Abstract

The purpose of this study is to provide initial validity evidence for a test of learning from text, that is, reading to learn. It is grounded in theoretical and empirical evidence that the ability to read for general comprehension is not the same as the ability to read to learn (e.g., Carver, 1990; Enright et al., 2000; Grabe, 2009; Koda & Yamashita, 2015). Most transfer of information in higher education occurs through text, making the ability to learn from text a crucial skill for university success. This study examines reading to learn in first and second language readers of English. It compares two tests, one of reading for general comprehension and one of reading to learn, in an effort to determine if these two tests are measuring the same construct or two related but different constructs. Performance on the tests as well as participants' perceptions of the tests are compared. Test performance supports the argument that reading for general comprehension and reading to learn are related but distinct constructs. Participant perceptions of the tests do not support this interpretation, but this may be a result of the limitations of using self-report data.

Is comprehending text the same as learning from text?

Deirdre J. Derrick: Northern Arizona University

Learning from written text is an important skill in educational and professional contexts. This skill has been called "reading to learn" by first and second language researchers (Carver, 1990; Grabe, 2009; Koda & Yamashita, 2015). Reading to learn can be described as the ability to absorb and retain information from text. It is a type of reading that is adopted by readers when their purpose for reading is to increase their knowledge about a topic. It is commonly used in academic settings, where students are frequently held responsible for knowing the content in assigned readings. It can also be used in professional settings, an example of which would be an employee learning from a manual. The ability to learn from text is therefore a crucial skill that students should possess.

This study builds upon work in first and second language reading research, and is therefore relevant to both. Within first language (L1) reading research, there has been increasing interest in how educators can provide language and reading support within content disciplines such as science and history (e.g., Goldman, Snow, & Vaughn, 2016; Greenleaf & Valencia, 2017). Several projects have received government grants to fund explorations of how to improve students' ability to learn from text (e.g., La Russo et al., 2016; Goldman et al., 2016; Vaughn et al., 2015). In second language (L2) reading research, work continues on how language, reading, and content instruction can be integrated to improve students' reading and learning in English (Koda & Yamashita, 2015).

One critical barrier in these studies is the absence of a test to measure the ability to learn from written text. Researchers are currently limited to using tests of reading for general comprehension to measure student improvements following interventions designed to support learning from text (e.g., Jones et al., 2016; Vaughn et al., 2015). Given the theoretical and empirical support for differences between reading for general comprehension and reading to learn from text, it is clear that information is potentially being lost.

This study seeks to address this issue by developing and providing validity support for a test of reading to learn from text. It will follow established recommendations for developing validity support by showing that the test is measuring what it has been developed to measure.

Literature Review

The review of literature frames the discussion of reading to learn by presenting an overview of how the theory of reading to learn has developed, first in L1 reading research and then in L2 reading research. Following from discussions of reading to learn, a construct definition is proposed. This definition presents reading to learn as a construct which is related to, yet distinct from, other types of reading, most notably reading for general comprehension.

Defining Reading to Learn

The discussion of reading to learn originated in L1 reading research. Table 1 presents three views of reading to learn from Chall (1983), Carver (1990), and Maclellan (1997). Chall's (1983) influential book introduced the distinction between *learning-to-read* and *reading-to-learn*. According to this framework, the first three years of formal schooling are spent mastering the basic skills of reading comprehension such as decoding and building vocabulary, and in the fourth year, students transition from learning to comprehend text to using text to learn content. In

this new *reading-to-learn* stage, students become responsible for knowing the information in their class textbooks and support for learning to read disappears (Cervetti, Jaynes, & Hiebert, 2009). Carver's (1990; 1997) conceptualization of reading to learn involves not only remembering ideas, but also the ability to decode as well as make meaning from sentences and then integrate them. Maclellan (1997) incorporates the importance of monitoring and the ability to construct a situation model. His explanation of reading to learn also includes the types of knowledge that are required in order to learn from text. Readers who read for the purpose of studying, or learning, make more explanatory and predictive inferences, and those who read for entertainment make more associative inferences (van den Broek, Lorch, Linderholm, & Gustafson, 2001). When reading for a study purpose, readers with low working memory make more paraphrases and readers with high working memory make more metacognitive comments (Linderholm & van den Broek, 2002).

Table 1
Component Knowledge and Processes in Reading to Learn (L1)

Chall (1983)	Carver (1990; 1997)	Maclellan (1997)	Linderholm & van den Broek (2002); van den Broek et al. (2001)
• Finding information in text efficiently • Mastering ideas • Vocabulary knowledge • Prior knowledge	• Lexical accessing • Semantic encoding • Sentence integrating • Idea remembering	• Cognitive and metacognitive monitoring • Constructing a situation model • Summarizing • Content knowledge • Discourse knowledge • Strategic knowledge	• Predictive inferences • Explanatory inferences • Metacognitive monitoring

In describing the component skills necessary in reading to learn, L2 reading researchers focus on text integration, situation-model construction, memory, and prior knowledge (Table 2). Enright et al. (2000) and Grabe (2009) both emphasize the importance of creating an organizational frame of the text to aid in the construction of a situation model. Grabe (2009) notes the importance of both remembering information and recalling the information for later use, thus introducing a temporal element to the discussion of reading to learn. Work by both Grabe (2009) and Koda and Yamashita (2015) acknowledge the role of prior knowledge in reading to learn.

