

The appearance of coherence: Using cohesive properties of readers' constructed responses to predict individual differences

La apariencia de coherencia: Uso de propiedades cohesivas de las respuestas construidas por los lectores para predecir las diferencias individuales

Lauren E. Flynn

UNIVERSITY OF NEW HAMPSHIRE
DURHAM, NH
USA
lauren.flynn@unh.edu

Danielle S. McNamara

ARIZONA STATE UNIVERSITY
TEMPE, AZ
USA
dsmcnamara1@gmail.com

Kathryn S. McCarthy

GEORGIA STATE UNIVERSITY
ATLANTA, GA
USA
kmcCarthy12@gsu.edu

Joseph P. Magliano

GEORGIA STATE UNIVERSITY
ATLANTA, GA
USA
jmagliano@gsu.edu

Laura K. Allen

UNIVERSITY OF NEW HAMPSHIRE
DURHAM, NH
USA
laura.allen@unh.edu

Recibido: 21-VI-2021 / **Aceptado:** 18-X-2021

DOI: 10.4067/S0718-09342021000301061

Abstract

Successful text comprehension requires readers to engage in a number of coherence-building processes. This study examined how analyzing the cohesion of students' constructed responses can be used to evaluate these coherence-building processes and the extent to which they vary across readers' individual differences and across types of texts. We posed two primary research questions: 1) Can we predict individual differences in working memory and reading skill based on the cohesion of students' constructed responses to text? 2) Do the relations between individual differences and cohesion vary as a function of genre? Participants ($n = 119$) generated constructed responses while reading history and science texts and completed reading skill and working memory assessments. The current study leveraged natural language processing (NLP) techniques to analyze the cohesion of readers' constructed responses, using cohesion as a proxy for assessing the coherence of their mental representations of the texts. Cohesion was measured at the sentence, paragraph, and synonym levels. Machine learning models showed that linguistic indices related to cohesion were significant predictors of both working memory and reading skill. Additional quantitative and qualitative inspection revealed that the relations between individual differences and coherence-building

processes varied depending on the text's genre. These findings indicate that the interaction between genre and individual differences may be used to model coherence-building processes during reading. This study has important implications for the realm of educational technology such as in the implementation of stealth assessments to predict students' cognitive abilities.

Key Words: Reading comprehension, individual differences, think-aloud methods, genre effects, coherence-building processes.

Resumen

La comprensión exitosa del texto requiere que los lectores se involucren en una serie de procesos de construcción de coherencia. Este estudio examina cómo se puede utilizar el análisis de la cohesión de las respuestas construidas por los estudiantes para evaluar estos procesos de construcción de coherencia y el grado en que varían entre las diferencias individuales de los lectores y entre los distintos tipos de textos. Planteamos dos preguntas principales de investigación: 1) ¿Podemos predecir las diferencias individuales en la memoria de trabajo y en las habilidades de lectura basadas en la cohesión de las respuestas construidas por los estudiantes al texto? 2) ¿Varían las relaciones entre las diferencias individuales y la cohesión en función del género? Los participantes ($n = 119$) generaron respuestas construidas mientras leían textos de historia y ciencia y completaron evaluaciones de habilidad de lectura y memoria de trabajo. El estudio actual aprovechó las técnicas de procesamiento del lenguaje natural (NLP) para analizar la cohesión de las respuestas construidas por los lectores, utilizando la cohesión como representación para evaluar la coherencia de sus representaciones mentales de los textos. La cohesión se midió en los niveles de frase, párrafo y sinónimo. Los modelos de aprendizaje automático mostraron que los índices lingüísticos relacionados con la cohesión eran predictores significativos tanto de la memoria de trabajo como de la capacidad de lectura. Una inspección cuantitativa y cualitativa adicional reveló que las relaciones entre las diferencias individuales y los procesos de creación de coherencia variaban según el género del texto. Estos hallazgos indican que la interacción entre el género y las diferencias individuales puede ser utilizada para modelar los procesos de construcción de coherencia durante la lectura. Este estudio tiene implicaciones importantes para el ámbito de la tecnología educativa, como en la implementación de evaluaciones sigilosas para predecir las habilidades cognitivas de los estudiantes.

Palabras Clave: Comprensión de lectura, diferencias individuales, métodos de pensar en voz alta, efectos de género, procesos de construcción de coherencia.

INTRODUCTION

Successful text comprehension occurs when a reader has constructed a 'coherent' and meaningful mental representation of a text (McNamara & Magliano, 2009). Coherence, or the interconnectedness of this mental representation, involves establishing how explicitly conveyed content is semantically related in the mental representation. These semantic relationships are established via explicit content in the text in combination with inference generation on the part of the reader (Graesser, Singer & Trabasso, 1994). In the context of reading simple narrative texts, inferences may be heavily supported by implicit processes that do not require conscious effortful processing (Myers & O'Brien, 1998). However, in the context of challenging texts or

texts in which readers have relatively little relevant prior knowledge, establishing these inferences may require substantial effort (McNamara, 2004; McNamara & Magliano, 2009). As such, there is considerable research on individual differences that are associated with ‘coherence-building’ (e.g., Whitney, Ritchie & Clark, 1991; Magliano & Millis, 2003; Klauda & Guthrie, 2008; Magliano, Higgs, Santuzzi, Tonks, O’ Reilly, Sabatini, Feller, Kopatich, Ray & Parker, 2020).

Text comprehension is supported by cognitive skills that are specific to reading as well as those that can be construed as relatively domain-general (Kopatich, Magliano, Millis, Parker & Ray, 2019). Indeed, prior work shows that reading skill (i.e., performance on standardized tests of comprehension proficiency) is a predictor of comprehension such that skilled readers are more likely to engage in increased coherence-building processes during reading as compared to less skilled readers (Magliano & Millis, 2003; Klauda & Guthrie, 2008). Additionally, working memory capacity is a relative domain-general aspect of cognition that may support the ability to develop coherence during reading (Turner & Engle, 1989). In the present study, we explored the extent to which a novel and informative approach to assessing coherence-building is predictive of individual differences in reading skill and working memory capacity.

To this end, we examined coherence-building by analyzing students’ constructed responses that they produced as they read. ‘Constructed responses’ such as think-aloud protocols are online measures that can provide a direct window into coherence-building processes during reading (Trabasso & Magliano, 1996; Magliano & Millis, 2003). Think-aloud procedures typically involve readers being intermittently interrupted and prompted to report their thoughts as they come to mind (Ericsson & Simon, 1980; Pressley & Afflerbach, 1995). One way of leveraging these constructed responses is to score their quality in terms of the extent to which the reader appears to be constructing a more global understanding of the text (Coté & Goldman, 1999). The quality of constructed responses has been associated with readers’ comprehension of the target texts (McCarthy et al., 2020) as well as various individual differences in reading skill (Magliano & Millis, 2003; Coiro, 2011), prior knowledge (Pressley, Wood, Woloshyn, Martin, King & Menke, 1992), and working memory (Bohn-Gettler & Kendeou, 2014; Whitney et al., 1991). An alternative approach to leveraging constructed responses is by analyzing their linguistic properties via natural language processing (NLP) methodologies (Magliano & Graesser, 2012; McNamara, Allen, McCarthy & Baylan, 2018). We used an NLP approach in the current study to assess the extent that measures of coherence-building as revealed by constructed responses are predictive of individual differences in reading skill and working memory. The study uses an approach developed by Allen, Snow and McNamara (2015), Allen, Jacovina and McNamara (2016) that uses the ‘cohesion’ of students’ constructed responses to model individual differences. More specifically, this study examines if linguistic features of constructed responses are

predictive of individual differences and the extent to which these relations vary across genres.

We conducted this study in honor of Giovanni Parodi Sweis, a leader, mentor, and generous colleague to innumerable students and researchers seeking to better understand language, comprehension, and literacy. This study embodies Giovanni by combining multiple methods germane to computational linguistics, corpus linguistics, and discourse analyses to further our understanding of individual differences in students' ability to comprehend different text genres. With heavy hearts, we remember Giovanni fondly, as a generous and kind mentor, and friend.

