Linguistic Features of Lectures and their Relationship with Student Performance

by

Torin Young

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirements for the degree of

Master of Arts

in

Psychology

Waterloo, Ontario, Canada, 2021

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Lectures are an important part of the post-secondary experience. Optimizing various aspects of this experience for the benefit of students' learning has been examined (Mayer, 2019). However, the linguistic features of lectures and how these features might affect student learning have been overlooked in the extant literature. Recent studies have utilised Coh-Metrix, an automated text analyzer, to examine discourse in both texts and lecture discourse (Graesser, McNamara, & Kulikowich 2011; McNamara, Graesser, McCarthy, & Cai, 2014; Medimorec, Palvik Jr, Oleny, Gaesser, & Risko, 2015; Morgan, Burkett, Bagley, Graesser, 2011). We extend this effort here by analyzing linguistic features of lectures and how they are associated with student performance. In particular, we were interested in determining whether (a) computationally generated measures of language are associated with student performance and (b) whether different associations are observed with different testing methods (multiple-choice vs. short answer). We demonstrate that a lecturer's narrativity, syntactic simplicity, and referential cohesion are associated with performance on multiple-choice tests. Preliminary results suggest a different pattern of association for short answer tests.

Acknowledgements

I would like to thank my supervisors, Dr. Katherine White and Dr. Derek Besner, for their guidance and support, not only through the research presented herein but throughout my graduate study thus far. I would like to thank Dr. Evan F. Risko for his helpful feedback and comments.

I would like to thank my colleagues in the Lab for Infant Development and Language (LIDL) for their comments and feedback for the various procedures presented herein. Special thanks go to Ashley Avarino for help coding the Experiment 2 data and Dr. Srdan Medimorec at Teesside University for his continue mentorship and guidance with the data analyses.

Finally, I would like to thank my mother for her everlasting patience and support during this process as well as my uncle for sparking the fire to pursue this degree and his continued support in my future endeavours.

To my love, Braelyne Crete, thank you for being the everlasting flame to light my way through any dark place I may be or have been and the omnipresent voice of reason whenever I may have doubt.

Table of Contents

List of Figures

List of Tables

Introduction

Lectures are a staple of academic life. The goal of lectures is to produce engagement with the course material and to facilitate learning within a structured environment. Ideally, a lecturer should strive to deliver the material in a way that reduces the amount of unnecessary cognitive load placed on the students and, at the same time, constructs a knowledge base that then can be actively worked on by the student (Mayer, 2019, 2011). Given the multimodal nature of lectures, which can include visual graphics, written text, and narration, there are multiple features that a lecturer could manipulate to achieve these goals. In the present work, we focus on an often-overlooked aspect of lectures: the variation in their language characteristics.

Although there is little work to date that provides a direct analysis of lecturer language (or its relationship with student learning), there is extensive research examining the linguistic features of another instructional medium: academic textbooks. This work has characterized the dimensions along which textbook language tends to vary, identifying four main dimensions of variation – the concreteness of the language, syntactic simplicity, narrativity, and the cohesiveness of ideas presented throughout the text (Graesser, McNamara, Louwerse, & Cai, 2004; Graesser et al., 2011; McNamara, Graesser, McCarthy, & Cai, 2014). I describe these dimensions below, as well as how they might relate to learning.

Linguistic Features of Text and Comprehension

The concreteness of a text is jointly determined by how concrete, meaningful and imageable its content words are. A text that is less concrete is characterized by more abstract concepts and ideas. Syntactic simplicity is determined by a number of factors, including the number of words per sentence and the number of words before the main verb of the main clause.

The third feature, narrativity, is associated with how informational or informal the text is. Higher narrativity involves a more informal style of writing, such as that often found in works of fiction, using words and events that are more familiar to the reader. In contrast, a more informational style of writing, such as that often found in science texts, uses domain-specific jargon, less familiar words, and a more formal style. The final dimension, cohesiveness, is associated with how connected concepts and ideas are throughout the text. More cohesive text includes the frequent use of causal (e.g., because, so), temporal (e.g., after, before), logical (e.g., and then, if-then), additive (e.g., moreover, however), and/or adversative (e.g., but, although) connectives, as well as repetition of particular nouns throughout the text.

These linguistic features of texts affect readers' processing, comprehension, and memory. For example, concrete words are more easily recognized, read, and recalled than more abstract words (e.g., Sadoski, Goetz, & Fritz, 1993; Sadoski & Paivio, 2013; Graesser et al., 2011). Shorter sentences with simpler structure place less load on working memory, in turn making recall of the information easier (Caplan & Waters, 1999). Information that is presented in a more story-like way is read faster, easier to comprehend, and recalled better than text that is purely informational (Graesser, Hauft-Smith, Cohen, & Pyles, 1980; Haberlandt & Graesser, 1985). And comprehension of material is easier when ideas are linked (e.g., Britton & Gulgoz, 1991; Zwaan & Radvansky, 1998; Kintsch, Kozminsky, Streby, McKoon, & Keenan, 1975; Vidal-Abarca, Martínez, & Gilabert, 2000).

The work on the linguistic features of academic texts has further demonstrated that textbook authors might modify certain dimensions of their language to balance others, presumably in an attempt to maximize engagement and comprehension (e.g., Graesser et al., 2014; McNamara, 2013). McNamara (2013) analyzed textbooks with varying grade levels and

genre (e.g., science, narrative) to examine whether there was a compensatory relationship between the different linguistic features related to text difficulty. Their analysis suggests that textbook authors compensate for more abstract concepts by having simpler syntax and more cohesion, and that the extent to which an author understands the cognitive abilities of their audience affects the degree of compensation (McNamara, 2013). This suggests that authors may engage in this type of compensatory behavior to aid students' learning. Indeed, a recent study suggests that when texts are more complex by virtue of having less concrete topics, students' performance on an explanation task is moderated by the degree to which the text provides a cohesive structure (Jacob, Lachner, & Sheiter, 2020). In other words, greater cohesion appears to compensate for less accessible content.

The Language of the Lecturer and Author

To date, there is almost no research on these sorts of linguistic properties in lectures. The goals of the lecturer and textbook author are similar, in that both strive to deliver material in a way that facilitates learning, by taking their audience's knowledge and background into account. However, there are some key differences between the two instructional mediums that are important to note. First, authors and lecturers differ in how much time they devote to preparing the material, and this may have a number of consequences for language properties. For example, Glass and colleagues (2004) demonstrate that the language used in lectures is less organized and less concise than the corresponding textbook material, and suggest that this is because lecture material is produced live in the lecture hall, rather than pre-planned. Lecturers have less time to plan and edit, which may also affect lexical and syntactic choices. In fact, even within a single modality, increased processing time can lead to more sophisticated language (e.g., as indexed by lexical frequency; see Medimorec, Young, & Risko, 2016 for examples of how processing time

affects lexical choice in writing). Further, spoken language is generally seen as being more informal relative to text (Graesser et al., 2014; Li et al., 2013). Indeed, in a study exploring the linguistic features of both typed and spoken conversations in an academic-oriented game, spoken language was less formal (higher in narrativity and cohesiveness) than written language (Morgan et al., 2011). Lastly, lecturers are afforded access to paralinguistic cues (e.g., looks of confusion, confirmatory head nods) from their audience, which could alter the language they decide to use in the moment.

Despite these differences, there is one study documenting how lecturers might manipulate their language in ways similar to textbook writers. Medimorec and colleagues (2015) examined the linguistic features of open-access pre-recorded lectures. They found that when concepts within a lecture are more abstract and involve lower frequency words, the lecturers' discourse compensated for this by having more informal language, shorter and simpler sentences, and used more connectives to link concepts throughout. These findings suggest that the language properties that are informative for characterizing textbooks are also informative for characterizing the language of lectures.

Language Use and Learning

To what extent do these dimensions actually affect student learning in a lecture setting? To our knowledge, there is little work examining this question. The existing literature examining how lecturer discourse affects student learning seems to be focused on either students' subjective reports of interest and affect (e.g., Tin, 2009; Weninger, Staudt, & Schuller, 2013; Shadiev & Huang, 2020) or how the lecturer organizes and contextualizes the material using pragmatic features such as contextualization markers (e.g., *"to sum up so far", "let me repeat myself"*) and organization markers (e.g., *"my point here is that", "I am going to briefly talk about"*) (e.g., Zare

& Keivanloo-Shahrestanaki, 2017; Zare, 2019). However, given that many of the processing and recall effects described above for various properties of text also extend to the spoken language modality (Favier, Meyer, & Huettig, 2021; Garlock, Walley, & Metsala, 2001; Griffin & Bock, 1998; Walker & Hulme, 1999), it is conceivable that the linguistic characteristics of lectures affect student learning.