Table 2
Component Knowledge and Processes in Reading to Learn (L2)

Enright et al. (2000)	Grabe (2009)	Koda & Yamashita (2015)
• Locating information • Basic comprehension of text	• Remembering main ideas and many supporting details	• Building text-meaning • Constructing personal meaning

• Integrating and connecting text information into a coherent whole • Forming links between an elaborated model of text construction and organizational frames (i.e., an efficient alignment of text model and situation model)	• Recalling information as needed • Using supporting information • Creating an organizing frame • Building a close integration with prior knowledge • Prior knowledge	• Refining knowledge

From the work done in both L1 and L2 reading research, a construct definition of reading to learn can be proposed. Reading to learn is a type of reading adapted by readers when the purpose of reading is to remember information from a text and later recall that information at a later point in time. Thus, reading to learn involves memory and time. Reading to learn includes knowledge of discourse structure, either sub-conscious or conscious (Grabe, personal communication, October 10, 2016). This knowledge is used to build a mental representation of the text that is organized. This organization, in turn, aids in memory and recall. Reading to learn also involves explanatory and predictive inferences, as well as paraphrases. Finally, reading to learn involves cognitive and metacognitive monitoring that guide reader strategies to aid in text retention and recall.

As stated before, the role of language proficiency is never absent from discussions of L2 reading to learn. Language instruction is noted as necessary in Koda and Yamashita's (2015; Table 3) framework of reading to learn. Their framework includes three stages. The first is text-meaning building skills, which include word recognition and integration, main idea identification, text model construction, and identification of the author's view. The second stage is personal-meaning construction, in which the reader relates text content with personal experience and content knowledge, as well as compares the author's view with his or her own. The third stage is knowledge-refinement, where the reader reflects on three things. The first two are the similarities and differences between the author's view and his or her personal view, and the third is the changes in his or her view that have occurred as a result of reading the text.

Reading to learn is distinct from reading for general comprehension. Reading for general comprehension is the type of reading adopted when the purpose is to read for pleasure (Carver, 1990; Grabe, 2009). The primary difference between reading to learn and reading for general comprehension is that the former requires a conceptual integration of the main ideas of a written text. Because reading for general comprehension does not require this, it is hypothesized to be less challenging that reading to learn.

Assessing Reading to Learn

Although a theoretical differentiation of reading to learn from other purposes or types of reading has existed for over twenty years, development of assessments that can measure ability in this type of reading have lagged. In L1 contexts, development of measures has begun with pre-adolescent and adolescent readers. In L2 contexts, the focus has typically been on developing assessments for university-aged students (i.e., admissions tests such as the TOEFL iBT).

L1 measures have focused on cognitive skills and comprehension of content areas. Davis and Guthrie (2015) used a measure of text relationships and validated it against free recalls. They found that 3rd and 9th grade students who were able to distinguish between closely related

concepts and less closely related concepts showed greater understanding of text. Uccelli Galloway, Barr, Meneses, and Dobbs (2015) developed a test of Core Academic Language Skills (CALS) that included unpacking dense information, organizing analytic texts, connecting ideas logically, tracking participants and themes, interpreting writers' viewpoints, and understanding metalinguistic vocabulary. With participants in grades 4-6, Uccelli et al. found that the ability to use these academic skills accounted for 50% of reading comprehension even after vocabulary, reading fluency, grade, SES, and English proficiency were accounted for. Both studies support the importance of discourse awareness as part of reading to learn.

Most discussion of L2 reading to learn has originated in language assessment literature, specifically in literature related to the development of the TOEFL iBT, an academic English proficiency test. The TOEFL Framework listed reading to learn as a purpose for reading, along with reading for basic comprehension and reading to integrate (Enright et al., 2000). The framework hypothesized a hierarchy of difficulty, with reading to learn being more difficult than reading for basic comprehension. This, however, was not born out in research, where items created to measure learning from text were easier than items created to measure basic comprehension (Cohen & Upton, 2007).

Other attempts to measure reading to learn have been made by using constructed response items in which examinees are required to write a response. Trites and McGroarty (2005) successfully operationalized reading to learn using graphic organizers. Weigle, Montee, and Yang (2013) found evidence of learning from text in examinee responses to short answer questions. Plakans (2009) successfully included use of text content in an analytic rubric used to score essays written in response to texts. In L2 contexts, these have limited utility as measures of reading to learn because they require written responses by examinees. High scores indicate that an examinee is able to read and write a response. Low scores, however, do not provide information about the source of examinees' difficulties. It is unclear if the source of a low score is a reading problem, a writing problem, or a combination of the two.

Although this is less of a concern with L1 readers, there are other advantages of using selected response item types (e.g., multiple choice) that are more applicable to this population. First, while L1 readers may have the language proficiency, they do not necessarily have the written proficiency to construct a well-developed written response. Second, use of selected response item formats allows for rapid scoring. Third, selected response item formats have the potential to be used with large numbers of examinees. Depending on the purpose of assessment, such ease of administration is potentially very useful.

There are arguments against selected response item formats, namely that they cannot be used to elicit higher order thinking. In the case of multiple choice reading tests, studies have shown that examinees engage in matching behaviors rather than building a textbase and situation model of the text (Farr, Pritchard, & Smitten, 1990; Rupp, Ferne, & Choi, 2006). Design of the test and the items must be done carefully to limit matching behavior and create an environment where examinees are required to display the skills associated with reading to learn. A multiple choice test may not elicit reading to learn itself, but if it can elicit behaviors associated with reading to learn, it can still provide useful information about examinees.

The purpose of this study is to develop a test to measure students' ability to learn from text. In doing so, it also provides support for the argument that reading for general comprehension and reading to learn are two related but distinct constructs. The quality and construct relevance of a reading to learn test can be assessed by answering the following research questions:

1. Does a test of reading to learn vary in predictable and systematic ways from a test of reading for general comprehension?
2. Do examinee perceptions of the two tests differ?
3. To what extent does background knowledge affect performance on a test of general comprehension versus a test of reading to learn?

Method

Participants

This study used a convenience sample of 62 undergraduate native (n = 54) and non-native (n = 8) English speakers from a single university in the Southwest of the United States. Students ages ranged from 18 to 24. Non-native English speakers included five Arabic speakers and three Chinese speakers.