1. Individual differences in reading comprehension

The current study focuses on two individual differences: reading skill and working memory. These individual differences have been linked to performance in a variety of reading comprehension contexts, such as multiple-document comprehension (Barzilai & Strømsø, 2018; Stadtler, Bromme & Rouet, 2018), science comprehension (Allen, Snow, Crossley, Jackson & McNamara, 2014; McNamara, Graesser, McCarthy & Cai, 20014; O'Reilly, Sinclair & McNamara, 2004), and narrative text processing (Bower & Morrow, 1990; van den Broek, 1994).

Reading skill refers to a number of specific cognitive skills that are associated with successful reading processes, ranging from the decoding of individual words (Perfetti, 1985) to syntactic knowledge and the skills necessary for making connections with prior knowledge (Gernsbacher, Varner & Faust, 1990; Hannon & Daneman, 2001). This ability to generate connections amongst items contained in the text, as well as with prior knowledge is critical for deep comprehension (Oakhill & Yuill, 1986; Hannon & Daneman, 2001). Reading skill is correlated with increased comprehension of texts and can be significantly impacted based on the properties of the text being read (Ozuru, Dempsey & McNamara, 2009). Importantly, reading skill can be improved through training which helps readers to generate stronger connections with the text and consequently build a more coherent representation during reading (e.g., McNamara, 2004).

In addition to reading-specific skills, differences in working memory capacity may also affect how readers process and store information from texts (Daneman & Carpenter, 1980). Performance on working memory tasks is positively correlated with measures of text comprehension. Researchers have argued that readers with lower performance on working memory capacity tasks may struggle to integrate texts into their mental representations due to weaker attention-control mechanisms (Daneman & Carpenter, 1980; Kane & McVay, 2012). For example, working memory has been found to be a necessary component in integrative computations of pronominal referents (Daneman & Carpenter, 1980). Another study found that children with higher working memory capacities performed better on reading comprehension tasks involving surface-

level and deep comprehension questions (Borella & de Ribaupierre, 2014). Thus, individuals who perform well on working memory tasks may be better equipped to comprehend texts, both at shallow and deep levels, by having prior information more readily accessible. This may, in turn, make it easier for readers to connect concepts and ideas across a text, aiding in the construction of coherent mental representations (Daneman & Carpenter, 1980).

Whitney et al. (1991) provided some evidence that individual differences in working memory capacity are related to coherence-building. They had college students think aloud while reading narrative texts. They also measured individual differences in working memory capacity with a span task. Lower span readers tended to produce thoughts that focused on establishing local coherence, whereas higher span readers tended to produce thoughts that emphasized global coherence until they got to the end of the story where they produced statements that reflected more specific interpretations of the narrative events. Whitney et al. argued that individual differences in reading span led to readers differentially trading off between establishing local (between immediate sentences) and global (between distally related sentence) coherence.

2. Assessing the cohesion of constructed responses

Although there are multiple linguistic features that can be used to characterize constructed responses, the cohesion of these responses is particularly relevant in providing insight into the coherence of the readers' mental representation of the text. 'Cohesion' refers to the explicit cues that establish connections amongst text content (McNamara et al., 2014). The cohesion of constructed responses can be assessed in terms of overlap between utterances, such as content words (e.g., nouns, verbs). Highly connected ideas about the text correspond to higher coherence, and in turn, deeper levels of comprehension (Allen et al., 2016). As such, the cohesion of constructed responses generated during reading has recently been proposed as a proxy for the coherence of the readers' mental representations (Allen et al., 2015, 2016). For this reason, the cohesion of constructed responses produced during the coherence-building process should be influenced by individual differences in the same way that it affects the coherence of readers' mental representations of text.

Allen et al. (2015) developed a model to predict performance on the Gates-MacGinitie Reading Test (MacGinitie, MacGinitie, Maria & Dreyer, 2000) based on the linguistic features of students' constructed responses to texts. In this study, participants were asked to produce self-explanations at key points while reading a science text. NLP techniques were then used to analyze these self-explanations along a number of linguistic dimensions (e.g., word, sentence, and cohesion). These linguistic indices were then used in a linear model to predict performance on the Gates-MacGinitie assessment. Skilled readers were more likely to generate self-explanations containing

greater semantically overlap and more explicit connections between their self-explanations. This study provided preliminary evidence that reading comprehension skills can be modeled based on the linguistic properties of readers' constructed responses. Allen et al. (2016) extended this work by examining how the cohesion of constructed responses varied across instructions to engage in different reading strategies (i.e., paraphrasing, self-explanations). When readers were prompted with self-explanation instructions, they produced more diverse language and global connectives in their constructed responses (Allen et al., 2016). This provides further evidence that the cohesion of constructed responses can serve as a proxy for coherence-building processes.

3. Genre effects on comprehension

Beyond individual differences, comprehension can also be impacted by the nature of the text being read. For example, recent meta-analyses have shown that readers process, understand, and recall narrative texts differently than expository texts (Clinton, Taylor, Bajpayee, Davison, Carlson & Seipel, 2020; Mar, Li, Nguyen & Ta, 2021).

While the nature of genre is debated, it is generally agreed upon that different text genres have different purposes, are marked by different linguistic features, and can be reliably distinguished by readers. For instance, social studies and science texts have more challenging words and sentence structures compared to narrative texts (McNamara, Graesser & Louwse, 2012). These aspects of text difficulty further translate to differences in text cohesion. For example, narratives tend to contain less referential cohesion, but more connectives compared to science texts. This linguistic variation drives differences in the coherence-building processes that afford comprehension (McNamara et al., 2012).

These features of texts also have top-down effects on individuals' coherence-building processes and subsequent comprehension (McNamara et al., 2012; Zwaan, 1994). For example, readers' genre expectations have been shown to affect the cognitive top-down processes that occur while they construct and update their mental representations during reading (Kintsch, 1992; Parodi, 2014; Schmitz, Gräesel & Rothstein, 2017). Zwaan (1994) asked participants to read a text with the expectation that it was either a literary story or a news story. Although the text was identical in both conditions, participants reading from the literary perspective had longer reading times and better memory for surface-level information but had a weaker memory for situational information compared to those reading from the news perspective. Genre has also been found to be an important factor in how well readers integrate text content with prior knowledge (Wolfe & Mienko, 2007). This, and similar studies, provide evidence that individuals' knowledge and expectations of genres can influence their processing and ultimate comprehension of texts (Rozimela, 2014).

Although the majority of work in genre differences has focused on narrative versus expository texts, work in genre studies and disciplinary literacies suggest that coherence-building processes vary across more nuanced genre categories. Indeed, Parodi (2010, 2014, 2015) identified key differences between the genres of social sciences/humanities and basic sciences in academic discourses. These differences can be distilled into three major components: 1) the ways in which pedagogic devices are articulated, 2) the types of methodological procedures being transferred, and 3) the extent to which the knowledge expressed is either abstract or concrete. Researchers interested in discipline-specific comprehension challenges have further identified that science and history texts tend to be written in ways that are inconsiderate of students' knowledge and reading skills (e.g., Moje & Speyer, 2008), but also that each type of text presents specific challenges in establishing and maintaining coherence (Beck & McKeown, 1988; Goldman & Bisanz, 2002). However, few studies have explicitly examined these coherence-building and processing differences side-by-side.

Importantly, an abundance of work in the text comprehension literature has studied these genres in relatively isolated contexts. For example, narrative texts are rarely measured in the same experimental contexts and studies as expository texts; science and history texts are generally studied in different contexts and by different researchers. Also, the methodologies and assessments that are used to examine the various genres tend to vary widely. It is therefore an open question as to whether and how these individual differences and coherence-building processes interact differently across genres. Thus, a novel contribution of the current study is our examination of whether individual differences and genre interact to impact coherence-building processes during reading, as measured by cohesion dimensions.

4. Current study

The current study assesses the coherence of readers' mental representations during the reading process using the cohesive properties of their constructed responses. Specifically, we examined the cohesion of responses constructed by participants while they read texts of different genres (i.e., history, science) from an archival data set (Magliano, Durik & Holt, 2011). We posed two primary research questions: 1) Can we predict individual differences in working memory and reading skill based on the cohesion of students' constructed responses to text? 2) Do the relations between individual differences and cohesion vary as a function of genre?