Student learning can be assessed using different testing methods, including multiple-choice and open-ended short answer questions. These different testing methods may tap into different degrees of learning. For example, it is argued that multiple choice questions place less demand on students as compared to open-ended style questions (Graesser et al., 2010; Kang, McDermott, & Roediger, 2007; Ozuru, Briner, Kurby, & McNamara, 2013). In particular, multiple choice questions rely on familiarity and automatic retrieval processes, whereas open-ended questions rely on more controlled retrieval processes that include multiple steps of goal-oriented, active searching through previously stored information (Jacoby, 1996). Thus, it is possible that different language features could affect different processes associated with these different testing methods. For example, the cohesiveness of a lecturer's language might predict better performance on open ended short answer questions because more global connections and links across ideas may provide stronger retrieval cues for that information. It may also be the case that some linguistic features would not be predictive in a free recall context. For example, narrative texts have been shown to predict worse performance on a free recall test, and it has been argued that the genre of the text affects the learner's goals when processing the text (Wolfe & Mienko, 2007; McDaniel & Einstein, 1989, 2005). Wolf and Mienko (2007) suggest that when given a text that is deemed to have a narrative genre, the learner focuses more on the structure of the narrative and commits the information into 'narrative memory' rather than integrating the

information with their prior knowledge, whereas a learner will integrate the information of an expository (informational) text with their prior knowledge. This possible dissociation leads us to examine different types of question types (multiple choice vs. open ended short answer) to see if different linguistic profiles might predict different performance outcomes.

Present Investigation

Taken together, previous research has 1) documented features of variation that can be used to characterize language use in an academic setting and 2) demonstrated that these features apply meaningfully to lectures as well as textbooks (despite the differences in modality). Moreover, in some contexts, these dimensions have effects on processing, comprehension, and recall. However, to our knowledge, there is no work directly examining whether these language properties influence students' comprehension and recall of lecture material. The present investigation aims to examine the relationship between the properties of a lecturer's language and participants' scores on a later quiz.

Experiment 1

In Experiment 1, we examine the linguistic features of a series of lectures and their relation to participants' scores on a multiple-choice quiz. This was done by examining the linguistic features of 18 lecture transcripts and their relationship to the participants' quiz scores for each lecture. We predicted that students' scores on the quiz would be predicted by specific features of the lecturer's language, such that language that is more easily processed and/or promotes encoding and retrieval processes would be associated with higher multiple-choice quiz scores.

Method

Participants

A total of 200 University of Waterloo students (Age: $M = 19.64$, $SD = 3.45$) participated in the experiment in exchange for course credit. Participants were recruited through SONA and completed the study online. All participants were either native English speakers or had learned English by the age of eight years old or younger. Given the overlapping subject matter, participants had to have not taken PSYCH 207: Cognitive Processes in the past or be enrolled in the course presently. The data from 11 participants were discarded due to failure to complete the task, leaving 189 participants for analysis.

Lectures

The video lectures presented were pre-recorded for the online version of *PSYCH 207: Cognitive Processes*, a course offered at the University of Waterloo. Permission to use these lectures was obtained from the lecturer prior to the study. The course was divided into 12 different learning modules (corresponding to the 12 weeks of the course). Each module was comprised of a varying number of short lectures (e.g., the material in module 1 was distributed across lectures 1, 2, 3, 4, and 5). From the full set, 18 lectures from 6 different learning modules were chosen, ranging in length from 5 minutes to 19 minutes. Lectures were grouped into sets of 3, such that each set contained lectures covering different topics (e.g., Set 1 had a lecture on the history of the cognitive revolution, an introduction to localisation of function in the brain, and a summary of utility models of decision making; see Appendix A for a full list of the lectures and descriptions). Due to the variation in the lengths of the individual lectures, each set of 3 lectures

was created to be similar in length (between 28 - 31 minutes). The lectures were uploaded as unlisted videos to a private YouTube channel for presentation.

Tests

Each lecture was followed by a six-question multiple-choice quiz. Each question contained 1 correct answer and 2 foils.

Working Memory tasks

Two working memory tasks (Oswald, McAbee, Redick, & Hembrick, 2015) were used as distractors between the lecture and multiple-choice quizzes. While data for these tasks were collected, they are not analyzed for this thesis. The first task was an *operation span* (O-span) task, in which the participants were presented with a mathematical equation and an answer that was either true or false (e.g., (5x2) - 6 = 4). The participant had to determine whether the answer was true or not and respond by clicking on the radial buttons labelled "TRUE" or "FALSE". After their response, a letter appeared on the screen for 750ms that the participants were told to remember for later recall. After several trials, participants were asked to recall the letters in the order they appeared. They were presented with a list of letters and were instructed to number the order of presentation by typing a number in the text box next to each letter. If participants could not recall the order of certain letters in the set, they were instructed to type the number in boxes labelled 'BLANK' as place holders for the forgotten letters.

The second task was a *reading span* (R-span) task, in which participants read statements that either made sense (e.g., "Dan walked around the streets posting signs and looking for his lost puppy.") or were nonsensical (e.g., "When it is cold, my mother always makes me wear a cape on my head."). Participants had to determine whether the statement made sense or not by clicking

on the radial buttons labelled "TRUE" or "FALSE". Following each statement, a letter was presented. The remainder of the procedure was identical to the O-span task.

The number of letters presented to participants within a trial for later recall ranged from 4 to 6 (Oswald et al., 2015). Participants completed 2 set sizes per task following each lecture (e.g., after the first lecture, a participant might complete the O-span task with set sizes of 4 and 6 and the R-span task with set sizes of 6 and 5; after viewing the second lecture, a set size of 5 and 4 for the O-span task and a set size of 4 and 6 for the R-span; and, after viewing the third lecture, a set size of 5 and 6 for the O-span and a set size of 4 and 5 for the R-span task).

Text Analyzer Description

The transcripts of the lectures were analyzed using the Coh-Metrix text analyzer (Graesser et al., 2004, 2011; McNamara et al., 2014; Medimorec et al., 2015). All reported variables were computationally extracted from the transcripts and no human coding was used. Coh-Metrix uses multiple text analyzing tools that are readily available (e.g., lexical databases, syntactic parsers, part-of-speech taggers, lexical coreference tools; for more details see McNamara et al., 2014, Chapter 3) and integrates them into one convenient location. There is a free online version of Coh-Metrix available at http:///tool.cohmetrix.com.

Coh-Metrix provides measures of five features (Word Concreteness, Narrativity, Deep Cohesion, Referential Cohesion, and Syntactic Simplicity) that are composed of 53 language-discourse measures. These five features accounted for over 54% of text variability among 37,520 paragraph-length texts (Graesser et al., 2011) and have been used extensively in past research on texts (Graesser et al., 2011,2014; McNamara, 2013; McNamara et al., 2012) and spoken discourse (Medimorec et al., 2015; Morgan et al., 2011).

In the present study, Coh-Metrix was used to determine the word count and the following five features for each lecture: Word Concreteness, Narrativity, Deep Cohesion, Referential Cohesion, and Syntactic Simplicity. Coh-Metrix compares the inputted text to the Touchstone Applied Sciences Associates (TASA) corpus (Landauer, Foltz, & Laham, 1998) and computes *z* scores for each of these five features (for values for the current lecture set, see Appendix B). The TASA corpus is comprised of texts from various genres and disciplines between kindergarten and college-ready, and it contains over 11 million words. The computed z scores provide us with a measure of 'ease' for each feature, where a higher score would mean that the feature is 'easier'. For example, a higher Narrativity score indicates that the transcript has a more story-like narrative and contains words that are more familiar. Descriptions of the five features and examples of the measures they are comprised of follow.

Word Concreteness reflects how concrete or abstract the concepts presented in the text are. Word Concreteness is jointly determined by the semantic measures of word concreteness, word meaningfulness (a function of a word's familiarity and its paired associations; See Noble, 1963), and word imageability determined by the MRC Psycholinguistic Database (Coltheart, 1981).