Participants were from three intact Freshman Composition classes, and the study was conducted during class time. All three classes were taught by the same instructor (not the researcher). Students were given the option to not participate in the study without penalty; students who chose to participate were offered ten extra credit points by the teacher. Students who did not wish to participate were allowed to leave the classroom.

Instruments

Three instruments were developed for this study. The first was a reading for general comprehension test (R4GC), the second was a reading to learn test (R2L), and the third was a questionnaire. The two tests used three of the same passages. Passage development is described first, followed by development of the R4GC test, development of the R2L test, and development of the questionnaire.

Passages. Passage development involved several stages. First, six passages were selected from a 12th grade Earth Science textbook. Three criteria for selection of passages were followed. First, passages had to be self-contained and not require specialized knowledge. Second, passages had to not be common knowledge. The first two criteria were ensured through discussion with a Science teacher (M.S.) what had experience teaching the topics and could provide information about what students would and would not typically know. Third, passages had to be short (~500 words). Once the passages were selected, they were further analyzed for the quality and number of test items that could be derived from them. From this analysis, four passages were selected. These passages were about the Permian extinction event, deep-sea hydrothermal vents, aerosols, and geoengineering.

The passages went through several rounds of review. First, they were shown to a Biology teacher with an M.S., who provided feedback about their scientific accuracy. Several minor revisions were made to the passages to update the information to reflect current scientific thought. The passages were also shown to an L2 reading expert (Ph.D. in Applied Linguistics) and an L2 assessment expert (Ph.D. in Applied Linguistics). Minor revisions were made following their feedback.

The passages were then piloted using the think aloud method with six graduate students who were both native and non-native English speakers. Based on their comments and feedback, the passages underwent major revisions. First, the passages were restructured. This was following comments by the participants that the passages were confusing and hard to understand. Also, the paragraphs were numbered. A sentence was added at the end of the first paragraph in the *Aerosols* passage to explicitly state that aerosols were not the cause of global warming. Pictures and bolded words were added to the passages to clarify concepts and highlight important

information. Following this, the passages, as part of the R4GC and R2L tests, were then piloted with two intact Freshman Composition classes (n = 16 and n = 19). The passages performed well and no further revisions were made.

Readability statistics for all of the passages and vocabulary statistics for the Permian Extinction passage can be found in Appendix A.

Reading for General Comprehension Test (R4GC). The purpose of the R4GC test was to measure general comprehension. The three passages selected for this test were *Permian Extinction, Deepsea Hydrothermal Vents,* and *Aerosols*. For the R4GC test, it was decided that passages and items would be presented simultaneously. Three R4GC testlets were developed. Each testlet had eight items. Item types were selected based on R4GC theory and existing test of general reading comprehension. Item types included main idea, detail, resolution of pronouns, and vocabulary. Items for the R4GC test were developed with the idea that the passage would be available to examinees. This meant that it was less important to consider what examinees would have in their mental model of the text and more important to consider what examinees would need to comprehend to answer an item correctly.

The R4GC test went through the same review and piloting processes as the passages. First, the R4GC test underwent reviews from the science content expert, the L2 reading expert, and the L2 assessment expert. Based on their feedback, revisions were made to the items to secure keys (i.e., correct answers), make distractors (i.e., incorrect options) non-keyable (i.e., there was nothing in the text to make them a possible correct answer), and update content. The R4GC test was then piloted with two intact Freshman Composition classes (n = 16 and n = 19). No further changes were made.

The first testlet, Permian Extinction, and the table of specifications for the complete R4GC test is included in Appendix B.

Reading to Learn (R2L) Test. The purpose of the R2L test was to measure the ability to learn from text. The four passages selected for this test were *Permian Extinction, Deepsea Hydrothermal Vents, Aerosols,* and *Geoengineering*. For the R2L test, it was decided that passages and items would be presented separately. Three R2L testlets were developed. Each testlet had eight items. Item development was guided by R2L theory about what R2L is and how it differs from R4GC. The skills that appear to distinguish reading to learn from reading for general comprehension are memory for text content, awareness of discourse and discourse organization, the ability to make predictive and explanatory inferences, and the ability to transfer what is learned from a text to a new context.

The final result was three testlets. The first two testlets consisted of a single passage (~500 words). These passages were *Permian Extinction* and *Deepsea Hydrothermal Vents*. Both testlets were developed with the intention that the passage would be presented first then removed before the questions were presented. The third testlet was developed using two passages (*Aerosols* and *Geoengineering*). The items were developed with the intention that the first passage would be presented and removed before the second passage and corresponding items were presented simultaneously.

The items (keys and options) underwent moderate revisions. Most revisions were to secure keys (i.e., correct answers) and make distractors (i.e., incorrect options) non-keyable (i.e., there was nothing in the text to make them a possible correct answer). Next, the testlets were shown to the L2 reading expert and the L2 assessment expert. Minor revisions to the testlets were made following their feedback.

The testlets were then piloted with two intact classes of Freshman Composition students (n = 16 for each class). No further changes were made to items.

The items from the first testlet, Permian Extinction, and the table of specifications for the complete R2L test is included in Appendix C.

Questionnaire. Development of the questionnaire also underwent several rounds of revision. First, a questionnaire was developed to measure three constructs: background knowledge, perception of learning, and perception of difficulty. The first draft of the questionnaire used a three-point scale for the first two constructs and a five-point scale for the final construct. Three versions of this questionnaire were created, one for the R4GC tests, one for the first two R2L testlets (*Permian Extinction* and *Deepsea Hydrothermal Vents*) and one for the third R2L testlet (*Aerosols and Geoengineering*). The primary difference in the questionnaires was the number of items, with the third questionnaire having the most (k = 11) because it asked about the relationship between the two passages.