5. Method

5.1. Participants

Participants from this archival data (Magliano et al., 2011) set were undergraduate students ($n = 119$) enrolled in an introduction to psychology course from a large

Midwestern University who received course credit for their participation. Of these participants, 45% were female. In addition, 55% were Caucasian, 17% were African American, 7% were Hispanic, 4% were Asian American, 1% were Native American, and 16% reported that their ethnicity did not fall into any of the above options provided. We include ethnicity in our description of participants for the purposes of replication and generalization across different populations in extensions of this work. Three participants did not complete the working memory task and so they were not included in any working memory analyses.

5.2. Materials

5.2.1. Text passages

Participants were randomly assigned to read one of two history texts (i.e., economic causes of the Civil War or garment worker labor strikes) and one of two science texts (i.e., erosion or geographic isolation and the evolution of species). The texts were relatively matched for length (299-390 words) and readability (Flesch-Kincaid Grade Level from 9.2-11.5). Participants each read only one history text and one science (see Appendix for example passages from the text) to prevent fatigue effects.

5.2.2. Reading strategy assessment tool (RSAT)

The protocols were collected in the context of the Reading Strategy Assessment Tool (RSAT), which is an online data collection tool for collecting constructed responses while reading texts (Magliano et al., 2011). RSAT contains instruction and practice for producing constructed responses. The instructions and practice were delivered by the experimenter and were presented on paper. Participants were instructed to ‘think aloud’ in response to the prompts. They were told that thinking aloud involves producing whatever comes to mind in terms of their understanding of the sentence that was just read and in the context of the larger text. The instructions specify that participants should avoid restricting their response to short phrases (e.g., ‘OK, makes sense’) and making comments for humorous purposes. Practice consisted of reading a short 5-sentence text on the expansion of railroads in the American West during the middle 1800s. The thoughts were produced in writing. The experimenter looked at responses and restricted corrective feedback to situations in which the participants produced short responses during practice (‘OK, makes sense’). When this happened, the experimenter restated the instructions on thinking aloud.

After completing practice, participants were transitioned to the RSAT interface to complete the task on the computer. In the context of the present study, sentences were presented one at a time in isolation and participants were prompted to generate constructed responses after every sentence. When a constructed response was solicited, the sentence that was just read was removed from the screen. This decision was made to force students to draw upon their mental model to produce the constructed response.

The sentences were presented at the top of the screen and a 'Next' button appeared above a textbox near the bottom. However, while a sentence was present, the textbox was not active. After pressing the next button when there was a sentence present, the sentence disappeared and 'What are you thinking now?' appeared above the text box. The textbox then became active, and participants could type their responses. Participants pressed the next button to advance to the next sentence. The phrase 'Next Paragraph' appeared when there were transitions to new paragraphs prior to the first sentences in the paragraphs.

5.2.3. Working memory test

A computerized version of the operation span (OSPAN) task was used to measure participants' working memory (Unsworth, Heitz, Schrock & Engle, 2005). Sets of mathematical operations are presented and participants are asked to indicate if the operation is true or false (e.g., Is $9 / 3 + 1 = 6$?; False). Between these mathematical operation trials, a random letter of the alphabet is shown on the screen for participants to remember for later recall. A practice trial was first administered to accustom subjects to the task before they moved on to the experimental trials. Working memory scores were computed by how many correct letters participants could recall in the same order they were originally presented in. Possible scores ranged from 0 to 75.

5.2.4. Reading skill assessment

Reading skill was measured using the Gates-MacGinitie Reading Test (Level 10/12, Form T; MacGinitie et al., 2000). This standardized test presents participants with 48 multiple-choice questions about expository passages. Each reading passage was associated with 2-6 comprehension questions that assess both shallow and deep comprehension by requiring them to make inferences. All participants were given practice questions beforehand and then had 20 minutes to complete the test. Raw scores were collected with 48 being the highest possible score.

5.3. Procedure

The study took place in two sessions. In session 1, participants completed a series of individual difference assessments that were administered in paper and pencil formats (the working memory assessment was distributed in session 2 because it did not fit the paper and pencil format of session 1). Though a range of additional measures were collected in the larger study, the current study focuses exclusively on reading skill and working memory (i.e., Gates-MacGinitie Reading Test and OSPAN). Session 2 involved activities that were administered on the computer, specifically OSPAN and RSAT. Participants first completed the OSPAN followed by the constructed response task in RSAT, wherein the order of presented texts were randomized.

5.4. Linguistic analyses

For the purpose of calculating the linguistic properties of students' constructed responses, the individual think-aloud responses were aggregated across participants and saved as text files (Allen et al., 2015, 2016). Specifically, the constructed responses produced at the ends of each sentence were aggregated into a single file for each genre of text. Paragraph breaks were added to the aggregated file to preserve the structure and mimic the organization of the source texts. Thus, the aggregated responses resembled the paragraph structure of the source texts. Each participant had two aggregated constructed response files – one for each of the texts they read.

The aggregated constructed response files were analyzed using the 'Tool for the Automatic Analysis of Text Cohesion' (TAACO; Crossley, Paquette, Dascalu, McNamara & Baker, 2016), an automated natural language processing tool that processes multiple text files to report various features related to cohesion. We analyzed the cohesion of the constructed responses in terms of the word lemmas and their synonyms at both the sentence and paragraph levels (see Table 1 for index descriptions). Sentence-level cohesion (i.e., adjacent sentence overlap, adjacent 2-sentence overlap) is intended to be indicative of local connections that readers make whereas paragraph-level cohesion (i.e., adjacent paragraph overlap, adjacent 2-paragraph overlap) is indicative of readers' generation of more distal connections. Whereas adjacent sentence overlap refers to the overlap of lemma types across consecutive sentences, adjacent paragraph overlap refers to the cohesive overlap of linguistic markers over consecutive paragraphs. Cohesion was assessed using both the words (i.e., lemmas) and their synonyms to provide estimates of both explicit repetitions from the text as well as their semantic associates.

Table 1. Cohesion analysis indices.

Cohesion Levels	Index	Description
Sentence	Adjacent Sentence Overlap	Number of lemma types that occur at least once in the next sentence
Sentence	Adjacent 2-Sentence Overlap	Number of lemma types that occur at least once in the next two sentences
Paragraph	Adjacent Paragraph Overlap	Number of lemma types that occur at least once in the next paragraph
Paragraph	Adjacent 2-Paragraph Overlap	Number of lemma types that occur at least once in the next two paragraphs
Synonym	Adj. Sentence Noun Synonym Overlap	Average sentence to sentence overlap of noun synonyms
Synonym	Adj. Sentence Verb Synonym Overlap	Average sentence to sentence overlap of verb synonyms
Synonym	Adj. Paragraph Noun Synonym Overlap	Average paragraph to paragraph overlap of noun synonyms
Synonym	Adj. Paragraph Verb Synonym Overlap	Average paragraph to paragraph overlap of verb synonyms

6. Results

All aggregated constructed responses with fewer than 100 words (i.e., 8 students' responses) were removed from the following analyses because estimates of cohesion are not reliable with responses including fewer than 100 words (Crossley, 2018; Crossley & Kostyuk, 2017; Crossley et al., 2016). The remaining constructed responses produced by participants ranged in length from 102 to 793 words ($M = 279.82$; $SD = 128.59$). Table 2 provides the means, standard deviations, and correlations between the linguistic indices.

Table 2. Means and correlations between linguistic indices.