Narrativity refers to how much a text is classified as informational or story-like. Narrativity is determined by measures of how frequently the text uses pronouns, the use of intentional events and actions (actions and events that are performed by an animate agent that is motivated by plans in pursuit of goals; Zwaan & Radvansky, 1998) or causal events and actions (events or actions that take place in the physical or psychological world, such as an earthquake or discovering a solution, that may or may not be driven by goals; Graesser et al., 1994, 2004;

Zwaan & Radvansky, 1998), and the familiarity of the words presented (e.g., All-word frequency, Content-word frequency, familiarity).

Deep Cohesion refers to the extent to which ideas and concepts are explicitly linked through the use of causal (e.g., because, so), temporal (e.g., after, before), logical (e.g., and then, if-then), additive (e.g., moreover, however), and/or adversative (e.g., but, although) connectives throughout the text. *Referential cohesion* is determined by measures of content word overlap: the proportion of content words that are the same between pairs of sentences (*content word overlap*), proportion of sentence pairs that share one or more common nouns (*noun overlap*), proportion of sentence pairs that share common nouns or pronouns and their morphological variants (e.g., *table/tables; argument overlap*), and proportion of sentence pairs in which a noun in one sentence has a semantic unit in common with a word in any other grammatical category in adjacent sentences (e.g., *photograph/photographed*; *stem overlap)* throughout the text (Graesser et al., 2011, *pp*. 226). Such repetition establishes links between ideas and concepts across sentences.

Syntactic Simplicity refers to how complex or simple the syntactic structure of a sentence is (e.g., the number of words per sentence, number of words before the main verb of the main clause).

Procedure

Participants were assigned to one of the six lecture sets. The study was administered through Qualtrics. After providing consent, participants completed 2 blocks of practice trials for each of the two working memory tasks. These practice blocks were constructed so that participants first practiced the components of the working memory tasks (responding to the

operation or sentence recalling a set of letters presented) for 6 trials and then the full task with the components together.

Participants were then instructed to watch a lecture in its entirety and to focus on the lecture itself, without taking notes. YouTube links to the lectures were embedded in the Qualtrics survey. The video player had all functions (e.g., speed, closed captions) disabled except the ability to pause and play the video. Following each lecture, participants completed two sets of the O-Span task followed by two sets of the R-Span task. Participants were then instructed they would have 6 minutes to complete a six-question multiple-choice quiz. This timing was imposed to limit participants' ability to search for the answers in another tab in their browser. Answers to each quiz were pseudo-randomly ordered (and consistent across participants receiving that quiz), with the constraint that it was never the case that a single option (e.g., a) was correct for all questions in a quiz. After completing the first multiple choice quiz, participants repeated the sequence (lecture – working memory tasks – quiz) for the remaining two lectures.

Results

To assess whether test scores were affected by differences in language characteristics, we fitted mixed-effects models using R version 4.0.5 (R Core Team, 2019) and the package lme4 (version 1.1-26; Bates, Maechler, Bolker, & Walker, 2015). We complement the frequentist analyses with the corresponding Bayesian analyses to additionally confirm the strength of effects. Bayesian analyses were performed using the brms package (Bürkner, 2017) and are presented in the text when critical.

The maximum model structure included the five linguistic features related to discourse difficulty (Word Concreteness, Narrativity, Deep Cohesion, Referential Cohesion, and Syntactic Simplicity) and word count as the fixed effects. As random effects, we entered intercepts for

both participants and topics. The dependent variable was test score (z-scored). Each model was tested against a null model containing only the intercept and the maximal model structure. The models were compared using the anova function and evaluated using the AIC criterion. The reported estimates are standardized coefficients (β).

Descriptive statistics are presented in Table 1, and correlations among the variables are presented in Table 2 and Figure 1. It is important to note that in the creation of these five linguistic features, they were constrained as orthogonal principal components (Graesser et al., 2011). This is not the case here or typically considered when analyzing 'new' texts from outside the TASA Corpus. Indeed, Medimorec and colleagues (2015) observed correlations among these linguistic features in their lecture corpus as well. These similarities and differences between the correlations reported here and Medimorec and colleagues are discussed later.

Table 1

Descriptive Statistics, All Predictors

	Mean	SD	Minimum	Maximum
Word Concreteness	-1.08	.37	-1.63	-.49
Narrativity	-.20	.42	-.88	.32
Deep Cohesion	.87	.99	-.97	2.96
Referential Cohesion	.20	.53	-1.09	.98
Syntactic Simplicity	-.09	.22	-.37	.38
Word Count	1673.33	640.57	846	3491

Table 2

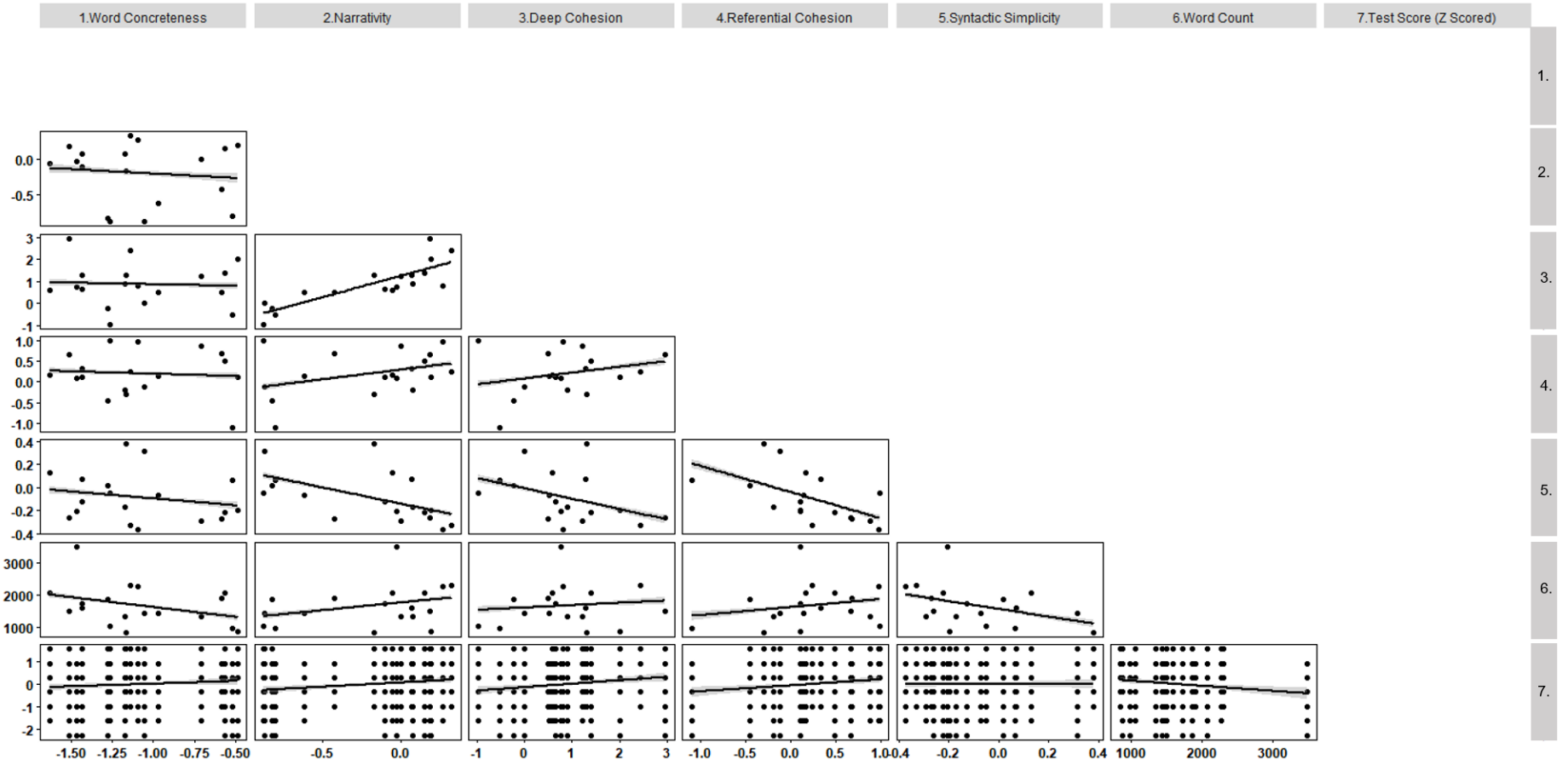
Experiment 1: Correlations Among All Variables

	1	2	3	4	5	6	7
1. Word Concreteness	—						
2. Narrativity	-.10 *	—					
3. Deep Cohesion	-.05	.84 ***	—				
4. Referential Cohesion	-.08	.37 ***	.26 ***	—			
5. Syntactic Simplicity	-.20 ***	-.56 ***	-.42 ***	-.56 ***	—		
6. Word Count	-.34 ***	.32 ***	.11 **	.20 ***	-.40 ***	—	
7. Test Z-Score	.09 *	.14 ***	.14 ***	.13 **	.01	-.15 ***	—

Note. * $p < .05$, ** $p < .01$, *** $p < .001$

Figure 1.