The questionnaires were piloted in the Freshman Composition classes at the same time as their corresponding testlets. Questionnaire data were entered into SPSS for analysis. Analyzing the data from different versions of the same questionnaire presented challenges, so the decision was made to create a single version of the questionnaire that could be administered after each testlet. Another construct was also added: perception of task relevance.

The final questionnaire consisted of 11 items to measure four constructs. Two items measure perception of task difficulty, three measure perception of learning, three measure perception of task relevance, and three measure background knowledge. The final version of the questionnaire is presented in Appendix D.

Procedure

This study used a within-subjects design. Figure 1 presents the research design. The R4GC test was administered first. Three questionnaires were administered, one after each testlet. Twelve days later, the R2L test was administered. Again, three questionnaires were administered, one following each testlet.

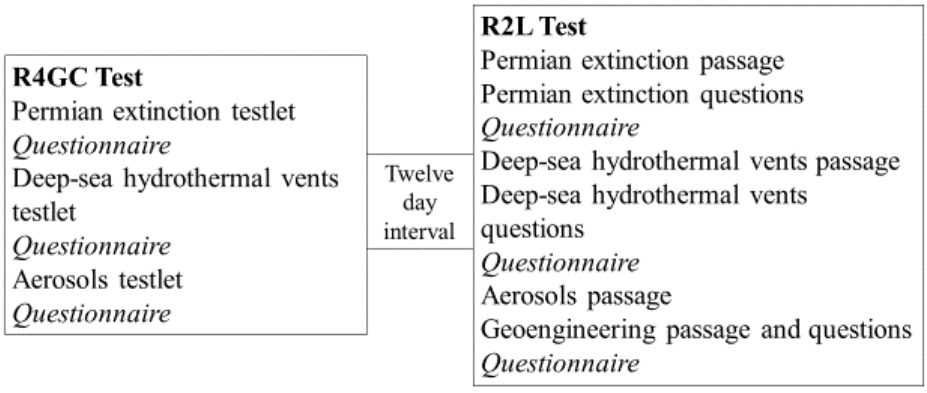

Figure 1. Research design.

Data Analysis

Data were entered into SPSS for analysis. First, two independent samples *t*-tests were conducted to determine if there were significant differences between English L1 and L participants in their performance on the R4GC and R2L tests. There was no statistical difference in performance on either the R4GC test or the R2L test, so data from the two groups were combined for subsequent analyses.

Reliability of the measures was calculated (Cronbach's alpha). Descriptive statistics were calculated, and statistical assumptions were checked. The R4GC and R2L tests met assumptions for parametric statistics. All questionnaire measures met assumptions for parametric statistics except for the R2L background knowledge measure (as indicated by skewness and kurtosis values and a bimodal distribution as shown on the histogram).

A series of paired sample *t*-tests and correlations were performed to answer the research questions.

Results

This study sought to develop a test to measure reading to learn. It did so by examining how a test of reading to learn differs from a test of reading for general comprehension, by examining students' perceptions of the two tests, and by examining the influence of background knowledge on both tests. This section presents the results of the data analysis.

Descriptive statistics of all measures are presented in Table 3. One of the items on the R4GC test contained an error, so it was not included in the analysis. The first two measures presented in the table are the reading for general comprehension test and the reading to learn test. The following measures were from the three R4GC questionnaires and the three R2L questionnaires. Question responses were combined to form four constructs: background knowledge, perception of learning, perception of difficulty, and perception of relevance.

Table 3
Descriptive Statistics

Measure	*k*	Range	Min	Max	Mean	S.D.	Reliability
R4GC	23	0-23	3	21	15.14	3.92	.77
R2L	24	0-24	4	21	13.75	3.86	.70
R4GC Background knowledge	9	9-36	9	13	16.31	5.57	.88
R2L Background knowledge	9	9-36	9	34	15.35	5.55	.91
R4GC Perception of learning	9	9-36	13	36	23.68	4.39	.84
R2L Perception of learning	9	9-36	11	35	22.70	4.80	.83
R4GC Perception of difficulty	6	6-24	12	24	18.66	2.61	.80
R2L Perception of difficulty	6	6-24	12	24	18.02	2.99	.80
R4GC Perception of relevance	9	9-36	9	36	24.34	5.78	.94
R2L Perception of relevance	9	9-36	9	36	24.36	6.41	.93

Several paired sample *t*-tests were conducted to examine differences between the two conditions (R4GC and R2L). Results are shown in Table 4. First, a paired sample *t*-test was conducted with test condition (R4GC and R2L) as the independent variable and test scores as the dependent variable. Results of the comparison showed that the two tests were statistically different, $t(61) = 3.458$, $p = .001$). Next, a paired sample *t*-test was conducted with test condition as the independent variable and perception of test difficulty as the dependent variable. The result was not statistically significant, $t(61) = -1.501$, $p = .14$. Third, a paired sample *t*-test was conducted with test condition as the independent variable and perception of test relevance as the

dependent variable. The result was not statistically significant, $t(61) = -.203$, $p = .84$. Fourth, a paired sample *t*-test was conducted with test condition as the independent variable and perception of learning as the dependent variable. The result was not statistically significant, $t(61) = -.928$, $p = .36$.

Table 4

Paired Sample t-tests

Comparison	t	Sig.	95% Confidence Intervals	
			Lower bound	Upper bound
R4GC – R2L	3.458	.001	.678	2.56
R4GC – R2L Perception of Difficulty	-1.501	.139	-1.126	.162
R4GC – R2L Perception of Relevance	-.203	.840	-1.72	1.40
R4GC – R2L Perception of Learning	-.928	.358	-1.703	.626

A series of correlations were performed. The first correlation was between the R4GC test and the R2L test. The two tests showed a moderate ($r = .63$, $p < .00$) correlation. Neither test correlated significantly with background knowledge. The correlation between R4GC scores and background knowledge was not statistically significant: $r = .052$, $p = .70$. A Spearman's rho correlation was performed between the R2L test scores and background knowledge. The result was not statistically significant ($\rho = .037$, $p = .79$).

