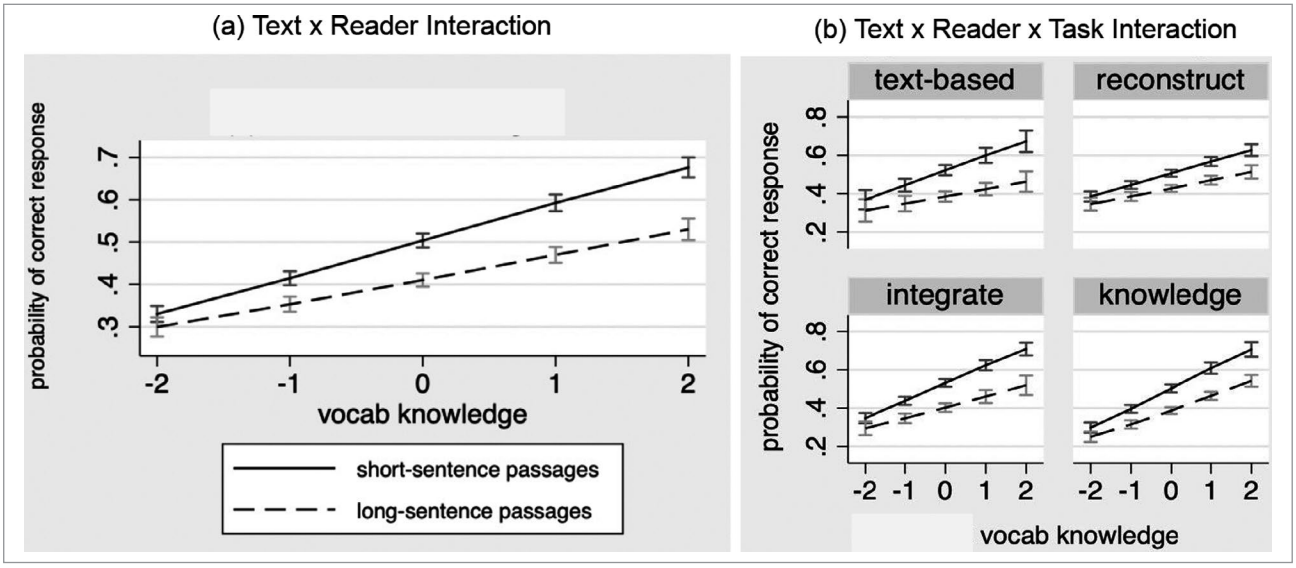
Discussion

This study offered ways to investigate variations in RC performance by simultaneously modeling explanatory variables about the reader, the text, the task, and their interactions, embracing, and making very explicit, the interactive view of comprehension offered by the RAND Reading Study Group's (2002) heuristic model. Importantly, the explanatory item response modeling uncovered possible sources of text processing difficulty in the InSight assessment that differentially affected RC performance, depending on students' general vocabulary knowledge and the item types.

One conclusion I drew was that it was the text features that had more explanatory power than the task features, after controlling for a reader's general vocabulary knowledge. This finding is contrary to what Embretson and her colleagues found using RC assessments for older examinees. Embretson and Wetzel (1987) examined the Armed Services Vocational Aptitude Battery, whereas Gorin and Embretson (2006) used the GRE-V. Both studies reported that the task features had a greater explanatory power than the text features did.

Two factors might have contributed to this difference. First, in the current study, I examined a wider range of

FIGURE 3
(a) A Line Plot Depicting Interactions Between Mean Sentence Length (the Text Feature) and General Vocabulary Knowledge, and (b) Line Plots Depicting Three-Way Interactions Among Mean Sentence Length, General Vocabulary Knowledge, and Four Separate Item Types



Note. Figure 2a shows that the text's main effect is moderated by general vocabulary knowledge, as evident with the widening of the gap between the two lines representing two levels of the text feature (±1 *SD* from the mean). Figure 2b shows no statistically significant three-way interaction. Hinge-like shapes along the lines indicate the 95% confidence intervals for the success rate estimates.

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

passages in which vocabulary demand was intentionally varied for grades 2–12, as compared with the prior studies. This variability in the source passages may have contributed for the text features to reach a greater explanatory power. Another factor had to do with the lack of access to the source passage while answering questions. This feature is a strong difference from most RC assessments, which typically allow students to access the passage. This feature means that the InSight assessment focuses more on memory-based RC than other assessments do.