Experiment 1: Correlation Plots of All Variables



The best fitting model included Narrativity, Referential Cohesion, Syntactic Simplicity and word count, but not Deep Cohesion. The summary output for the model indicated that test scores increased with increased Narrativity, $\beta = .91$, $SE = .34$, $t = 2.67$, $p = .019$. Similarly, increased test scores were (marginally) related to increased Syntactic Simplicity $\beta = .61$, $SE = .32$, $t = 1.94$, $p = .055$. In the corresponding Bayesian analysis, the Syntactic Simplicity effect was statistically significant as indicated by the 95% credible interval (CI) not containing zero, $\beta = .66$, 95% CI [.033, 1.33]. Increased test scores were also related to increased Referential Cohesion, $\beta = .25$, $SE = .12$, $t = 2.12$, $p = .035$. On the other hand, test scores decreased with increased word count, $\beta = -.0004$, $SE = .0001$, $t = -6.16$, $p < .001$.

Additional Analysis: Test Questions

In our final analysis we investigated whether linguistic features of test questions were related to test scores. Given our question of interest (whether lecture features are related to test scores), it is important to rule out the possibility that our results reflect language differences during test alone. For example, it is possible that more linguistically challenging lectures also had more linguistically challenging questions. Therefore, to ensure that the effects reflect differences in processing or learning during encoding of the material, we additionally asked whether the linguistic features of the test questions themselves related to test outcomes.

We focused on two linguistic features of test questions: Syntactic Simplicity and Word Concreteness (other features were not assessed since questions were relatively short and thus not conducive to text-level feature analyses). We entered question Syntactic Simplicity and Word Concreteness as the fixed factors and intercepts for participants and topics as random effects. The DV was test score. The summary output for the model indicated that test scores were not

related to the predictors: Word Concreteness $\beta = .003$, $SE = .05$, $t = .06$, $p = .956$; Syntactic Simplicity $\beta = .004$, $SE = .07$, $t = .06$, $p = .950$. Therefore, test performance was related to linguistic features that were present at the time of encoding, and not retrieval.

Discussion

The results of Experiment 1 revealed that higher narrativity, syntactic simplicity, and referential cohesion in lectures were related to better performance on a multiple-choice test. Increased word count was related to a decrease in performance. In other words, when the lecturer's language was more informal, used shorter and simpler sentences, and had greater overlap in ideas and concepts across sentences, students' performance at test increased. In contrast, two other linguistic features of the lectures, deep cohesion and word concreteness, were unrelated to students' performance. Therefore, the use of more concrete words and increased use of connectives (e.g., but, so, although) did not provide a benefit, despite past research showing that these two properties are associated with encoding, retrieval, and recall.

Experiment 2

In Experiment 2, we examine whether the same linguistic properties are related to performance on a different type of test question. This was done by altering our procedure so that participants responded to short answer questions instead of multiple-choice questions. One possibility is that we will see that Narrativity, Syntactic Simplicity, and Referential Cohesion are once again associated with student performance at test. However, another possibility is that we will see a different set of associations because of the different demands/processes involved in answering short answer questions.

Method

Participants

A total of 41 University of Waterloo students (Age: $M = 22.51$, $SD = 5.59$) participated in the experiment in exchange for course credit (data collection is still ongoing). Participants were recruited through SONA and completed the study online. All participants were either native English speakers or had learned English by the age of eight years old or younger. Given the overlapping subject matter, participants had to have not taken PSYCH 207: Cognitive Processes in the past or be enrolled in the course presently.

Lectures

The lectures and lecture groups were identical to those used in Experiment 1.

Tests

Each lecture was followed by 2 short answer questions. The first question was the same for all lectures and asked "*List and describe 3 important points made in the lecture you just watched*". The second question was focused on the content of each lecture. For example, after the lecture on brain imaging techniques, the second question was "*Compare and contrast MRI and CAT scans as neuroimaging techniques in terms of their methods and the information they provide*".

The participants' responses were graded by two separate coders. Grading was based on marking guides created by the researchers. These marking guides contained information from each lecture as well as a grade breakdown for each correct answer. Each question was worth 3 marks. Coders were instructed to code based only on the marking guides provided. To ensure

consistency, both coders first (independently) coded the data from 12 participants. A Pearson's correlation across the coders' scores for these 12 participants was $r = .85$, $p < .001$ and any variation in coding for a single question was within half a point (0.5). Following this reliability check, the remaining participants were divided between the two coders for marking.

Procedure

The procedure was exactly the same as in Experiment 1, with the exception that the multiple-choice questions were replaced with 2 short answer questions. The questions were presented individually, and participants had 5 minutes to respond to each question. This time limit was imposed to prevent participants from using external sources (e.g., Google) to look up the answer.

Preliminary Results

The maximum model structure included the five linguistic features related to discourse difficulty (Word Concreteness, Narrativity, Deep Cohesion, Referential Cohesion, and Syntactic Simplicity) and word count as the fixed effects. As random effects, we entered intercepts for both participants and topics. The dependent variable was the total test score. As before, the models were compared using the anova function and AIC criterion. The reported estimates are standardized coefficients (β). Descriptive statistics are the same as Experiment 1 (see Table 1) and correlations among the variables are presented in Table 3 and Figure 2.

Table 3.

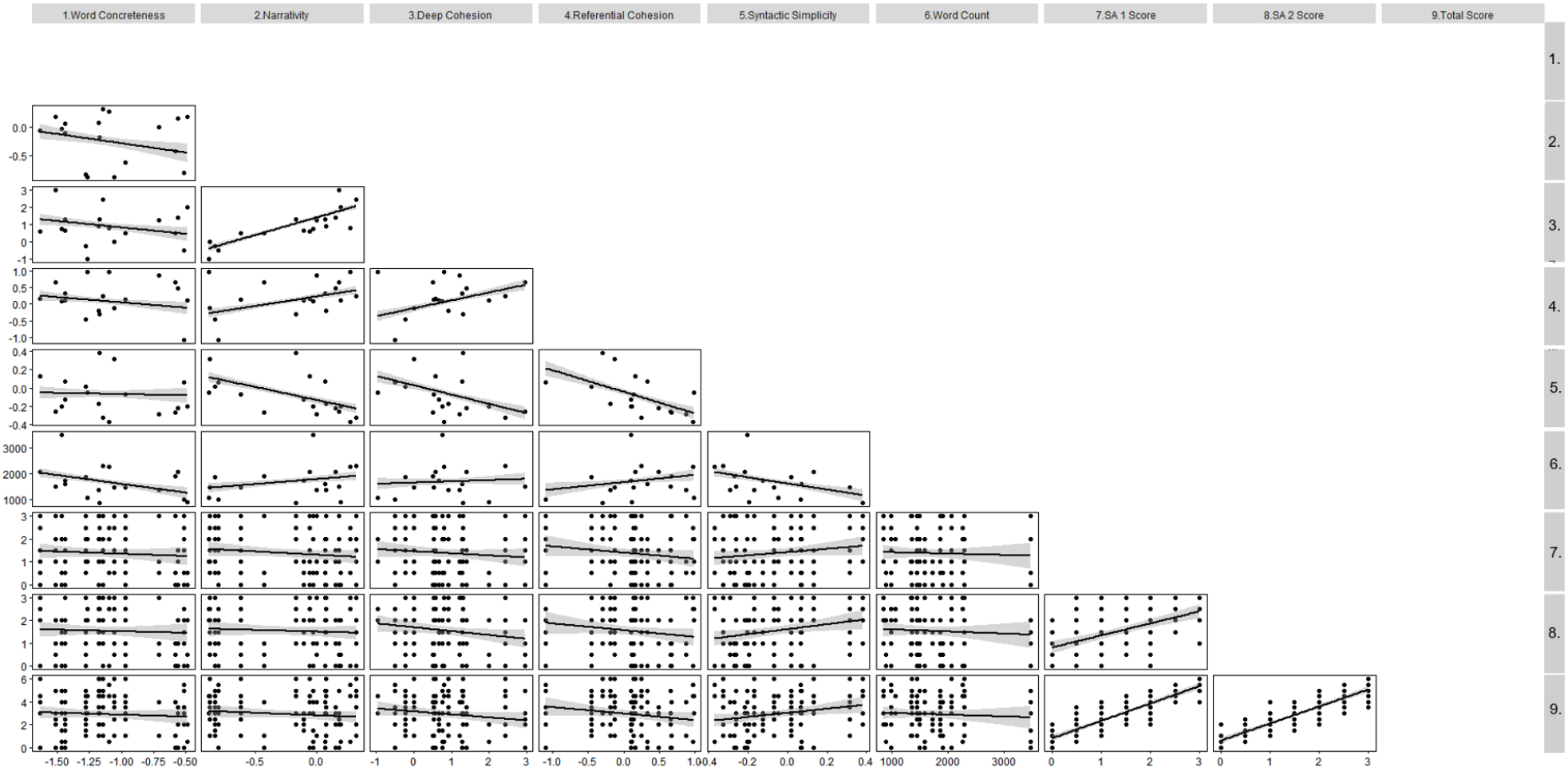
Experiment 2: Correlations Among All Variables