	Index	Mean	SD	1	2	3	4	5	6	7	8	9
1.	Number of Words	276.06	104.85	–								
2.	Adjacent Sentence Overlap	1.84	.98	.80	–							
3.	Adjacent 2-Sentence Overlap	2.79	1.35	.83	.99	–						
4.	Adjacent Paragraph Overlap	2.12	1.16	.92	.94	.95	–					
5.	Adjacent 2-Paragraph Overlap	3.17	1.62	.94	.92	.94	.99	–				
6.	Adjacent Sentence Noun Synonym Overlap	.36	.29	.56	.81	.78	.72	.70	–			
7.	Adjacent Sentence Verb Synonym Overlap	.34	.29	.71	.81	.81	.78	.78	.53	–		
8.	Adjacent Paragraph Noun Synonym Overlap	.42	.34	.67	.81	.80	.81	.79	.96	.55	–	
9.	Adjacent Paragraph Noun Synonym Overlap	.42	.38	.79	.74	.76	.82	.82	.47	.93	.56	–

Participants scored 52.79% ($M = 39.49$; $SD = 16.47$) and 58.81% ($M = 28.01$; $SD = 8.34$) on the working memory and reading skill assessments, respectively. Performance on these two measures was not significantly correlated ($r = 0.10$, $p = 0.27$).

6.1. Relations between individual differences and constructed response cohesion

Our first research question examined the magnitude of the relations between individual differences (i.e., working memory, reading skill) and the cohesion of participants' constructed responses generated during reading. Table 3 provides the correlations between the individual difference measures and cohesion indices aggregated across genre. The overall correlations indicated that the cohesion observed in readers' constructed responses was more strongly related to their reading skill than their scores on the working memory task. Indeed, six of the eight cohesion variables were correlated with reading skill whereas none of the cohesion variables were significantly related to working memory. This provides preliminary support that we can

predict individual differences amongst readers based on the cohesion of their constructed responses generated during reading.

Table 3. Correlations between individual differences and cohesion indices.

Cohesion Index	Working Memory	Reading Skill
Adjacent Sentence Overlap	-.14	.22*
Adjacent 2-Sentence Overlap	-.10	.23*
Adjacent Paragraph Overlap	-.14	.25**
Adjacent 2-Paragraph Overlap	-.11	.27**
Adjacent Sentence Noun Synonym Overlap	-.05	.25**
Adjacent Sentence Verb Synonym Overlap	-.15	.14
Adjacent Paragraph Noun Synonym Overlap	-.07	.27***
Adjacent Paragraph Verb Synonym Overlap	-.17	.16

Notes: $p < .0001$ ***; $p < .01$ **; $p < .05$ *

6.1.1. Machine learning models

To determine how predictive the cohesion indices were of the individual differences, we extended the correlation analyses by developing machine learning models to predict scores on the individual difference measures from the constructed response cohesion variables (see Table 1 for the list of variables used). We used four different models (Linear Regression, Random Forest, Support Vector Machine Polynomial, Bayesian Regularized Neural Networks; see Table 4 for model details) because linguistic features often exhibit nonlinear and complex relations, which are better able to be captured by different models (Kuhn, 2019). The four models were selected because they have been shown to successfully model individual differences from linguistic features in texts (Öncel, Flynn, Sonia, Barker, Lindsay, McClure, McNamara & Allen, 2021).

Linear Regression models predict continuous variables using slope-intercepts (Rong & Bao-wen, 2018). Random Forest models are used to avoid the overfitting problem commonly associated with decision trees by constructing multiple decision trees and using majority voting (Breiman, 2001). Support Vector Machine (SVM) models work by creating separate planes to break data up by their features. However, the current study uses the polynomial kernel, a more generalized representation of the typical linear SVM model (Tong & Koller, 2001). Finally, the Bayesian Regularized Neural Networks (BRNN) functions by assigning weights and biases to a probability distribution (Karklin & Lewicki, 2005). Table 4 provides additional information about the models presented in this study and classifies them by type and tuning parameter (Kuhn, 2019). Aside from linear regression, all models analyze data nonlinearly to account for the nonlinearity of language data. Predictions were then evaluated using R^2 and the root-mean-square error (RMSE). To avoid overfitting and maximize the chance of generalizability, all models in the current study utilized a 10-fold cross-validation procedure and were repeated three times.

Table 4. Machine learning model descriptions.

Model	Type	Tuning Parameter
Linear Regression	Regression	intercept
Random Forest	Classification, Regression	mtry
Support Vector Machine Polynomial (SVM Poly)	Classification, Regression	degree, scale, C
Bayesian Regularized Neural Networks (BRNN)	Regression	neurons

As shown in Tables 5 and 6, the cohesion indices were able to account for approximately 16-18% of the variance in the working memory and reading skill scores, respectively. In particular, the BRNN model ($R^2 = 0.18$, $RMSE = 16.81$) performed best for the working memory scores and the Random Forest model ($R^2 = 0.16$, $RMSE = 8.04$) performed best for reading skill. This indicates that individuals who performed better on the working memory and reading skill tests were engaging in different types of coherence-building processes compared to those with lower scores. Critically, the most important variable in each of the best-performing models was not significantly correlated with the dependent variable. This suggests the presence of nonlinear relations in the data that may have otherwise been missed by relying solely on linear models, such as correlations and regressions.

Table 5. Machine-learning models for working memory.

Model	R^2	RMSE	Most Important Variable
Linear Regression	0.12	16.77	Adjacent Sentence Overlap
Random Forest	0.13	17.19	Adjacent Paragraph Verb Synonym Overlap
SVM Poly	0.17	16.44	Adjacent Paragraph Verb Synonym Overlap
BRNN	0.18	16.81	Adjacent Paragraph Verb Synonym Overlap

Table 6. Machine-learning models for reading skill.

Model	R^2	RMSE	Most Important Variable
Linear Regression	0.09	8.30	Adjacent Sentence Noun Synonym Overlap
Random Forest	0.16	8.04	Adjacent Paragraph Verb Synonym Overlap
SVM Poly	0.14	8.06	Adjacent Paragraph Verb Synonym Overlap
BRNN	0.14	8.07	Adjacent Paragraph Verb Synonym Overlap

6.2. Exploration of genre effects

We next examined the correlations between individual differences and cohesion as a function of genre (i.e., history, science). As shown in Table 7, the stability of these correlations varied across genre. We found significant, weak, negative relations between cohesion and working memory (WM) for history texts, but not for science texts. Conversely, significant, but weak positive correlations were observed between cohesion and reading skill (RS; see Table 7) for both the history and science texts, albeit with stronger relations within the history genre. These results suggest that history and science

texts may have subtle differential relations to coherence-building processes across skill levels (i.e., individual differences).

Table 7. Correlations between cohesion and individual differences by genre.

Cohesion Index	WM History	WM Science	RS History	RS Science
Adjacent Sentence Overlap	-.25**	.02	.19*	.15
Adjacent 2-Sentence Overlap	-.25**	.05	.20*	.17
Adjacent Paragraph Overlap	-.25**	.04	.23*	.17
Adjacent 2-Paragraph Overlap	-.23*	.06	.25**	.17
Adjacent Sentence Noun Synonym Overlap	-.10	.00	.21*	.20*
Adjacent Sentence Verb Synonym Overlap	-.24**	.06	.10	.12
Adjacent Paragraph Noun Synonym Overlap	-.12	.01	.22*	.22*
Adjacent Paragraph Verb Synonym Overlap	-.26**	.08	.11	.13

To complement the correlational analyses, linear mixed-effects models were used to examine interactions between individual differences and genre in the cohesion of participants' constructed responses. Participants were treated as random effects because they each had an observation for both a history and science text. Individual difference measures and genre were treated as fixed effects to predict cohesion indices. We selected the cohesion variables with the highest correlation from each level (i.e., sentence-level, paragraph-level, synonym cohesion) and individual difference measure (i.e., working memory, reading skill; see Table 9). For all three models (at each level of cohesion), there was a significant interaction between working memory and genre on constructed response cohesion ($p < .01$), indicating that relations between working memory and cohesion were moderated by genre. Conversely, we found no significant interactions between reading skill and genre. These results suggest that reading skill was consistently related to constructed response cohesion, regardless of the genre of the text. Overall, these results indicate that both individual differences and genre impact the constructed response cohesion at the sentence-level, paragraph-level, and synonym-levels.

To provide further insights into why the individual differences exhibited such different relations to cohesion in the history domain, we conducted an exploratory, qualitative analysis of the constructed responses with the cohesion variable that correlated the highest. The first five constructed profiles from samples of both low and high scorers on individual difference assessments (working memory, reading skill) appear in Table 8.