Discussion

The purpose of this study was to answer three research questions. The first question asked if and how a test of reading for general comprehension varies from a test of reading to learn. The second question asked if examinee perceptions of the two tests were different. The third question asked to what extent background knowledge would affect scores on a reading for general comprehension test and a reading to learn test.

Three limitations of this study should be acknowledged. First, the design was not counterbalanced. This was done for practical reasons related to the timing of the tests within the curriculum. The R2L test took longer than the R4GC test, so test administration had to match the teacher's schedule for the class. Overall, this limitation was not overly important to the results of the study. The R2L test was given second, so even if participants had remembered any of the content from the R4GC test, that memory would have been simply part of their "learning." It is unlikely that participants remembered text, however. First, their performance on the R2L test was lower than on the R4GC test, indicating that participants did not remember text. Second, the mean for the background knowledge measure was lower for the R2L test (though not significantly so), also indicating that participants did not remember text.

A second limitation is that this study used a convenience sample. This presents problems for generalizability, as conclusions can only be drawn about the participants and not a larger population. Results can be informative, however, for teachers and researchers who wish to consider how reading to learn may be researched within different contexts. Results of the study allow for hypotheses to be formed concerning different groups who have similar characteristics.

The third limitation is the use of self-report data. While in some studies self-report data can provide important insights, it is unlikely that it did so in this study. The questions related to background knowledge were straight-forward and therefore most likely easy for participants to answer. The other questions related to perceptions (of difficulty, of learning, and of relevance) might not have yielded accurate results. This can be seen in the lack of a statistical difference between perceptions of R4GC and R2L difficulty, despite significant difference in test results. In

R2L, it appears that participants may have overestimated their abilities. This merits further research, particularly if students tend to overestimate their ability to learn from text. This tendency may have implications for class performance. This is conjecture, however, requiring targeted research.

Despite these limitations, results of the study provide information about reading to learn and how it is related to reading for general comprehension. Results of the *t*-test and correlation showed that the reading for general comprehension and reading to learn tests are statistically different and moderately correlated. This provides preliminary support for the argument that the tests measuring two related but distinct constructs. Both the R4GC and the R2L tests were developed using reading theory and existing tests, and performance on the tests showed the hypothesized relationship. This result is support for the construct validity of the R2L test.

The results of this study are similar to results from Trites and McGroarty (2005) and Weigle et al. (2013). Both studies were able to measure reading to learn, and Trites and McGroarty were able to show that reading to learn and reading for general comprehension are distinct constructs. The difference between this study and those studies is that their items required participants to write their responses. The introduction of writing into a reading measure means that low scores that are the result of weakness in writing could be interpreted as indicative of a weakness in reading to learn. The measure in this study avoids that potential issue by using only reading.

Results of this study also show that reading to learn is more difficult than reading for general comprehension. This is in direct contrast to Cohen and Upton (2007), who found that reading to learn items were easier that reading for general comprehension items. There are two possible explanations for the difference. The first is that in Cohen and Upton's study, participants had access to the text while they were responding to items and in this study the participants did not. In this study, items were designed to test important information that ought to have been present in a reader's mental model of the text. The second difference is with the items themselves. In Cohen and Upton's study, one reading to learn item required participants to select sentences to create a prose summary of the reading. The other item required participants to select sentences to put into a table. The first item would have required that participants be able to distinguish between important and relatively unimportant information, and the second item would have required that participants be able to understand how information is related. These two skills are part of the construct definition of reading to learn used for this study, but they are not alone. The construct definition also includes the ability to remember information, the ability to make explanatory inferences, the ability to make predictive inferences, and the ability to transfer what has been learned to a new context. The addition of items to test these skills may have increased the difficulty of the reading to learn test used in this study.

Despite empirical evidence that the two tests were different, participants' perceptions of the tests were not. Results of the *t*-tests showed that examinee perceptions of the two tests did not differ in terms of learning, difficulty, and relevance to university reading. In other words, participants did not believe that they had learned more from the reading to learn test. They also did not believe that the reading to learn test was more difficult than the reading for general comprehension test, nor did they believe that the type of reading they did for the reading to learn test was more similar to the type of reading that they do for their university classes than the type of reading they did for the reading for general comprehension test.

These results are difficult to explain without making the mistake of reading too much into them. It might be that the participants did not have any real awareness of how and why they read

or if they have really learned from text. It may be that items were not always clear to participants. Or, it may be that participants were more focused on answering the reading test questions and paid less attention to the questionnaire questions. Or, maybe the participants did not read the passages differently. This may explain why they did not perform as well on the reading to learn test as on the reading for general comprehension test. As stated above, further research about readers' perceptions related to reading for general comprehension and reading to learn should be conducted to better understand these issues.

Results of the correlations showed that background knowledge did not significantly correlate with either of the two tests (R4GC and R2L). This result is not surprising. Tests were constructed in such a way that readers would not be able to use their background knowledge to answer the items. Statistically significant correlations between background knowledge and either of the tests might have been indicative of a weakness in the tests. Furthermore, it can be argued that while reading to learn does require background knowledge, skilled learners should have strategies in place that help them learn unfamiliar or new material. In other words, one of the characteristics that separates a skilled from an unskilled learner may be the ability to overcome limited background knowledge in a content area. A reader without conceptual science knowledge may be able to substitute personal experience to help them understand and learn text.

Conclusion

Results of this study provide preliminary support for the argument that reading to learn is related to yet distinct from reading for general comprehension. The R4GC and R2L tests, which were developed based on reading theory, showed the hypothesized relationship with each other. These results have implications for L1 and L2 reading instruction and assessment. Further research should be conducted to more clearly delineate where the distinctions are. This research should also include other types of reading, such as reading to integrate and reading to evaluate. Theoretical classifications of types of reading will benefit from careful empirical research to support them.