	1	2	3	4	5	6	7	8
1. Word Concreteness	—							
2. Narrativity	-.25 **	—						
3. Deep Cohesion	-.24 **	.84 ***	—					
4. Referential Cohesion	-.22 *	.52 ***	.51 ***	—				
5. Syntactic Simplicity	-.04	-.59 ***	-.50 ***	-.55 ***	—			
6. Word Count	-.37 ***	.27 **	.08	.22 *	-.41 ***	—		
7. SA1 Score	-.07	-.13	-.10	-.13	.15	-.04	—	
8. SA2 Score	-.04	-.06	-.17	-.14	.22 *	-.05	.51 ***	—
9. Total Score	-.06	-.11	-.16	-.16	.22 *	-.05	.87 ***	.87 ***

*Note. * p < .05, ** p < .01, *** p < .001; SA1 Score refers to the participants' scores on the first short answer question; SA2 Score refers to the participants' scores on the second short answer question.*

Figure 2.

Experiment 2: Correlation Plots of All Variables



The best fit model included Word Concreteness and Referential Cohesion. Test scores increased with decreased Word Concreteness, β = -1.23, *SE* = .50, *t* = -2.49, *p* = .014, and with decreased Referential Cohesion, β = -1.44, *SE* = .39, *t* = -3.71, *p* = <.001.

Short Answer Variation?

The two short answer questions presented to participants can be seen as two types of open-ended questions. The first could be considered, in memory research terms, a free recall question whereas the second question could be considered a cued recall. To examine the possible differences in how the linguistic properties affected performance on these questions, we ran two separate analyses identical to the one above where the dependent variable was the test score for the first question for one analysis and the dependent variable was the test score for the second question for the other analysis.

When the first short answer question was the DV, the best fit model included Syntactic Simplicity, Referential Cohesion, and word count. Specifically, test scores increased with increased Syntactic Simplicity, β = .84, *SE* = .40, *t* = 2.08, *p* = .040, and with increased word count, β = .0003, *SE* = .0001, *t* = 2.18, *p* = .032. On the other hand, test scores increased with decreased Referential Cohesion (marginally), β = -.41, *SE* = .21, *t* = -1.98, *p* = .051.

When the second short answer question was the DV, the best fit model included only Deep Cohesion. Specifically, test scores increased with decreased Deep Cohesion, β = -.33, *SE* = .11, *t* = -2.89, *p* = .007.

Additional Analysis: Test Questions

Again, we investigated whether linguistic features of test questions were related to test scores.

However, since question one was the same for all lecture sets, we only focused on the second test question. We again considered two linguistic features of the questions: Syntactic Simplicity and Word Concreteness (other features were not assessed since questions were relatively short and thus not conducive to text-level feature analyses). We entered question Syntactic Simplicity and Word Concreteness as the fixed factors and intercepts for participants and topics as random effects. The DV was test score for the second short answer question.

We found, once again, no effects of Syntactic Simplicity, $\beta = .04$, $SE = .08$, $t = .55$, $p = .586$, or Word Concreteness, $\beta = -.02$, $SE = .02$, $t = -1.21$, $p = .267$ for questions. Therefore, as in Experiment 1, test performance was related to linguistic features that were present at the time of encoding, and not retrieval.

Discussion

The preliminary results for Experiment 2 revealed that lower Word Concreteness and Referential Cohesion in lectures are associated with better performance on the short answer questions overall. These data are very preliminary (only approximately 20% of the full sample) and the pattern is likely to change. However, if this pattern holds with additional participants, this suggests that when a lecturer uses fewer concrete words (i.e., uses more abstract, low frequency words) and has less content overlap between sentences, students' performance on short answer questions increases. In contrast, Narrativity, Syntactic Simplicity, and Deep Cohesion were unrelated to students' performance overall. Therefore, the use of more informal language, shorter simpler sentences, and the use of connectives (e.g., but, so, although) to form connections across concepts did not provide a benefit to the students' overall score.

Interestingly, however, there were different patterns observed for the two short answer question types. When students needed to generate an answer about important concepts with no

support, they performed better for longer lectures (as indexed by word count) that contain shorter and simpler sentences and less explicit overlap in content across sentences. On the other hand, when asked to recall specific concepts from the lecture, they performed better for lectures with fewer connectives (e.g., but, so, although) linking concepts. It will be interesting to see whether these patterns hold in the full dataset.

General Discussion

In the present investigation, we set out to examine the linguistic properties of a set of lectures and their relation to students' performance using two types of testing methods (multiple choice and short answer) across two experiments. The results from Experiment 1 revealed that higher narrativity, syntactic simplicity, and referential cohesion in lectures were related to better performance on a multiple-choice test. Increased word count was associated with a decrease in performance. In other words, when the lecturer's language was more informal, used shorter and simpler sentences, and had greater explicit overlap in ideas and concepts across sentences, students' performance at test increased. In contrast, two other linguistic features of the lectures, deep cohesion and word concreteness, were unrelated to students' performance. Therefore, the use of more concrete words and increased used of connectives (e.g., but, so, although) did not provide a benefit, despite past research showing that these two properties are associated with encoding, retrieval, and recall.

The preliminary results from Experiment 2 show that lower word concreteness and referential cohesion are associated with better performance on the short answer questions overall. If this pattern holds, then this suggests that when a lecturer uses fewer concrete words and less content overlap between sentences, students' performance on short answer questions is better. In

contrast, narrativity, syntactic simplicity, and deep cohesion were unrelated to students' performance overall. Therefore, the use of more informal language, shorter and simpler sentences, and the use of connectives (e.g., but, so, although) to link concepts did not provide a benefit to the students' overall score. However, we observed different patterns for the two short answer question types. For the more unconstrained question in which students had to generate important concepts unassisted, higher syntactic simplicity and word counts were associated with better performance, and higher referential cohesion with worse performance. On the other hand, when asked to recall specific information from the lecture, lower deep cohesion was associated with better performance. To our knowledge, the present study is the first attempt to directly examine linguistic properties of a lecturer's language and relate these properties to student performance.

In the following sections, we discuss how narrativity and syntactic simplicity as significant predictors relate to the extant literature in terms of the cognitive processes associated with learning. Further, we explore why high referential cohesion and low word counts may be beneficial for answering multiple choice questions whereas low referential cohesion and higher word counts may be beneficial for answering short answer style questions. Also, we examine why word concreteness was not associated with students' performance in Experiment 1 but was associated with performance in Experiment 2. Our discussion is primarily focused on Experiment 1, given the preliminary nature of the Experiment 2 data. Lastly, we discuss potential limitations of the study and propose directions for further research in this area.