Proponents have argued that memory-based RC assessments better capture automatized comprehension processes for building text representations by preventing examinees from engaging in more effortful, nonlinear text processing and test-taking strategies (Artelt, Schiefele, & Schneider, 2001; Higgs, Magliano, Vidal-Abarca, Martínez, & McNamara, 2017), although students' performance in this without-text condition is known to be driven more by their prior knowledge about the topic of the passage than in the with-text condition (Ozuru, Best, Bell, Witherspoon, & McNamara, 2007). Findings from this study suggest that the lack of access to the source passage likely altered the nature of the reading task and, thereby, the nature of the target construct being measured, which explains why the current study found a larger explanatory power for the text features that are thought to influence building text representations, rather than the task features that affect reasoning and mapping processes during the response decision phase.

In addition, it seems clear that what makes reading difficult entails the three factors—the text, the task, and the reader—which interact in complex ways. Specifically, the reverse effect of temporality found in this study lends further support to the argument put forth by McNamara and colleagues (1996): Good texts are not always better or easier; it depends on the reader and task characteristics. Ultimately, understanding these complex interactions among the reader, the passage, the task, and their interactions will help in identifying students with RC difficulties and in designing targeted interventions that best support their RC development.

NOTES

[1] The terms *text* and *passage* are used interchangeably throughout this article.

[2] For each passage, four types of RC questions were asked: literal recall, gap-filling, text-connecting, and main idea.

[3] The assessment developer determined the vocabulary demand with Lexile's mean log word frequency (A. Spichtig, personal communication, July 14, 2018).

[4] To identify the seven testlets, two sets of item responses were calibrated concurrently with the Rasch model. The first set was for the 240 items with the responses to the testlets given by the adaptive logic, and the second set was for the same 240 items but only with the responses to the randomly given fifth testlet. As anchors, I selected those with the least discrepancy in item difficulty between the two sets.

[5] Embretson (1983) specified this as $\Delta^2 = (\ln L_0 - \ln L_m) / (\ln L_0 - \ln L_S)$, where $\ln L_0$ is the log-likelihood for the null model with just an intercept, which assumes a constant difficulty value for all items; $\ln L_m$ is the log-likelihood for the model to be evaluated, which includes a set of text, task, and reader predictors—the doubly explanatory item response model; and $\ln L_S$ is the log-likelihood for the saturated model in the Rasch-latent regression model with item dummies and the student's general vocabulary knowledge as the person covariate. The denominator in the equation shows the maximum amount of variance in item difficulty that can be modeled, and the numerator shows how much improvement a set of predictors in the double explanatory item response model makes as compared with the null model.

REFERENCES

Adlof, S.M., Catts, H.W., & Lee, J. (2010). Kindergarten predictors of second versus eighth grade reading comprehension impairments. *Journal of Learning Disabilities*, *43*(4), 332–345. https://doi.org/10.1177/0022219410369067

Anderson, R.C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research*, *42*(2), 145–170. https://doi.org/10.3102/00346543042002145

Artelt, C., Schiefele, U., & Schneider, W. (2001). Predictors of reading literacy. *European Journal of Psychology of Education*, *16*(3), 363–383. https://doi.org/10.1007/BF03173188

Bormuth, J.R. (1969). *Development of readability analysis: Final report*. Washington, DC: Bureau of Research, Office of Education, U.S. Department of Health, Education, and Welfare.

Cervetti, G.N. (2020). The nature and development of reading for understanding. In P.D. Pearson, A.S. Palincsar, G. Biancarosa, & A. Berman (Eds.), *Reaping the rewards of the Reading for Understanding initiative* (pp. 41–66). Washington, DC: National Academy of Education.

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.

Duke, N.K. (2005). Comprehension of what for what: Comprehension as a nonunitary consruct. In S.G. Paris & S.A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 93–104). Mahwah, NJ: Erlbaum.

Embretson, S.E. (1983, June). *An incremental fit index for the linear logistic latent trait model*. Paper presented at the annual meeting of the Psychometric Society, Los Angeles, CA.

Embretson, S.E., & Wetzel, C.D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*(2), 175–193. https://doi.org/10.1177/014662168701100207

Gorin, J.S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement*, *42*(4), 351–373. https://doi.org/10.1111/j.1745-3984.2005.00020.x

Gorin, J.S., & Embretson, S.E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*(5), 394–411. https://doi.org/10.1177/0146621606288554

Graesser, A.C., & McNamara, D.S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*(2), 371–398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Graesser, A.C., McNamara, D.S., & Kulikowich, J.M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

Higgs, K., Magliano, J.P., Vidal-Abarca, E., Martínez, T., & McNamara, D.S. (2017). Bridging skill and task-oriented reading. *Discourse Processes*, *54*(1), 19–39. https://doi.org/10.1080/0163853X.2015.1100572

Hua, A.N., & Keenan, J.M. (2014). The role of text memory in inferencing and in comprehension deficits. *Scientific Studies of Reading*, *18*(6), 415–431. https://doi.org/10.1080/10888438.2014.926906

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Kendeou, P., van den Broek, P., White, M.J., & Lynch, J.S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology*, *101*(4), 765–778. https://doi.org/10.1037/a0015956

Kincaid, J.P., Fishburne, R.P. Jr, 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 No. 8-75). Millington, TN: Research Branch, Naval Education and Training Support Command.