Table 8. History constructed response examples (corrected for spelling mistakes).

Individual Difference	Constructed Response Example
Low Working Memory (<i>High Cohesion</i>)	[1] People came to America hoping to get a better job and more money so that they can live a better life and have things that they might not have had before. But when they arrived they could only get the jobs that did not pay a lot and that nobody else would do.
	[2] In America it was about quality and quantity so they wanted to get things done good the first time so that they would not have to do them again.
	[3] These girls were used to sewing on a certain machine and had a hard time learning to use different methods of sewing. This caused them to work harder to get the tasks accomplished.
	[4] Since the women were highly skilled they were able to get the better positions because they were able to get the work completed.
	[5] They did not make as much money as they would have liked for the labor that they were doing. They did not get the better life that they thought they would.
High Working Memory (<i>Low Cohesion</i>)	[1] What were some of the many issues
	[2] Is this a wrong assumption
	[3] What were the other causes
	[4] When or how did these rifts become present
	[5] What was the U.S. like
Low Reading Skill (<i>Low Cohesion</i>)	[1] more then one thing caused war
	[2] people assume slavery
	[3] slavery was not single issue
	[4] there were a lot of reasons
	[5] what the US was like in 1800 s
High Reading Skill (<i>High Cohesion</i>)	[1] I'm thinking about Ellis island and the movies I've seen about people traveling over seas to find jobs and live the American dream
	[2] I'm thinking about my Victorian class because we just talking about the singer sewing machine yesterday and how it made life easier and helped the industrial revolution
	[3] I'm thinking that that s a lot of girls and given the title of the article that they are going to all work with these sewing machines and end up striking because of poor working conditions
	[4] I'm thinking that this sentence is accurate because if you were skilled you could find a way to make more money because you'd be more in demand
	[5] I'm thinking about what I've learned about the working class and how poor labor conditions and living conditions would be different from how they dreamed they'd be living

Note. The low working memory example with high cohesion (centered WM = -23.37; adj. paragraph verb synonym overlap = 3.00) is from the garment workers text and the high working memory with low cohesion (centered WM = 35.63; adj. paragraph verb synonym overlap = .04) example is from the Civil War text. The low reading skill example with low cohesion (centered RS = -10.25; adj. 2-paragraph overlap = .82) is from the Civil War text and the high reading skill with high cohesion (centered RS = 9.75; adj. 2-paragraph overlap = 9.09) example is from the garment workers text.

Table 9. Mixed-effects models.

Predictors	adjacent overlap all sent div seg			adjacent overlap all para div seg			syn overlap para verb			adjacent overlap 2 all sent div seg			adjacent overlap 2 all para div seg			syn overlap para noun		
	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	1.98	1.78 – 2.18	<0.001	2.30	2.05 – 2.55	<0.001	0.47	0.39 – 0.56	<0.001	2.99	2.72 – 3.26	<0.001	3.44	3.11 – 3.77	<0.001	0.37	0.30 – 0.43	<0.001
ospan_cent	-0.02	-0.03 – -0.01	0.005	-0.02	-0.04 – -0.01	0.004	-0.01	-0.01 – -0.00	<0.001									
genre [science]	-0.26	-0.45 – -0.07	0.006	-0.32	-0.58 – -0.07	0.014	-0.10	-0.20 – 0.00	0.057	-0.35	-0.60 – -0.10	0.006	-0.46	-0.81 – -0.10	0.011	0.13	0.05 – 0.20	0.001
ospan_cent * genre [science]	0.02	0.01 – 0.03	0.001	0.03	0.01 – 0.04	0.002	0.01	0.01 – 0.02	<0.001									
gates_cent										0.04	0.00 – 0.07	0.025	0.06	0.02 – 0.10	0.005	0.01	0.00 – 0.02	0.034
gates_cent * genre [science]										-0.00	-0.03 – 0.03	0.874	-0.02	-0.06 – 0.03	0.437	0.00	-0.01 – 0.01	0.521
Random Effects																		
σ^2	0.51			0.96			0.14			0.93			1.84			0.08		
τ_{00}	0.71	snum		0.87	snum		0.07	snum		1.26	snum		1.52	snum		0.06	snum	
ICC	0.58			0.48			0.34			0.57			0.45			0.43		
N	116	snum		116	snum		116	snum		119	snum		119	snum		119	snum	
Observations	224			224			224			230			230			230		
Marginal R ² / Conditional R ²	0.047 / 0.603			0.049 / 0.502			0.066 / 0.386			0.050 / 0.596			0.062 / 0.487			0.075 / 0.476		

As shown in Table 8, the student with lower working memory scores seemed to provide more of a broad paraphrase of the content, rather than engage in more self-reflective thought. The individual with high working memory scores, conversely, wrote fewer words but asked more targeted, specific questions about the text. Question asking has been shown to be a particularly beneficial reading strategy that can aid in coherence-building and comprehension (Palincsar & Brown, 1984; Rosenshine, Meister & Chapman, 1996; Graesser, McNamara & VanLehn, 2005). Thus, this person may be more efficiently leveraging their working memory capacity with more useful responses to the text, as further evidenced by our quantitative results.

Regarding reading skill, Table 8 reveals that the less skilled readers made very few connections across the text and did not provide any indication of strategic processing. Conversely, the high skilled reader engaged in more metacognitive monitoring while reading, as the reader described thoughts regarding how the text related to personal knowledge and experiences. Overall, this qualitative examination of the constructed responses illustrates how individual differences in working memory and reading skill may manifest differently in the properties of constructed responses.

7. Discussion

A central assumption of the present study is that constructed responses reveal important insights into individual difference factors associated with comprehension (Allen et al., 2015, 2016). The present study builds on the robust body of work that has examined relations between individual difference measures and constructed responses (Whitney et al., 1991; Bohn-Gettler & Kendeou, 2014; Magliano et al., 2020) by leveraging novel NLP and machine learning methodologies. We leveraged an NLP approach developed by Allen et al. (2015, 2016) to analyze the cohesion of readers' constructed responses across two different text genres (i.e., history and science).

Cohesion was calculated for the constructed responses at three levels: sentence, paragraph, and synonym cohesion. Our primary objective was to examine if readers' individual differences in working memory and reading skill could be reliably predicted from the cohesion of their constructed responses while reading and the extent that these relations varied across genres.

7.1. Predicting individual differences

The results of our correlation analyses indicated that all but two of the cohesion indices were correlated with reading skill; however, none of the indices were correlated with working memory. This is consistent with theoretical frameworks that emphasize the importance of the proficiencies in linguistic processes over general resources like working memory (Perfetti & Stafura, 2014). Although working memory capacity has implications on the waxing and waning of activated knowledge during reading (Just & Carpenter, 1991), accurately representing content serves as the primary retrieval cue for that knowledge (Myers & O'Brien, 1998). Thus, proficiency in processes that support the construction of strong mental representations of the explicitly conveyed content have direct effects on coherence-building (Gernsbacher, 1997; Kintsch, 1998).

Despite these differences in the correlation analyses, the subsequent machine learning analyses indicated that both reading skill and working memory could be predicted by the constructed response cohesion variables. This suggests that variance associated with both reading skill and working memory was related to the ways in which students generated connections during reading. Importantly, the most important variable in the models was 'adjacent paragraph verb synonym', which was not significant in the correlational analyses. This suggests a more complex, nonlinear relationship between the cohesion indices and individual differences, which should be explored in future studies.

Finally, the qualitative analysis of the constructed responses yielded important insights into how individuals who scored lower and higher on the working memory and reading skill assessments approached coherence-building. With respect to working memory capacity, lower span readers tended to produce responses that reflected the local context, whereas higher span readers produced relatively shorter responses but reflected that they were sensitive to the global context. These results are similar to those reported by Whitney et al. (1991) in the context of thinking aloud while reading narratives. With respect to reading skill, less skilled readers did not demonstrate strategies that are typically reflective of coherence-building, such as bridging across explicitly conveyed ideas and elaborating based on knowledge external to the texts (McNamara, 2004). More skilled readers demonstrated evidence of metacognitive strategies that tend to be associated with successful comprehension (Pressley & Afflerbach, 1995). These strategies are important for determining when specific coherence-building strategies are important and need to be deployed (McNamara &

Magliano, 2009). The combination of computational analysis and more qualitative discourse analysis used in this study was critical in understanding the complex nature of the results.