References

Bråten, I., & Strömsó, H. I. (2006). Effects of personal epistemology on the understanding of multiple texts. *Reading Psychology, 27,* 457-484.

Carver, R. P. (1990). *Reading rate: A review of research and theory.* San Diego: Academic Press, Inc.

Cervetti, G. N., Jaynes, C. A., & Hiebert, E. H. (2009). Increasing opportunities to acquire knowledge through reading. In E. H. Hiebert (Ed.) *Reading more, reading better* (pp. 79-100). New York: Guilford Press.

Chall, J. S. (1983). *Stages of reading development.* New York: McGraw-Hill.

Cohen, A. D., & Upton, T. A. (2007). *'I want to go back to the text'*: Response strategies on the reading subtest of the new TOEFL. *Language Testing, 24,* 209-250.

Davis, M. H., & Guthrie, J. T. (2015). Measuring reading comprehension of content area texts using an assessment of knowledge organization. *The Journal of Educational Research, 108,* 148-164.

Enright, M. K., Grabe, W., Koda, K., Mosenthal, P., Mulcahy-Ernt, P., & Schedl, M. (2000). *TOEFL 2000 reading framework: A working paper.* MS-17. Princeton, NJ: Educational Testing Service.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. *Journal of Educational Measurement, 27,* 209-226.

Ferguson, L. E., Bråten, I., & Strömsó, H. I. (2012). Epistemic cognition when students read multiple documents containing conflicting evidence: A think-aloud study.

Goldman, S. R., Snow, C., Vaughn, S. (2016). Common themes in teaching reading for understanding: Lessons from three projects. *Journal of Adolescent and Adult Literacy, 60,* 255-264.

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

Greenleaf, C., & Vilencia, S. (2017). Missing in action: Learning from texts in subject-matter classrooms. In D. Appleman & K. Hinchman (Eds.), *Adolescent literacies: A handbook of practice-based research.* New York, NY: Guilford.

Koda, K. & Yamashita, J. (2015). An integrated approach to EL instruction and assessment. Manuscript in preparation.

Leinhardt, G., & Young, K. M. (1996). Two tests, three readers: Distance and expertise in reading history. *Cognition and Instruction, 14,* 441-486.

Linderholm, T., Therriault, D. J., & Kwon, H. (2014). Multiple science text processing: Building comprehension skills for college student readers. *Reading Psychology, 35,* 332-356.

Linderholm, T., & van den Broek, P. (2002). The effects of reading purpose and working memory capacity on the processing of expository text. *Journal of Educational Psychology, 94,* 778-784.

Maclellan, E. (1997). Reading to learn. *Studies in Higher Education, 22,* 277-288.

McCrudden, M. T., & Kendeou, P. (2014). Exploring the link between cognitive processes and learning from refutational text. *Journal of Research in Reading, 37,* S116-140.

Moje, E. B., Stockdill, D., Kim, K., & Kim, H. (2011). The role of text in disciplinary learning. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.) *Handbook of Reading Research vol. IV* (pp. 453-486). New York: Routledge.

Plakans, L. (2009). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes, 8,* 252-266.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23,* 441-474.

Scott, D., B., & Dreher, M. J. (2016). Student thinking processes while constructing graphic representations of textbook content: What insights do think-alouds provide? *Reading Psychology, 37,* 286-317.

Strómsó, H. I., Bråten, I., & Samuelstuen, M. S. (2003). Students' strategic use of multiple sources during expository text reading: A longitudinal think-aloud study. *Cognition and Instruction, 21,* 113-147.

Strómsó, H. I., Bråten, I., Britt, M. A., & Ferguson, L. E. (2013). Spontaneous sourcing among students reading multiple documents. *Cognition and Instruction, 31,* 176-203.

Trites, L., & McGroarty, M. (2005). Reading to learn and reading to integrate: New tasks for reading comprehension tests? *Language Testing, 22,* 174-210.

Uccelli, P., Galloway, E. P., Barr, C. D., Meneses, A., Dobbs, C. L. (2015). Beyond vocabulary: Exploring cross-disciplinary academic-language proficiency and its association with reading comprehension. *Reading Research Quarterly, 50,* 337-356.

Van den Broek, P., Lorch, R. F., Linderholm, T., & Gustafson, M. (2001). The effects of readers' goals on inference generation and memory for texts. *Memory and Cognition, 29,* 1081-1087.

Weideman, A. & van Dyk, T. (2014). *Academic literacy: test your competence.* Bloemfontein: Geronimo Distribution Ltd.

Weigle, S. C., Yang, W., & Montee, M. (2013). Exploring reading processes in an academic reading test using short-answer questions. *Language Assessment Quarterly, 10,* 28-48, essays.

Wineburg, S.S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology, 83,* 73–87.

Appendix A

Passage Readability and Vocabulary Statistics

Table C.1
Readability Statistics for R2L Passages

	Permian Extinction	Deepsea Hydrothermal Vents	Aerosols	Geoengineering
Flesch reading ease	34.3	40.3	47.6	40.3
F-K grade level	12.6	11.8	10.4	12.0
Lexile measure	1480L	1360L	1170L	1220L
Mean sentence length	16.6	17.6	15.2	19.7
Word Count	530	517	596	482

Table C.2
Vocabulary Statistics for Permian Extinction (Final Draft)