Narrativity and Syntactic Simplicity

The linguistic properties used in our analysis are derived from the Coh-Metrix text analyzing tool, which compares the submitted text to the TASA corpus and outputs a z-score for

each language property based on this comparison (Graesser et al., 2004, 2011; McNamara et al., 2014). The TASA corpus is comprised of texts from various genres and disciplines between kindergarten and college-ready grade levels and contains over 11 million words (Landauer, Foltz, & Laham, 1998). Coh-Metrix has primarily been used to examine the features of textbooks and provide a multidimensional framework for assessing readability and difficulty (e.g., Follmer and Sperling, 2018; Jacob, Lachner, and Sheiter, 2020; McNamara, 2013). However, the tool has also been shown to be useful for characterizing the properties of spoken language (e.g., Li et al., 2013; Medimorec et al., 2015). In the present investigation, we asked whether the linguistic properties of lectures (as assessed by this tool) affect students' learning.

Past research has suggested that, individually, language features associated with the dimensions we looked at (narrativity, referential cohesion, syntactic simplicity, word concreteness, and deep cohesion) affect performance on tasks involving encoding, recall, and comprehension. In terms of narrativity, existing research has focused on the qualitative distinction between narrative (story-like) vs. expository (informational) texts. Whether a text's genre is narrative or expository has been shown to predict student test performance (Graesser, Hauft-Smith, Cohen, & Pyles, 1980; Haberlandt & Graesser, 1985; Ginns, Martin, & Marsh, 2013; Mar et al., 2021). For example, Graesser and colleagues (1980) examined whether the genre (e.g., narrative, expository) of a textbook predicted students' retention of prose on a later multiple-choice test. They observed that more narrative texts were associated with greater retention. Although this effect has been challenged by experiments suggesting that narratives are detrimental to the learning process (Wolfe & Mienko, 2007; Wolfe & Woodwyk, 2010), meta-analyses do provide support for claims that narrative texts are related to better student recall and comprehension (see Ginns, Martin & Marsh, 2013; Mar et al., 2021).

We report an effect of narrativity on student performance in Experiment 1. However, there is a distinction between the extant literature examining genre differences and the work on narrativity as a linguistic property, as determined by Coh-Metrix, though sometimes the term is used interchangeably. As defined by Coh-Metrix (and Graesser et al., 2011), narrativity is comprised of individual linguistic measures that cluster to determine how 'narrative-like' a text is. A text may be expository in genre, but when an individual passage is submitted to Coh-Metrix it could return a high narrativity score. This would indicate that the particular passage has characteristics that make it more narrative-like (such as the use of more familiar words and actions/events that are performed by an animate agent that is motivated by plans in pursuit of goals). This can be seen in the linguistic profile of the present lecture set. Although all of the lectures were from a single domain (psychology), they vary in narrativity as a linguistic property. Thus, it is not only the case that different genres are more or less narrative and thus easier/more difficult to process and/or encode, but also true that variations in narrativity *within* a genre can affect these processes. To our knowledge, this is the first such demonstration that within-genre narrativity may affect students' performance.

A great deal of literature has documented that shorter and simpler sentences are better recalled, understood, and read (Caplan & Waters, 1999). This literature proposes that processing and comprehension of syntactic structure is guided by working memory (Lewis, Vasishth & Van Dyke, 2006; Vasishth, Nicenboim, Engelman, & Burchert, 2019). Predominant models of syntactic processing suggest that complex sentences take longer to process and cause more errors because there is more interference between the target words and similar words stored in memory and an increased load associated with storing more words (see Lewis, Vasishth, & Van Dyke, 2006 for a review). Simpler syntax likely reduces the load on working memory, facilitating

comprehension (Caplan & Waters, 1999; MacDonald, 1997). Consistent with this, we found that increases in the syntactic simplicity of a lecture were associated with increased student performance.

Across the two experiments, we found effects of narrativity for multiple-choice questions and effects of syntactic simplicity for both multiple choice and short answer questions. This means that when a lecturer presents the material in a story-like way, coupled with the use of shorter and simpler sentences, students' performance on multiple-choice questions increases. For short answer questions, only the presence of shorter, simpler sentences is associated with better performance. It may be that narrativity and syntactic simplicity facilitate better comprehension - a lecturer presenting a concept or theory within a narrative like structure coupled with shorter and simpler sentences likely allows students to better understand the material as it is being produced. This is supported by literature that calls for the use of more narrative-like language to facilitate better comprehension (Ginns, Martin & Marsh, 2013; Mar et al., 2021) and demonstrates how simpler syntax reduces the cognitive resources used for processing (Caplan and Waters, 1999; Lewis, Vasishth, & Van Dyke, 2006). Together, these findings support prevailing theories about effective student instruction, in which managing students' cognitive load is at the forefront of lecture design (see Mayer, 2019 for a review).

Cohesion: Two Different Patterns

Experiment 1 revealed that higher referential cohesion was associated with an increase in multiple-choice test performance. In contrast, the results of Experiment 2 suggest that higher cohesion is associated with worse short answer performance. Why is it that we see such a dissociation between testing methods in the effects of cohesion? A recent study examined how the linguistic properties of middle school textbooks affect students' comprehension using a

Think-aloud task (Dahl et al., 2021). The study revealed that, for texts low in referential cohesion, students drew inferences based on key words directly from the text, as well as inferences based on relevant prior knowledge. In contrast, for texts high in referential cohesion, there were more evaluative comments about the content of the text, paraphrasing, and metacognitive comments (e.g., reflecting on their understanding). It could be that higher referential cohesion facilitates encoding and comprehension in a way that allows students to succeed on multiple-choice questions (where less retrieval is involved; Kang, McDemott, & Roediger, 2007; Ozuru et al., 2013), but that lower cohesion leads to more inferencing from prior knowledge that is beneficial in self-generated recall (as is required for short answer questions; Jacoby, 1996).

These findings might relate to the *reverse cohesion effect* (O'Reilly & McNamara, 2007), in which the benefits of cohesion depend on the student's prior knowledge. For example, if a student has minimal prior knowledge, they benefit from discourse that is more cohesive, as it allows them to fill in the gaps in their knowledge. In contrast, students with more prior knowledge benefit from less cohesion, because they can use their prior knowledge to draw inferences and conclusions (O'Reilly & McNamara, 2007; McNamara, 2013). Even though we sought to control for prior knowledge by limiting the participant pool to students who had not taken the course, it is possible that some participants had prior knowledge about the topics presented. Although high referential cohesion might aid participants in learning the concepts sufficiently to recognise the correct answer on multiple choice questions (e.g., referential cohesion might increase the familiarity of certain multiple-choice items; Ozuru et al., 2013), lower cohesion might force participants to engage more with the material and actively integrate

that information with their prior knowledge base. For knowledgeable participants, lower cohesion might therefore lead to better performance on more challenging short answer questions.

More Words, Worse Performance?

In Experiment 1, lectures with higher word counts led to worse performance. One reason for a higher word count is that a lecture includes more material. It is unsurprising that performance might suffer when there is more material to be learned. Indeed, one of the main instructional recommendations in the online learning literature is that instructors should manage essential processing of the information (see Mayer, 2019 for a review). Essential processing refers to the cognitive processing needed for the student to mentally represent important information presented in the lecture. For online lectures, managing essential processing could take the form of user paced segments that allow the student to focus on the critical material of one section and continue at their own pace to the next section to build upon what they learned (Low & Sweller, 2014; Mayer & Pilegard, 2014). What is surprising is that we were able to detect an effect of word count within quite a limited range of lecture lengths (5 to 15 minutes). These were not the standard 90 to 120 minute lectures that are typical of in-class lectures. It is also important to note that the effects of narrativity, syntactic simplicity, and referential cohesion held in Experiment 1 even when we controlled for word count, demonstrating that these properties influence student learning above and beyond the load caused by the amount of material being processed. However, if the pattern of results from Experiment 2 holds, this would suggest that lectures with higher word counts are actually beneficial for short answer questions that ask participants to generate and describe a list of important concepts. The intuitive interpretation here may be that, if lectures with higher word counts have more material, the student can draw on more content to generate and describe when it comes to time to take the test.

However, it is possible that this benefit of having more content is only observed when the content is easier to process, as demonstrated by the contribution of syntactic simplicity to student performance for these types of questions.