7.2. Genre effects

The genre-level correlation analyses again revealed that constructed response cohesion was significantly correlated with reading skill for both history and science texts, indicating that the skilled readers generated more connections in their constructed responses regardless of the genre. However, the results were less clear for working memory – participants' scores on the working memory test were negatively related to constructed response cohesion in the history texts and unrelated to science texts. The mixed-effects models corroborated this finding as genre and working memory interacted for every level of cohesion (sentence, paragraph, synonym). These results are in line with prior work in text comprehension that has found differential processing amongst text genres (McNamara et al., 2012; Clinton et al., 2020; Mar et al., 2021). However, none of the interactions between reading skill and genre were significant.

Given the robust feature- and function-based differences found across text genres (Parodi, 2010, 2014, 2015), the history and science texts here likely had varying task demands from the readers. Despite these differences, however, reading skill seemed to have consistent and stable relations to the coherence-building processes measured by the cohesion analyses. Given that reading skill is a malleable skill (unlike working memory), this suggests that difficulties associated with coherence-building may be more easily remedied through strategy instruction (McNamara, 2004). Therefore, researchers should focus on reading skill interventions to improve reading comprehension, rather than attempting to focus on capacity limitations due to working memory capacity. Based on our findings, this approach may be more likely to be fruitful, as it would be more likely to transfer to new and varied discourse contexts. However, future research might consider systematically examining these relations between working memory, genre, and coherence-building processes to develop a more thorough understanding of how such individual differences impact the coherence-building process.

7.3. Implications for theories of reading and comprehension

Theories of comprehension universally specify that coherence-building is a foundation of comprehension (McNamara & Magliano, 2009), but most are agnostic regarding the extent that individual difference factors support coherence-building. In contrast, theories of reading typically specify these relationships (e.g., Cromley & Azevedo, 2007), but underspecify the nature of and importance of coherence-building (Magliano et al., 2020). The results of the present study indicate that a comprehensive model that describes how individual differences support coherence-building is warranted but, to our knowledge, there is no formal model that describes how reading and coherence-building operate. The Reading Systems Frameworks (Perfetti & Stafura,

2014) is a step in this direction but was proposed to provide a heuristic for formulating research questions rather than a formal model. Gernsbacher's structure building model provides an exception as it was proposed to describe the basic processes of coherence-building and how some individual difference factors (e.g., knowledge suppression) support comprehension (Gernsbacher, 1997; Gernsbacher et al., 1990). The present study underscores the need for further development of a comprehensive model of reading and comprehension (e.g., McNamara & Magliano, 2009).

7.4. Implications for the computational assessment of constructed responses

There are two basic approaches to the computational analysis of constructed responses to assess coherence-building. One approach involves analyzing the semantic overlap between the constructed responses and the texts (Magliano & Millis, 2003; Magliano et al., 2011; Magliano & Graesser, 2012). This approach rests on the assumption that level of overlap with different aspects of the text is indicative of strategies associated with the prior discourse context. For example, constructed responses with greater overlap with the current sentence were more likely to reflect responses that paraphrased that sentence. In contrast, constructed responses with greater overlap with prior discourse sentences were more likely to reflect the use of bridging strategies to specify how the current sentence is related to the prior discourse context. There is evidence that this approach is related to both language-specific (Feller, Magliano, Sabatini, O'Reilly & Kopatich, 2020; Kopatich et al., 2019; Magliano & Millis, 2003; Magliano et al., 2011, 2020) and domain-general (Kopatich et al., 2019) individual difference factors.

The present study reflects a different approach to assess coherence-building. Namely, we were able to leverage automated, NLP techniques to detect cohesion in students' responses. Rather than comparing the constructed responses to the text, these analyses examine the relations across students' responses. This provided a means to model the connections that students were making while reading without having to rely on human raters' judgments of the accuracy of these connections. Cohesion analyses, thus, allow us to examine connections made at the level of the individual student.

Rather than competition across these approaches, we see the combination of these methods as a critical future direction for more meaningfully evaluating how readers engage in coherence-building and for understanding how these processes vary across readers and across contexts. Further, reading comprehension skills are better modeled when taking both the quality and features of readers' online constructed responses into account (Allen et al., 2015). The current study only examined cohesive features of readers' responses, but future extensions should focus on further analyzing the qualitative quality of the constructed responses.

One of the primary limitations of the current study is its sample size. The analyses discussed above may be further strengthened if future studies replicate these findings with a larger and more diverse sample. A larger sample size would also afford examining a broader array of linguistic features such as syntax, semantics, or lexicality, which may provide additional power to predict individual differences. A second limitation is the nature of the general reading comprehension measure. There is a current debate as to whether the Gates-MacGinitie Reading Test accurately assesses deep comprehension skills (Cutting & Scarborough, 2006; van den Broek & Espin, 2012). For this reason, additional measures of deep reading comprehension should be used to assess reading skill in future studies.

CONCLUSIONS

Despite these limitations, the current study has important implications for the realm of computational analysis and educational technology. Our results suggest that constructed response cohesion may aid researchers in the implementation of stealth assessments (e.g., Shute & Ventura, 2013). Stealth assessments refer to evaluations of student learning that occur during practice or game-play rather than through summative evaluations. Stealth assessments afford real-time in situ evaluation of students' reading processes. Therefore, future studies should examine how these NLP analyses can be deployed in real-time to provide students with feedback on their text comprehension strategies. Implementing these assessments may allow educational technologies to better model students' individual differences, which can lead to the development of more adaptive and personalized feedback that promotes the use of the right coherence-building strategies at the right time.

REFERENCES

- Allen, L. K., Snow, E. L. & McNamara, D. S. (2015). Are you reading my mind? Modeling students' reading comprehension skills with natural language processing techniques. *Proceedings of the Fifth International Conference on Learning Analytics and Knowledge*, 246-254.
- Allen, L. K., Jacovina, M. E. & McNamara, D. S. (2016). Cohesive features of deep text comprehension processes. In J. Trueswell, A. Papafragou, D. Grodner & D. Mirman (Eds.), *Proceedings of the 38th Annual Meeting of the Cognitive Science Society in Philadelphia* (pp. 2681-2686). Austin, TX: Cognitive Science Society.
- Allen, L. K., Snow, E. L., Crossley, S. A., Jackson, G. T. & McNamara, D. S. (2014). Reading comprehension components and their relation to writing. *L'Année Psychologique*, 114(4), 663-691.

- Barzilai, S. & Strømsø, H. (2018). *Individual differences in multiple document comprehension*. In J. L. G. Braasch, I. Bråten & M. T. McCrudden (Eds.), *Handbook of multiple source use* (99-116). New York: Routledge.
- Beck, I. L. & McKeown, M. G. (1988). Toward meaningful accounts in history texts for young learners. *Educational Researcher*, 17(6), 31-39. DOI: <https://doi.org/10.3102/0013189X017006031>
- Bohn-Gettler, C. M. & Kendeou, P. (2014). The interplay of reader goals, working memory, and text structure during reading. *Contemporary Educational Psychology*, 39(3), 206-219. DOI: <https://doi.org/10.1016/j.cedpsych.2014.05.003>
- Borella, E. & de Ribaupierre, A. (2014). The role of working memory, inhibition, and processing speed in text comprehension in children. *Learning and Individual Differences*, 34, 86-92. DOI: <https://doi.org/10.1016/j.lindif.2014.05.001>
- Bower, G. H. & Morrow, D. G. (1990). Mental models in narrative comprehension. *Science*, 247(4938), 44-48.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. DOI: <https://doi.org/10.1023/A:1010933404324>
- Clinton, V., Taylor, T., Bajpayee, S., Davison, M. L., Carlson, S. E. & Seipel, B. (2020). Inferential comprehension differences between narrative and expository texts: A systematic review and meta-analysis. *Reading and Writing*, 33(9), 2223-2248. DOI: <https://doi.org/10.1007/s11145-020-10044-2>
- Coiro, J. (2011). Predicting reading comprehension on the internet: Contributions of offline reading skills, online reading skills, and prior knowledge. *Journal of Literacy Research*, 43(4), 352-392. DOI: <https://doi.org/10.1177/1086296X11421979>
- Coté, N. & Goldman, S. R. (1999). Building representations of informational text: Evidence from children's think-aloud protocols. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 169-193). Lawrence Erlbaum.
- Cromley, J. G. & Azevedo, R. (2007). Testing and refining the direct and inferential mediation model of reading comprehension. *Journal of Educational Psychology*, 11(2), 311-325. DOI: <https://doi.org/10.1037/0022-0663.99.2.311>
- Crossley, S. A. (2018). How many words needed? Using natural language processing tools in educational data mining. *Proceedings of the 10th International Conference on Educational Data Mining*, 630-633.