Frequency Level	Families	Types	Tokens	Cumulative Token %
K-1 Words	120 (52.63%)	137 (50.74%)	332 (63.36%)	63.36
K-2 Words	39 (17.11%)	46 (17.04%)	62 (11.83%)	75.19
K-3 Words	24 (10.53%)	27 (10.00%)	34 (6.49%)	81.68
K-4 Words	18 (7.89%)	19 (7.04%)	36 (6.87%)	88.55
K-5 Words	10 (4.39%)	11 (4.07%)	20 (3.82%)	92.37
K-6 Words	7 (3.07%)	7 (2.59%)	7 (1.34%)	93.71
K-7 Words				
K-8 Words	2 (0.88%)	2 (0.74%)	2 (0.38%)	94.09
K-9 Words	2 (0.88%)	2 (0.74%)	2 (0.38%)	94.47
K-10 Words	4 (1.75%)	4 (1.48%)	6 (1.15%)	95.62
K-11 Words				
K-12 Words				
K-13 Words				
K-14 Words	1 (0.44%)	1 (0.37%)	1 (0.19%)	95.81
K-15 Words	1 (0.44%)	1 (0.37%)	1 (0.19%)	96.00
Off-List	??	13 (4.81%)	21 (4.01%)	100.00

Appendix B

Permian Extinction

1) Several mass extinctions have occurred, but the worst occurred 65 million years ago. By the close of the **Permian period**, a **mass extinction** had destroyed ***70 percent*** of all vertebrate species on land, and perhaps ***as much as 90 percent*** of all marine organisms. The late **Permian extinction** was the greatest of at least five mass extinctions to occur over the past 500 million years.

2) Several mechanisms have been proposed to explain these ancient mass extinctions. Initially, paleontologists believed they were gradual events caused by a combination of ***climate change*** and ***biological forces***, such as predation and competition. Then, in the 1980s, a research team proposed that the mass extinction that happened occurred swiftly as a result of an ***explosive impact by a meteorite*** about 10 kilometers in diameter, similar to what killed the dinosaurs in the Cretaceous-Paleogene extinction event. However, scant evidence could be found of debris that would have been generated by an impact large enough to destroy many of Earth's lifeforms.

3) Another possible mechanism for the Permian extinction was the voluminous eruptions of basaltic lavas which began about 251 million years ago and are known to have covered thousands of square kilometers of the land. This was the same period of volcanism that produced the Siberian Traps located in northern Russia (Figure 1). The release of carbon dioxide would certainly have enhanced greenhouse warming, and the emissions of ***sulfur dioxide*** probably resulted in copious amounts of acid rain.

Figure 1. Siberian

4) A recent hypothesis begins with this period of volcanism and the ensuing period of global warming but adds a new twist. These researchers agree that the additional carbon dioxide released into the atmosphere would cause rapid greenhouse warming. This alone, however, would not destroy most plants because they tend to be heat tolerant and consume CO_2 in photosynthesis. They contend, instead, that the trouble began in the ocean rather than on land.

5) Most organisms on Earth use oxygen to metabolize food, as do humans. However, some forms of bacteria employ ***anaerobic*** (without oxygen) metabolism. Under normal conditions, oxygen from the atmosphere is readily dissolved in seawater, and is then evenly distributed to all depths by water currents. This oxygen "rich" water relegates "oxygen-hating" anaerobic bacteria to ***anoxic*** (oxygen free) environments found in deep-water sediments.

[*Continued on next page*]

6) The greenhouse warming associated with the vast outpouring of volcanic debris, however, would have warmed the ocean surface, thereby significantly reducing the amount of oxygen that seawater would absorb. This condition favors deep-sea anaerobic bacteria, which generate toxic ***hydrogen sulfide*** as a waste gas. As these organisms proliferated, the amount of hydrogen sulfide dissolved in seawater would have steadily increased. Eventually, the concentration of hydrogen sulfide would have reached a threshold and great bubbles of this toxin would have exploded into the atmosphere. On land, hydrogen sulfide was lethal to both plants and animals, but oxygen-breathing marine life would have been hit hardest.

7) How plausible is this scenario? Remember that the ideas that were just described represent a hypothesis, a tentative explanation regarding a set of observations. Additional research about this and other hypotheses that relate to the Permian extinction continues.

Tarbuck, E. J., & Lutgens, F. K. (2012). *Earth Science* (12th ed.). New York: Learning Solutions.

Permian Extinction: R4GC

This is a test of your reading comprehension. Read the question and select the best answer. Circle your answer.

1. What is the main idea of the passage?

    a. A description of the results of the Permian Extinction
    b. An introduction to the possible causes of the Permian Extinction
    c. A description of the events that led to the Siberian Traps
    d. An argument for continued research on extinction events

2. "Another possible mechanism for the Permian extinction was the voluminous eruptions of basaltic lavas which began about 251 million years ago..."

    In paragraph 3, what does the word **voluminous** mean?

    a. Explosive
    b. Extremely hot
    c. Loud and disruptive
    d. Having great size

3. "…and the emissions of ***sulfur dioxide*** probably resulted in copious amounts of acid rain."

    In paragraph 3, what does the word **copious** mean?

    a. Generous
    b. Wide-spread
    c. Dangerous
    d. Large in quantity

[*continued on the next page*]

4. "This condition favors deep-sea anaerobic bacteria, which generate toxic ***hydrogen sulfide*** as a waste gas."

   In paragraph 6, what does **this condition** refer to?

   a. Increased volcanism
   b. Warmed ocean
   c. Meteorite impact
   d. Anaerobic metabolism

5. What is the main idea of paragraph 5?

   a. An explanation of normal ocean conditions
   b. An introduction to a limitation to a hypothesis
   c. A description of a result of a meteorite impact
   d. An argument for a new interpretation of data

6. What are proposed causes of the Permian Extinction discussed in this passage? [*select all that apply*]

   a. Dinosaurs
   b. Overpopulation
   c. Increase in ocean size
   d. Predation
   e. Climate change

7. What is the author's opinion toward the hypotheses?

   a. The last hypothesis discussed is the most probable.
   b. The Permian Extinction was most likely caused by a combination of everything.
   c. More research is needed to prove or disprove the hypotheses.
   d. None of the hypotheses can account for the events of the Permian Extinction.