Word Concreteness

Past research led us to expect that higher concreteness would lead to improved performance (*Word Concreteness*: e.g., Garlock, Walley, & Metsala, 2001; Griffin & Bock, 1998; Graesser et al., 2011; McNamara, 2013; Nelson & Schreiber, 1992; Sadoski & Paivio, 2013). However, we did not observe significant effects of this feature for Experiment 1 and, interestingly, we observed a reverse effect in Experiment 2.

One possible explanation for the absence of a word concreteness effect in Experiment 1 is that the lectures contained relatively abstract words overall. In the present set of lectures, word concreteness scores ranged from -1.63 to -.49 when compared to the TASA corpus used by Coh-Metrix. However, generally low word concreteness is not surprising in this context given that it has been shown to be generally low in texts from grade levels 11 on (Graesser et al., 2011). But, intriguingly, lower word concreteness was related to increased performance on short answer questions overall. If this pattern holds, then we speculate that it may be the case that students pay more attention when unfamiliar words are presented in the lecture. This heightened attention may lead to stronger learning that promotes better performance on the harder short answer questions (when students cannot rely on familiarity of the presented answers).

Limitations

The results of Experiment 2 are preliminary and are presented here to allow us to speculate about any differences in how different linguistic features might affect students'

performance on different testing methods. Data is still being collected and the final results may look different than the ones reported here. Thus, the effects of linguistic features reported for Experiment 2 and any comparisons with Experiment 1 should be taken with a grain of salt.

The present set of lectures was created for an asynchronous online course (e.g., all resources available at the beginning of the course, no meetings or interaction with students during the sessions). This type of delivery may lead to features of both live lectures and text. For example, the language used in live lectures is less organized and less concise than the corresponding textbook material due to the amount of planning involved (Glass et al., 2005). If an asynchronous online lecture is planned ahead of time, then it may have greater syntactic organization and a more topic focussed scope (e.g., if the transcript is created ahead of time and read word-for-word). However, we asked the lecturer to explain their process for generating the lectures we used here. They stated that they produced the lectures over the accompanying slides as if the lectures were taking place in a live setting and that the corresponding transcripts were generated after the fact (Personal Communications, Johnathan Fugelsang, 2021). Therefore, there is good reason to believe that the properties found here are in some respects characteristic of spoken lectures produced live as well. However, one clear difference between a live lecture and an asynchronous online one is the presence (or absence) of paralinguistic cues (e.g., looks of confusion, confirmatory head nods) from the audience. These cues may alter the language a live lecturer decides to use in a given moment. It would be interesting for further research to examine the effects of these cues on the linguistic properties of lectures.

The lectures used in our experiment were produced by a single lecturer from a single course. However, there are reasons to be confident that the language properties observed here would generalize to other lecturers. In particular, the linguistic profile of our 18-part lecture

corpus shares similarities with the 54-lecture corpus analyzed in Medimorec et al. (2015). That corpus spanned multiple disciplines (Humanities = 26; Natural Sciences = 28) and institutions (Yale, MIT, and University of Michigan). In particular, we observed that word concreteness and syntactic simplicity were negatively correlated with all other linguistic features, and observed positive correlations between narrativity, referential cohesion and deep cohesion. These correlations were similar to those observed in their humanities lectures (which included Psychology). These similarities suggest the patterns we observed are not lecturer-specific. However, our significant correlations are stronger than those reported in the humanities lecturers in Medimorec et al. (2015).

Conclusion

In conclusion, the current study represents the first attempt to directly examine the influence of a lecture's language properties on students' performance at test. The results suggest that more informal language, with shorter and simpler sentences and greater overlap across sentences leads to better test performance on some types of assessments. These results have implications for the study of academic language use and its effects on student learning. If we want to optimize learning in lectures, then we need to consider the language of the lecture itself in the current conceptualization of the lecture experience, not just at a surface level (e.g., "friendly" tone) but at the linguistic feature level as well.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1 - 48.

Britton, B. K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*, 329–345.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences, 22*, 77–94.

Cain, K., & Nash, H. M. (2011). The influence of connectives on young readers' processing and comprehension of text. *Journal of Educational Psychology, 103,* 429–441.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A, 33*(4), 497-505.

Dahl, A. C., Carlson, S. E., Renken, M., McCarthy, K. S., & Reynolds, E. (2021). Materials Matter: An Exploration of Text Complexity and Its Effects on Middle School Readers' Comprehension Processing. *Language, Speech, and Hearing Services in Schools*, *52*(2), 702-716.

Favier, S., Meyer, A. S., & Huettig, F. (2021). Literacy can enhance syntactic prediction in

spoken language processing. *Journal of Experimental Psychology: General*. Advanced online publication.

Follmer, D. J., & Sperling, R. A. (2018). Interactions between reader and text: Contributions of cognitive processes, strategy use, and text cohesion to comprehension of expository science text. *Learning and Individual Differences, 67*, 177-187.

Garlock, V. M., Walley, A. C., & Metsala, J. L. (2001). Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults. *Journal of Memory and Language, 45*(3), 468-492.

Ginns, P., Martin, A. J., & Marsh, H. W. (2013). Designing instructional text in a conversational style: A meta-analysis. *Educational Psychology Review*, *25*(4), 445-472.

Glass, J., Hazen, T. J., Hetherington, L., & Wang, C. (2004). Analysis and processing of lecture audio data: Preliminary investigations. In: *Proceedings of the Workshop on Interdisciplinary Approaches to Speech Indexing and Retrieval at HLT-NAACL 2004* (pp. 9–12). Stroudsburg, PA: Association for Computational Linguistics.

Graesser, A. C., Hauft-Smith, K., Cohen, A. D., & Pyles, L. D. (1980). Advanced outlines, familiarity, and text genre on retention of prose. *The Journal of Experimental Education, 48*(4), 281-290.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal, 115*(2), 210-229.

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

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193-202.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101*(3), 371.

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, *38*(3), 313-338.

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114,* 357–374.

Jacob, L., Lachner, A., & Scheiter, K. (2020). Learning by explaining orally or in written form? Text complexity matters. *Learning and Instruction*, *68,* 101344.

Jacoby, L. L. (1996). Dissociating automatic and consciously-controlled effects of study/test compatibility. *Journal of Memory and Language, 35*, 32–52.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology*, *19*, 528–558.

Kintsch, W., Kozminsky, E., Streby, W. J., McKoon, G., & Keenan, J. M. (1975). Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior, 14*, 196–214.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259–284.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447-454.

Li, H., Cai, Z., & Graesser, A. C. (2013, May). Comparing two measures for formality. In *FLAIRS Conference* (pp. 220–225). Palo Alto, California: AAAI Press.

Liu, C. J., & Rawl, S. M. (2012). Effects of text cohesion on comprehension and retention of colorectal cancer screening information: A preliminary study. *Journal of Health Communication*, *17*(sup3), 222-240.

Low, R., & Sweller, J. (2014). The Modality Principle in Multimedia Learning. In R. Mayer (Ed.), *The Cambridge Handbook of Multimedia Learning* (Cambridge Handbooks in Psychology, pp. 227-246). Cambridge: Cambridge University Press.

Mar, R. A., Li, J., Nguyen, A. T., & Ta, C. P. (2021). Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review*, *28*, 732-749.

Mayer, R. E. (2011). Applying the science of learning. Boston: Pearson.

Mayer, R. E. (2019). Thirty years of research on online learning. *Applied Cognitive*

Psychology, *33*(2), 152-159.

Mayer, R. E., & Pilegard, C. (2014). Principles for managing essential processing in multimedia learning: Segmenting, pre‐training, and modality principles. In R. E. Mayer (Ed.*), The Cambridge Handbook of Multimedia Learning* (2nd ed.) (pp. 316‐345). New York: Cambridge University Press.

McDaniel, M. A., & Einstein, G. O. (1989). Material-appropriate processing: A contextualist approach to reading and studying strategies. *Educational Psychology Review*, *1*(2), 113-145.

McDaniel, M. A., & Einstein, G. O. (2005). Material Appropriate Difficulty: A Framework for Determining When Difficulty Is Desirable for Improving Learning. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 73–85). American Psychological Association.

McNamara, D. S. (2013). The epistemic stance between the author and reader: A driving force in the cohesion of text and writing. *Discourse Studies*, *15*, 579–595.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In J. P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89–116). Plymouth, UK: Rowman & Littlefield Education.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

Medimorec, S., Pavlik Jr, P. I., Olney, A., Graesser, A. C., & Risko, E. F. (2015). The language of instruction: Compensating for challenge in lectures. *Journal of Educational Psychology*, *107*(4), 971.