- Crossley, S. & Kostyuk, V. (2017). Letting the genie out of the lamp: Using natural language processing tools to predict math performance. *International Conference on Language, Data, and Knowledge*, 10318, 330-342. DOI: https://doi.org/10.1007/978-3-319-59888-8_28
- Crossley, S. A., Kyle, K. & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 48(4), 1227-1237. DOI: <https://doi.org/10.3758/s13428-015-0651-7>
- Crossley, S., Paquette, L., Dascalu, M., McNamara, D. S. & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, 6-14. DOI: <https://doi.org/10.1145/2883851.2883931>
- Cutting, L. E. & Scarborough, H. S. (2006). Prediction of reading comprehension: Relative contributions of word recognition, language proficiency, and other cognitive skills can depend on how comprehension is measured. *Scientific Studies of Reading*, 10(3), 277-299. DOI: https://doi.org/10.1207/s1532799xssr1003_5
- Daneman, M. & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4), 450-466. DOI: [https://doi.org/10.1016/S0022-5371\(80\)90312-6](https://doi.org/10.1016/S0022-5371(80)90312-6)
- Ericsson, K. A. & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, 87(3), 215-251. DOI: <https://doi.org/10.1037/0033-295X.87.3.215>
- Feller, D. P., Magliano, J., Sabatini, J., O'Reilly, T., & Kopatich, R. D. (2020). Relations between component reading skills, inferences, and comprehension performance in community college readers. *Discourse Processes*, 57(5-6), 473-490. DOI: <https://doi.org/10.1080/0163853X.2020.1759175>
- Gernsbacher, M. A. (1997). Coherence cues mapping during comprehension. In J. Costermans & M. Fayol (Eds.), *Processing interclausal relationships: Studies in text production and comprehension of text* (pp. 3-21). Hillsdale, NJ: Lawrence Erlbaum.
- Gernsbacher, M. A., Varner, K. R. & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(3), 430-445. DOI: <https://doi.org/10.1037//0278-7393.16.3.430>
- Goldman, S. R. & Bisanz, G. L. (2002). Toward a functional analysis of scientific genres: Implications for understanding and learning processes. In J. Otero, J. A. León & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 19-50). Hillsdale, NJ: Lawrence Erlbaum.

- Graesser, A. C., Singer, M. & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, 101(3), 371-395. DOI: <https://doi.org/10.1037/0033-295X.101.3.371>
- Graesser, A. C. McNamara, D. S. & VanLehn, K. (2005). Scaffolding deep comprehension strategies through Point&Query, AutoTutor, and iSTART. *Educational Psychologist*, 40(4), 225-234. DOI: https://doi.org/10.1207/s15326985ep4004_4
- Hannon, B. & Daneman, M. (2001). A new tool for measuring and understanding the individual differences in the component processes of reading comprehension. *Journal of Educational Psychology*, 93, 103-128.
- Just, M. A. & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99(1), 122-149. DOI: <https://doi.org/10.1037/0033-295x.99.1.122>
- Kane, M. J. & McVay, J. C. (2012). What mind wandering reveals about executive-control abilities and failures. *Current Directions in Psychological Science*, 21(5), 348-354. APA PsycInfo. DOI: <https://doi.org/10.1177/0963721412454875>
- Karklin, Y. & Lewicki, M. S. (2005). A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Computation*, 17(2), 397-423. DOI: <https://doi.org/10.1162/0899766053011474>
- Kintsch, W. (1992). A cognitive architecture for comprehension. In H. L. Pick, Jr. P. W. van den Broek & D. C. Knill (Eds.), *Cognition: Conceptual and methodological issues* (pp. 143-163). Washington: American Psychological Association. DOI: <https://doi.org/10.1037/10564-006>
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.
- Klauda, S. L. & Guthrie, J. T. (2008). Relationships of three components of reading fluency to reading comprehension. *Journal of Educational Psychology*, 100(2), 310-321. DOI: <https://doi.org/10.1037/0022-0663.100.2.310>
- Kopatich, R. D., Magliano, J. P., Millis, K. K., Parker, C. P. & Ray, M. (2019). Understanding how language-specific and domain-general resources support comprehension. *Discourse Processes*, 56(7), 530-552. DOI: <https://doi.org/10.1080/0163853X.2018.1519358>
- Kuhn, M. (2019). Available Models. In *The caret Package* [on line]. Retrieved from: <https://topepo.github.io/caret/>

- MacGinitie, W. H., MacGinitie, R. K., Maria, K. & Dreyer, L. G. (2000). *Gates-MacGinitie Reading Test* (4th ed.). Itasca, Ill: *The Riverside Publishing Company*.
- Magliano, J. P. & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure. *Cognition and Instruction*, 21, 251-283.
- Magliano, J. P. & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavioral Research Methods*, 44, 608-621.
- Magliano, J. P., Durik, A. M. & Holt, J. K. (2011). *Assessing the impact of topic interest on comprehension processes* (LSAC Research Report Series) (pp. 1-78). Law School Admissions Council, Inc.
- Magliano, J. P., Higgs, K., Santuzzi, A., Tonks, S. M., O'Reilly, T., Sabatini, J., Feller, D., Kopatich, R. D., Ray, M. & Parker, C. (2020). Testing the inference mediation hypothesis in a post-secondary context. *Contemporary Educational Psychology*, 61, 101867. DOI: <https://doi.org/10.1016/j.cedpsych.2020.101867>
- Mar, R. A., Li, J., Nguyen, A. T. P. & Ta, C. P. (2021). Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review*, 28(3), 732-749. DOI: <https://doi.org/10.3758/s13423-020-01853-1>
- McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes*, 38(1), 1-30.
- McNamara, D. & Magliano, J. (2009). Chapter 9 Toward a Comprehensive Model of Comprehension. *Psychology of Learning and Motivation*, 51, 297-384. DOI: [https://doi.org/10.1016/S0079-7421\(09\)51009-2](https://doi.org/10.1016/S0079-7421(09)51009-2)
- McNamara, D. S., Graesser, A. C. & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In J. P. Sabatini & E. Albro (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89-116). Lanham, MD: R&L Education.
- McNamara, D. S., Graesser, A. C., McCarthy, P. & Cai, Z. (2014). *Automated evaluation of text and discourse with Cob-Metrix*. Cambridge: Cambridge University Press.
- McNamara, D. S., Allen, L. K., McCarthy, K. S. & Balyan, R. (2018). NLP: Getting Computers to Understand Discourse. In K. Millis, D. Long, J. Magliano & K. Wiemer (Eds.), *Deep learning: Multi-Disciplinary Approaches* (pp. 224-236). New York: Routledge.
- Moje, E. B. & Speyer, J. (2008). The reality of challenging texts in high school science and social studies: How teachers can mediate comprehension. In K. A. Hinchman & H. K. Sheridan-Thomas (Eds.), *Best practices in adolescent literacy instruction* (pp. 185-211). New York/ London: Guilford Press.