Kintsch, W. (1988). The role of knowledge in discourse processing: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Klare, G.R. (1984). Readability. In P.D. Pearson, R. Barr, M.L. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (pp. 681–744). Mahwah, NJ: Erlbaum.

Kolen, M.J., & Brennan, R.L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer.

Kulesz, P.A., Francis, D.J., Barnes, M.A., & Fletcher, J.M. (2016). The influence of properties of the test and their interactions with reader characteristics on reading comprehension: An explanatory item response study. *Journal of Educational Psychology*, *108*(8), 1078–1097. https://doi.org/10.1037/edu0000126

Language and Reading Research Consortium. (2015). Learning to read: Should we keep things simple? *Reading Research Quarterly*, *50*(2), 151–169. https://doi.org/10.1002/rrq.99

McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*(1), 1–43. https://doi.org/10.1207/s1532690xci1401_1

McNamara, D.S., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes*, *22*(3), 247–288. https://doi.org/10.1080/01638539609544975

Oakhill, J.V., & Cain, K. (2012). The precursors of reading ability in young readers: Evidence from a four-year longitudinal study. *Scientific Studies of Reading*, *16*(2), 91–121. https://doi.org/10.1080/10888438.2010.529219

O'Reilly, T., & McNamara, D.S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43*(2), 121–152. https://doi.org/10.1080/01638530709336895

Ozuru, Y., Best, R., Bell, C., Witherspoon, A., & McNamara, D.S. (2007). Influence of question format and text availability on the assessment of expository text comprehension. *Cognition and Instruction*, *25*(4), 399–438. https://doi.org/10.1080/07370000701632371

Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D.S. (2008). Where's the difficulty in standardized reading tests: The passage or the question? *Behavior Research Methods*, *40*(4), 1001–1015. https://doi.org/10.3758/BRM.40.4.1001

Pearson, P.D., Hansen, J., & Gordon, C. (1979). The effect of background knowledge on young children's comprehension of explicit and implicit information. *Journal of Reading Behavior*, *11*(3), 201–209. https://doi.org/10.1080/10862967909547324

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

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

Rinker, T.W. (2013). qdap [Computer software]. Retrieved from http://trinker.github.io/qdap/

Shanahan, T., Kamil, M.L., & Tobin, A.W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17*(2), 229–255. https://doi.org/10.2307/747485

StataCorp. (2017). Stata (Version 15) [Computer software]. Retrieved from https://www.stata.com/

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.

Taylor, W.L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism Quarterly*, *30*(4), 415–433. https://doi.org/10.1177/107769905303000401

Toyama, Y. (2019). *What makes reading difficult? An investigation of the contributions of passage, task, and reader characteristics on item difficulty, using explanatory item response models* (Doctoral dissertation). Retrieved from https://escholarship.org/uc/item/36j1p7hr

van den Broek, P., & Espin, C. (2012). Connecting theory and assessment: Measuring individual differences in reading comprehension. *School Psychology Review*, *41*(3), 315–325. https://doi.org/10.1080/02796015.2012.12087512

YUKIE TOYAMA earned her PhD in quantitative methods and evaluation in education from the University of California, Berkeley, USA. Her dissertation was supported by her committee members: Mark Wilson (chair), P. David Pearson, Susanne Gahl, and Elfrieda H. Hiebert. Toyama is a research scientist at the UC Berkeley Evaluation and Assessment Research Center, USA; email yukie.toyama@berkeley.edu. Her work centers at the intersection of the learning sciences and measurement in science, technology, engineering, arts, and mathematics, with an emphasis on text and language complexity and comprehension.

INTERNATIONAL LITERACY ASSOCIATION

ILA INTENSIVE

Supporting Multilingual Learners With Translanguaging Strategies

Carla España | Ofelia García | Luz Yadira Herrera | Mark B. Pacheco
Shakina Rajendram | Kate Seltzer | Heather H. Woodley

Register for On-Demand Access

literacyworldwide.org/ILAIntensives

19362722, 2021, 4, Downloaded from https://ila.onlinelibrary.wiley.com/doi/10.1002/rrq.440 by Stanford University, Wiley Online Library on [08/03/2025]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License