Medimorec, S., Young, T. P., & Risko, E. F. (2017). Disfluency effects on lexical selection. *Cognition*, *158*, 28-32.

Morgan B., Burkett C., Bagley E., Graesser A. (2011) Typed versus Spoken Conversations in a Multi-party Epistemic Game. In: Biswas G., Bull S., Kay J., Mitrovic A. (eds) *Artificial Intelligence in Education*. AIED 2011. Lecture Notes in Computer Science, vol 6738. Springer, Berlin, Heidelberg.

Nelson, D. L., & Schreiber, T. A. (1992). Word Concreteness and word structure as independent determinants of recall. *Journal of Memory and Language, 31,* 237–260.

Noble, C. E. (1963). Meaningfulness and Familiarity. In *Conference on Verbal Learning and Verbal Behavior, 2nd, Jun, 1961, Ardsley-on-Hudson, NY, US*. McGraw-Hill Book Company.

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

Oswald, F. L., McAbee, S. T., Redick, T. S., & Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behavior Research Methods*, *47*(4), 1343-1355.

Ozuru, Y., Briner, S., Kurby, C. A., & McNamara, D. S. (2013). Comparing comprehension measured by multiple-choice and open-ended questions. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *67*(3), 215.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sadoski, M., Goetz, E. T., & Fritz, J. B. (1993). Impact of concreteness on comprehensibility, interest, and memory for text: Implications for dual coding theory and text design. *Journal of Educational Psychology, 85*, 291–304.

Sadoski, M., & Paivio, A. (2013). *Imagery and text: A dual coding theory of reading and writing*. New York, NY: Routledge.

Shadiev, R., & Huang, Y. M. (2020). Investigating student attention, meditation, cognitive load, and satisfaction during lectures in a foreign language supported by speech-enabled language translation. *Computer Assisted Language Learning*, *33*(3), 301-326.

Tin, T. B. (2009). Features of the most interesting and the least interesting postgraduate second language acquisition lectures offered by three lecturers. *Language and Education*, *23*(2), 117-135.

Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, *23*(11), 968-982.

Vidal-Abarca, E., Martínez, G., & Gilabert, R. (2000). Two procedures to improve instructional text: Effects on memory and learning. *Journal of Educational Psychology, 92,* 107–116

Walker, I., & Hulme, C. (1999). Concrete words are easier to recall than abstract words: Evidence for a semantic contribution to short-term serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1256.

Weninger, F., Staudt, P., & Schuller, B. (2013). Words that fascinate the listener: Predicting affective ratings of on-line lectures. *International Journal of Distance Education Technologies (IJDET), 11*(2), 110-123.

Wolfe, M. B., & Mienko, J. A. (2007). Learning and memory of factual content from narrative and expository text. *British Journal of Educational Psychology*, *77*(3), 541-564.

Wolfe, M. B., & Woodwyk, J. M. (2010). Processing and memory of information presented in narrative or expository texts. *British Journal of Educational Psychology*, *80*(3), 341-362.

Zare, J. (2019). Awareness of discourse organizers and comprehension of academic lectures: The effect of using concordancers. *Current Psychology*, 1-9.

Zare, J., & Keivanloo-Shahrestanaki, Z. (2017). The language of English academic lectures: The case of field of study in highlighting importance. *Lingua*, *193*, 36-50.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123,* 162–185

Appendix A

List of the lecture sets, lecture topics, and descriptions of the content. Each participant was assigned to one set.

Lecture Set	*Topic*	*Description*
1	Antecedents to Cognitive Revolution	Examines the theories (e.g., Empiricism, Nativism, Behaviourism, Gestalt, Structuralism, Functionalism) that preceded cognitive psychology as a discipline.
1	Early theories of localisation in the brain	Examines theories of and evidence for localisation of function in the brain, including work by Franz Gall (Phrenology), double dissociations in aphasia, and Penfield's Montreal procedure.
1	Utility models of decision making	Examines expected utility theory, Image theory, and Recognition primed decision making.
2	Introduction to brain imaging techniques	Examines flaws with lesion studies and provides background on imaging techniques, including static (CAT and MRI) and Dynamic (fMRI, EEG, PET) techniques. Explains BOLD function and subtractive logic.
2	General problem-solving techniques	Examines techniques of problem solving, such as generate and test, means end analysis, working backwards, and reasoning by analogy.
2	Introduction to categories and concepts	Defines what a concept and a category are and explains categorization as a process. Explains why we have concepts and categories.
3	Paradigms of cognitive psychology	Defines what a paradigm is and highlights the main paradigms of cognitive psychology: Information processing, connectionism, ecological, and evolutionary.
3	Short-term memory	Defines short-term memory. Examines various effects associated with short-term memory (e.g., primacy vs. recency effects) and theories of forgetting (decay, interference).
3	Heuristics and biases	Defines heuristics and biases and explains how they are a reflection of normal cognitive processes. Lists six heuristics and biases and provides examples for each one.
4	Neurological studies of memory	Defines semantic memory, episodic memory, and procedural memory with examples from patients with acquired brain injury. Defines anterograde and retrograde amnesia.

4	Precursors to the Cognitive Revolution	Talks about several key historical events that acted as precursors to the cognitive revolution: limitations of human cognition highlighted by human factors engineering, developments in linguistics that challenged the behaviourist view on language acquisition, developments in neuroscience leading to localization of function in the brain, and technological advancements leading to the computer metaphor of the mind.
4	Theoretical descriptions of the nature of concepts	Describes multiple views about how concepts and categories are formed: Classical, Prototype, Exemplar, and Schemata.
5	Sensory memory	Defines encoding and retrieval of information. Explains the Atkins-Shiffrin model of memory. Examples from Sperling's research to show the capacity limitations of sensory memory.
5	Blocks to problem solving	Provides examples of tasks where the typical responses are hindered by conventional thinking and imposing rules that are not there. Defines mental set and functional fixedness. Talks about how expertise affects problem solving to highlight domain-specific and domain-general problem solving.
5	Introduction to brain structure	Defines the phylogenetic division with a focus on the forebrain region. Talks about sub cortical regions of the brain (thalamus, hypothalamus, hippocampus, amygdala) and the four lobes in the cerebral cortex.
6	Long-term memory	Defines what long-term memory is in terms of capacity and how items are stored. Talks about retrieval cues and forgetting. Highlights encoding specificity principle.
6	Forming new concepts and classifying new instances	Examines strategies of concept formation, including: successive scanning, simultaneous scanning, and conservative focussing. Explains the trade-off between working memory resources and efficiency of the strategy. Describes neural evidence for different strategies in early vs. late learning. Discusses implicit and explicit learning.
6	Reasoning	Defines reasoning and how it is different from thinking. Defines and provides examples of inductive and deductive reasoning, and how they can be flawed. Defines rule-based and mental model approaches to reasoning.

Appendix B

Table of each lecture's linguistic features (word count and Coh-Metrix Principal Components z score)

Lecture Set	Word Count	Narrativity	Syntactic Simplicity	Word Concreteness	Referential Cohesion	Deep Cohesion
1	1860	-0.822	0.019	-1.281	-0.452	-0.23
1	1499	0.18	-0.26	-1.515	0.659	2.961
1	1449	-0.617	-0.071	-0.968	0.141	0.519
2	886	0.191	-0.198	-0.486	0.11	2.003
2	2068	0.15	-0.22	-0.561	0.492	1.405
2	1902	-0.425	-0.269	-0.585	0.673	0.497
3	2300	0.32	-0.329	-1.141	0.242	2.429
3	1352	0.073	-0.171	-1.177	-0.191	0.91
3	1443	-0.868	0.316	-1.057	-0.124	-0.007
4	846	-0.172	0.38	-1.168	-0.292	1.306
4	978	-0.8	0.067	-0.515	-1.091	-0.509
4	3491	-0.031	-0.205	-1.468	0.101	0.757
5	1054	-0.875	-0.048	-1.266	0.982	-0.971
5	2266	0.267	-0.37	-1.096	0.968	0.806
5	1341	0.001	-0.291	-0.707	0.873	1.224
6	1591	0.068	0.071	-1.436	0.332	1.296
6	1725	-0.108	-0.126	-1.436	0.106	0.659
6	2069	-0.059	0.13	-1.633	0.165	0.579