- Myers, J. L. & O'Brien, E. J. (1998). Accessing the discourse representation during reading. *Discourse Processes*, 26(2-3), 131-157. DOI: <https://doi.org/10.1080/01638539809545042>
- Oakhill, J. & Yuill, N. (1986). Pronoun resolution in skilled and less-skilled comprehenders: Effects of memory load and inferential complexity. *Language and Speech*, 29(1), 25-37. DOI: <https://doi.org/10.1177/002383098602900104>
- Öncel, P., Flynn, L. E., Sonia, A. N., Barker, K. E., Lindsay, G. C., McClure, C. M., McNamara, D. S. & Allen, L. K. (2021). Automatic student writing evaluation: Investigating the impact of individual differences on source-based writing. In M. Scheffel, N. Dowell, S. Joksimovic & G. Siemens (Eds.), *Proceedings of the 11th International Learning Analytics & Knowledge Conference*, (pp. 620-625). NY, USA: Association for Computing Machinery. DOI: <https://doi.org/10.1145/3448139.3448207>
- O'Reilly, T., Sinclair, G. P. & McNamara, D. S. (2004). ISTART: A web-based reading strategy intervention that improves students' science comprehension. In Kinshuk, D. G. Sampson & P. Isaias (Eds.), *Proceedings of the LADIS International Conference Cognition and Exploratory Learning in Digital Age: CELDA 2004* (pp. 173-180). Lisbon, Portugal: IADIS Press.
- Ozuru, Y., Dempsey, K. & McNamara, D. S. (2009). Prior knowledge, reading skill, and text cohesion in the comprehension of science texts. *Learning and Instruction*, 19(3), 228-242. DOI: <https://doi.org/10.1016/j.learninstruc.2008.04.003>
- Palincsar, A. S. & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, 1(2), 117-175. DOI: https://doi.org/10.1207/s1532690xci0102_1
- Parodi, G. (2010). Academic and professional discourse genres in Spanish. Amsterdam: John Benjamins. DOI: <https://benjamins.com/catalog/scl.40>
- Parodi, G. (2014). Genre organization in specialized discourse: Disciplinary variation across university textbooks. *Discourse Studies*, 16(1), 65-87. DOI: <https://doi.org/10.1177/1461445613496355>
- Parodi, G. (2015). Variation across university genres in seven disciplines: A corpus-based study on academic written Spanish. *International Journal of Corpus Linguistics*, 20(4), 469-499. DOI: <https://doi.org/10.1075/ijcl.20.4.03par>
- Perfetti, C. A. (1985). *Reading ability*. Oxford: Oxford University Press.

- Perfetti, C. & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18(1), 22-37. DOI: <https://doi.org/10.1080/10888438.2013.827687>
- Pressley, M. & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum.
- Pressley, M., Wood, E., Woloshyn, V. E., Martin, V., King, A. & Menke, D. (1992). Encouraging mindful use of prior knowledge: Attempting to construct explanatory answers facilitates learning. *Educational Psychologist*, 27(1), 91-109. DOI: https://doi.org/10.1207/s15326985ep2701_7
- Rong, S. & Bao-wen, Z. (2018). The research of regression model in machine learning field. *MATEC Web of Conferences*, 176, 01033. DOI: <https://doi.org/10.1051/matecconf/201817601033>
- Rosenshine, B., Meister, C. & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, 66(2), 181-221. DOI: <https://doi.org/10.2307/1170607>
- Rozimela, Y. (2014). The students' genre awareness and their reading comprehension of different text types. *International Journal of Asian Social Science*, 4(4), 460-469.
- Schmitz, A., Gräsel, C. & Rothstein, B. (2017). Students' genre expectations and the effects of text cohesion on reading comprehension. *Reading and Writing*, 30(5), 1115-1135. DOI: <https://doi.org/10.1007/s11145-016-9714-0>
- Shute, V. & Ventura, M. (2013). *Stealth Assessment: Measuring and Supporting Learning in Video Games*. Cambridge: The MIT Press. DOI: <https://doi.org/10.7551/mitpress/9589.001.0001>
- Stadtler, M., Bromme, R. & Rouet, J.-F. (2018). Learning from multiple documents: How can we foster multiple document literacy skills in a sustainable way? In *Promoting spontaneous use of learning and reasoning strategies: Theory, research, and practice for effective transfer* (pp. 46-61). New York: Routledge.
- Tong, S. & Koller, D. (2001). Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research*, 2, 45-66.
- Trabasso, T. & Magliano, J. P. (1996). Conscious understanding during comprehension. *Discourse Processes*, 21(3), 255-287.
- Turner, M. L. & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language*, 28(2), 127-154. DOI: [https://doi.org/10.1016/0749-596X\(89\)90040-5](https://doi.org/10.1016/0749-596X(89)90040-5)

- Unsworth, N., Heitz, R. P., Schrock, J. C. & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, 37, 498-505.
- van den Broek, P. (1994). Comprehension and memory of narrative texts: Inferences and coherence. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 539-588). San Diego, California: Academic Press.
- van den Broek, P. & Espin, C. A. (2012). Connecting cognitive theory and assessment: Measuring individual differences in reading comprehension. *School Psychology Review*, 41(3), 315-325. DOI: <https://doi.org/10.1080/02796015.2012.12087512>
- Whitney, P., Ritchie, B. G. & Clark, M. B. (1991). Working-memory capacity and the use of elaborative inferences in text comprehension. *Discourse Processes*, 14(2), 133-145. DOI: <https://doi.org/10.1080/01638539109544779>
- Wolfe, M. B. W. & Mienko, J. A. (2007). Learning and memory of factual content from narrative and expository text. *The British Journal of Educational Psychology*, 77(3), 541-564. DOI: <https://doi.org/10.1348/000709906X143902>
- Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(4), 920-933. DOI: <https://doi.org/10.1037/0278-7393.20.4.920>

APPENDIX

Example Text (Labor Strikes of Garment Workers):

During the late 1800s and early 1900s, thousands of people from all over the world came to America in hopes of a better life, a brighter future, a way to better their chances of living comfortably. In Russia, the Singer sewing machine had been introduced 2 decades earlier and symbolized the guarantee of a good livelihood. Among those who came were 50,000 young Jewish girls from Russia. Most of the young women who came over were highly skilled. However, the reality of the working class in America was not conducive to their plans.

Constituting 1/3 of the garment industry's workforce in 1910, these women labored for long hours in intensely cramped and unsafe working conditions. To complicate matters, wages paid in these jobs were not sufficient to survive. For many of these women, their dream of America was shattered.

By 1910, small spontaneous strikes began erupting at several of the garment factories. Although the women asked certain unions to help unify the strikers, these

unions lacked the resources to mobilize such a massive action. The power to do so had to come from the people themselves.

On the night of November 22, thousands of workers attended a mass meeting organized by the women strikers. In the packed hall, their bodies were restless and taut with anger, they heard speaker after speaker advise them to be patient and act cautiously. Clara Lemlich, a fragile looking teenager, rushed suddenly to the platform and passionately articulated the pent-up feelings of the audience.

The next morning, 15,000 workers were on strike. The East Side was a seething mass of agitated women, girls, and men. The strikers swelled in number, to over 20,000; they were overwhelmingly Jewish. The demands of the strikers included a 52 hour week, overtime pay, and union recognition. As they picketed, the strikers were arrested by the police and beaten by thugs. The courageous strikers impressed the community; they were proudly described as ‘our wonderful fervent girls.’

The strike was powerful, intimidating, and by February, more than 300 of some 450 firms in the New York industry had been forced to make some kind of settlement. These labor struggles represented a watershed in Jewish-American history. The ‘uprisings’ of this era sharpened a shared sense of ethnicity, a Jewish immigrant identity in America.

ACKNOWLEDGEMENTS

The authors declare that there are no conflicts of interest. This research was supported in part by IES Grant R305A190063 as well as the Office of Naval Research (Grants: N00014-20-1-2627, N00014-19-1-2424, and N00014-20-2623). Opinions, conclusions, or recommendations do not necessarily reflect the view of the Department of Education, IES, or the Office of Naval Research.

Correspondence should be addressed to Lauren Flynn at lauren.flynn@unh.edu or Laura Allen at laura.allen@unh.edu.