8. According to the passage, what might have killed oxygen-breathing marine life?
   a. Hydrogen sulfide gas
   b. Sulfur dioxide gas
   c. Carbon dioxide
   d. Predators

Reading for General Comprehension Answer Key:

Permian Extinction
1. b [mi]
2. d [voc]
3. d [voc]
4. b [pro]
5. a [mi]
6. d, e [det]
7. c [det]
8. a [det]

R4GC Table of Specifications						
	Main idea	Detail	Vocabulary	Pronoun	Total	Percentage
Testlet						
Permian extinction	2	3	2	1	8	33.30%
	1, 5	6, 7, 8	2, 3	4		
Deep-sea hydrothermal vents	2	3	2	1	8	33.30%
	1, 4	3, 6, 7	2, 5	8		
Aerosols	2	3	2	1	8	33.30%
	1, 8	2, 3, 4	6, 7	5		
Total	6	9	6	3	24	
Percentage	25%	37.50%	25%	12.50%		100%

Appendix C

Participant Number: _____

Permian Extinction: R2L

This is a test of your ability to learn from what you have read. Read the questions and circle the correct answer or answers. You have 12 minutes.

1. How is the passage organized?

    a. It lists the effects of the Permian Extinction.
    b. It presents several hypotheses for the Permian Extinction.
    c. It describes the pros and cons of several hypotheses.
    d. It provides a timeline of the events leading up to the Permian Extinction.

2. Why does the passage mention the Siberian Traps?

    a. To exemplify the results of volcanism
    b. To introduce a possible cause of the Permian Extinction
    c. To describe one of the causes of the dinosaur extinction
    d. To provide an explanation for increased competition among species

3. What can be inferred about plants during the Permian Extinction?

    a. They were unable to survive.
    b. They were a factor that led to the Permian Extinction.
    c. They were not affected because they are heat tolerant.
    d. They removed too much carbon dioxide from the air.

[*continued on the next page*]

4. Scientists who believe that the dinosaurs were killed by an impact cite the discovery that a layer of rock dates precisely to the extinction event and which is rich in the metal iridium. This layer is found all over the world, on land and in the oceans. Iridium is rare on Earth but it's found in meteorites at the same concentration as in this layer. This led scientists to postulate that the iridium was scattered worldwide when a comet or asteroid struck somewhere on Earth and then vaporized.

   What hypothesis will receive support if scientists find a second layer of meteorite debris?

   a. That the Permian extinction event was not as serious as previously assumed.
   b. That dinosaurs were not destroyed by an impact event.
   c. That the Permian extinction event was caused by an impact event.
   d. That iridium caused the Permian extinction event.

5. What caused the reduction of oxygen in the ocean?

   a. The spread of anaerobic bacteria
   b. The increase in hydrogen sulfide
   c. The impact of a meteorite
   d. The increased warmth of the water

6. What are two proposed causes of the Permian extinction?
   [*select two*]

   a. Climate change
   b. Overpopulation
   c. Volcanism
   d. Increase in ocean size

7. What is the author's view about the cause of the Permian Extinction?

   a. It resulted from the rapid release of sulfur dioxide.
   b. It resulted from a combination of predation and environmental factors.
   c. Discovering the cause will require additional research.
   d. Scientists will never truly know what caused it.

[*continued on the next page*]

8. This is a quote taken from a reporter who wrote a National Geographic article about the Permian Extinction:

   "Each scientist I met left me thinking that he or she was a clue or two away from solving the crime. But as Doug Erwin of the Smithsonian cautioned me, 'the truth is sometimes untidy.' The Permian extinction reminds him of Agatha Christie's *Murder on the Orient Express*, in which a corpse with 12 knife wounds is discovered on a train. Twelve different killers had conspired to slay the victim. "

   What can be inferred about Doug's view of the Permian Extinction:

   a. The mystery will most likely never be completely solved.
   b. It was probably the result of multiple factors.
   c. It was a messy event.
   d. He believes he will solve the mystery of the Permian Extinction.

Reading to Learn Answer Key

Permian Extinction:

1. b [as]
2. a [as]
3. a [p-inf]
4. c [ui]
5. d [e-inf]
6. a, c [mem]
7. c [mem]
8. b [ui]

R2L Table of Specifications							
	Awareness of structure	Memory for text	Use of information	Explanatory inference	Predictive inference	*Total*	*Percentage*
Testlet							
Permian extinction	2	2	2	1	1	8	33.33%
	1, 2	6, 7	4, 8	5	3		
Deep-sea hydrothermal vents	2	2	2	1	1	8	33.33%
	9, 15	10, 13	14, 16	12	11		
Aerosols & geoengineering	2	2	2	1	1	8	33.33%
	21, 23	17, 19	18, 24	20	22		
Total	6	6	6	3	3	24	
Percentage	25%	25%	25%	12.50%	12.50%		100%

Appendix D
Post-test Questionnaire

The purpose of this questionnaire is to find out your opinions about the reading test that you just took. Your responses will be used to help improve the test. All responses are confidential.

Please put an 'x' in the appropriate box:

	Strongly Disagree	Disagree	Agree	Strongly Agree
	1	2	3	4
1. The reading was easy for me.				
2. The questions were easy for me.				
3. I have a lot of prior knowledge of the topic.				
4. I will remember the information in the reading.				
5. I am learning about this topic in one of my university classes this semester.				
6. I have previously studied this topic in school.				
7. I have learned about this topic from the reading.				
8. I understand this topic as a result of the reading.				
9. I read this passage similarly to how I read for my university classes.				
10. I read this passage similarly to how I read for my university tests.				
11. I read this passage similarly to how I read to prepare for group discussions or study sessions.				